1 A Review of Correlation and Regression

SW, Chapter 12

Suppose we select $n = 10$ persons from the population of college seniors who plan to take the MCAT exam. Each takes the test, is coached, and then retakes the exam. Let X_i be the pre-coaching score and let Y_i be the post-coaching score for the i^{th} individual, $i = 1, 2, \cdots, n$. There are several questions of potential interest here, for example: Are Y and X related (associated), and how? Does coaching improve your MCAT score? Can we use the data to develop a mathematical model (formula) for predicting post-coaching scores from the pre-coaching scores? These questions can be addressed using **correlation** and **regression** models.

The **correlation coefficient** is a standard measure of **association** or relationship between two features Y and X. Most scientists equate Y and X being correlated to mean that Y and X are associated, related, or **dependent** upon each other. However, correlation is only a measure of the strength of a **linear relationship.** For later reference, let ρ be the correlation between Y and X in the population and let r be the sample correlation. I define r below. The population correlation is defined analogously from population data.

Suppose each of n sampled individuals is measured on two quantitative characteristics called Y and X. The data are pairs of observations $(X_1,\ Y_1)$, $(X_2,\ Y_2)$, $\cdots (X_n,\ Y_n)$, where $(X_i,\ Y_i)$ is the $(X,\ Y)$ pair for the i^{th} individual in the sample. The sample correlation between Y and X, also called the **Pearson product moment correlation coefficient**, is

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}},$$

where

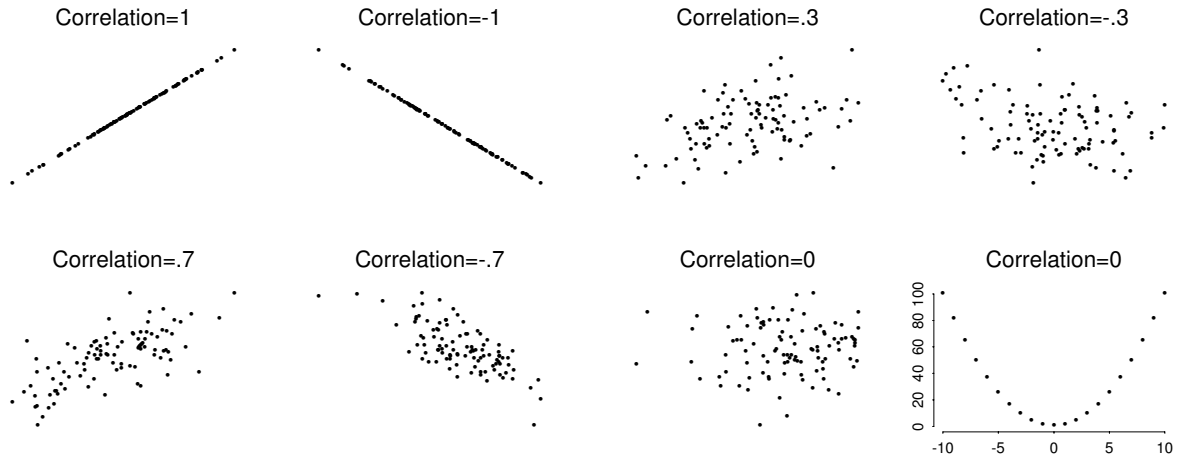
$$S_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

is the **sample covariance** between Y and X, and $S_Y = \sqrt{\sum_i (Y_i - \bar{Y})^2/(n-1)}$ and $S_X = \sqrt{\sum_i (X_i - \bar{X})^2/(n-1)}$ are the standard deviations for the Y and X samples. Here are eight important properties of r:

1. $-1 \leq r \leq 1$.
2. If Y_i tends to increase linearly with X_i then $r > 0$.
3. If Y_i tends to decrease linearly with X_i then $r < 0$.
4. If there is a perfect linear relationship between Y_i and X_i with a positive slope then $r = +1$.
5. If there is a perfect linear relationship between Y_i and X_i with a negative slope then $r = -1$.
6. The closer the points (X_i, Y_i) come to forming a straight line, the closer r is to ± 1.
7. The magnitude of r is unchanged if either the X or Y sample is transformed linearly (i.e. feet to inches, pounds to kilograms, Celsius to Fahrenheit).
8. The correlation does not depend on which variable is called Y and which is called X.

If r is near ± 1, then there is a strong linear relationship between Y and X in the sample. This suggests we might be able to accurately predict Y from X with a linear equation (i.e. linear regression). If r is near 0, there is a weak linear relationship between Y and X, which suggests that a linear equation provides little help for predicting Y from X. The pictures below should help you develop a sense about the size of r.

Note that $r = 0$ does not imply that Y and X are not related in the sample. It only implies they are not linearly related. For example, in the lower right plot on the following set of plots, $r = 0$ yet $Y_i = X_i^2$.



Testing that $\rho = 0$

Suppose you want to test $H_0 : \rho = 0$ against $H_A : \rho \neq 0$, where ρ is the population correlation between Y and X. This test is usually interpreted as a test of no association, or relationship, between Y and X in the population. Keep in mind, however, that ρ measures the strength of a linear relationship.

The standard test of $H_0 : \rho = 0$ is based on the magnitude of r. If we let

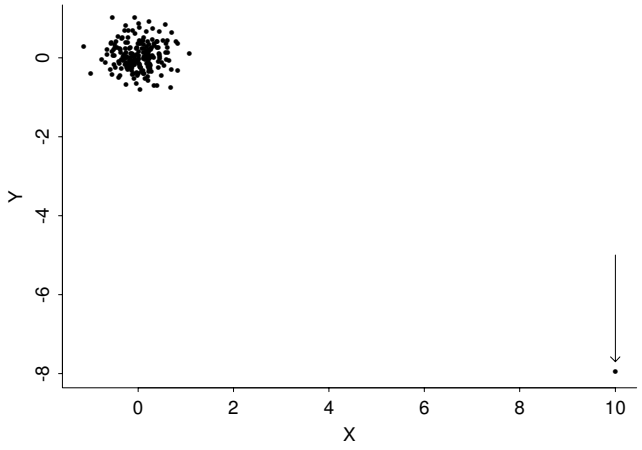
$$t_s = r\sqrt{\frac{n-2}{1-r^2}},$$

then the test rejects H_0 in favor of H_A if $|t_s| \geq t_{crit}$, where t_{crit} is the two-sided test critical value from a t-distribution with $df = n-2$. The p-value for the test is the area under the t-curve outside $\pm t_s$ (i.e. two-tailed test p-value).

This test assumes that the data are a random sample from a **bivariate normal population** for $(X,\ Y)$. This assumption implies that all linear combinations of X and Y, say $aX + bY$, are normal. In particular, the (marginal) population frequency curves for X and Y are normal. At a minimum, you should make boxplots of the X and Y samples to check marginal normality. For large-sized samples, a plot of Y against X should be roughly an elliptical cloud, with the density of the points decreasing as the points move away from the center of the cloud.

The Spearman Correlation Coefficient

The Pearson correlation r can be highly influenced by outliers in one or both samples. For example, $r \approx -1$ in the plot above. If you delete the one extreme case with the largest X and smallest Y value then $r \approx 0$. The two analyses are contradictory. The first analysis (ignoring the plot) suggests a strong linear relationship, whereas the second suggests the lack of a linear relationship. I will not strongly argue that you should (must?) delete the extreme case, but I am concerned about any conclusion that depends heavily on the presence of a single observation in the data set.



Spearman's rank correlation coefficient r_S is a sensible alternative to r when normality is unreasonable or outliers are present. Most books give a computational formula for r_S. I will verbally describe how to compute r_S. First, order the X_is and assign them ranks. Then do the same for the Y_is and replace the original data pairs by the pairs of ranked values. The Spearman rank correlation is the Pearson correlation computed from the pairs of ranks.

The Spearman correlation r_S estimates the **population rank correlation coefficient**, which is a measure of the strength of linear relationship between population ranks. The Spearman correlation, as with other rank based methods, is not sensitive to the presence of outliers in the data. In the plot above, $r_S \approx 0$ whether the unusual point is included or excluded from the analysis. In samples without unusual observations and a linear trend, you often find that $r_S \approx r$.

An important point to note is that the magnitude of the Spearman correlation does not change if either X or Y or both are transformed (monotonically). Thus, if r_S is noticeably greater than r, a transformation of the data might provide a stronger linear relationship.

Example

Eight patients underwent a thyroid operation. Three variables were measured on each patient: weight in kg, time of operation in minutes, and blood loss in ml. The scientists were interested in the factors that influence blood loss.

weight	time	blood loss
44.3	105	503
40.6	80	490
69.0	86	471
43.7	112	505
50.3	109	482
50.2	100	490
35.4	96	513
52.2	120	464

We are using **Stata** for the computations in this course. As with most packages, there are many ways to get data into the package. With such a small data set it is reasonable to enter it directly, but that is usually a fairly awkward way to do it. Another (older) method is to place the data in a flat text file and use commands to read it in. The method I prefer is to import from an Excel spreadsheet, since most researchers record data that way. A sketch of these methods follows:

Direct: Follow the menu path in Stata of `Data -> Data Editor` to bring up the spreadsheet-like Stata Editor. You have to enter data before you can name columns.

Flat Text File: Put the data into a flat text file (just the numbers), named say `C:\biostat\bloodloss.txt`. In the Stata Command window type

```
infile weight time loss using "c:\biostat\bloodloss.txt"
```

This names the variables while reading them in. The " marks let you use modern path structures in the name (spaces, etc.). There also are options under `File -> Import` to do this.

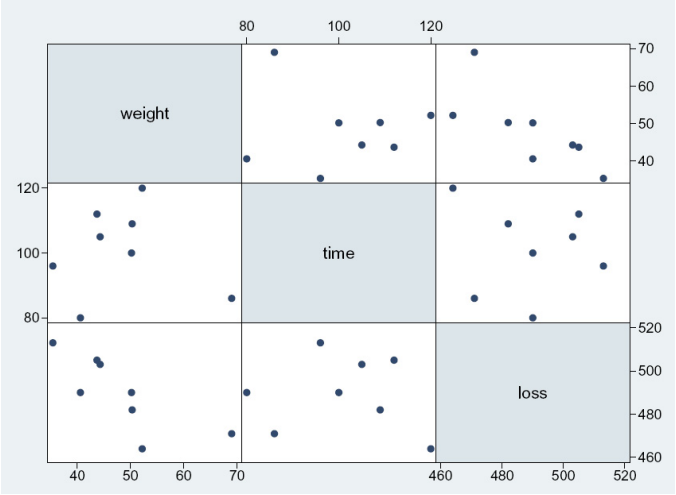
Excel Import: Put the data into an Excel spreadsheet with names in the first row of columns. You would think you could find `Excel spreadsheet` in `File -> Import` but no such luck. Open the spreadsheet, highlight the data (or the whole sheet), follow the menu path in Stata of `Data -> Data` Editor and paste into the data editor.

There are other ways to get data into Stata. I use a program named DBMSCopy to import SAS and Splus datasets into Stata. The easiest method I have found is the Excel Import above. No matter how you get it in, do follow the path `File -> Save` As to save a copy of your data in Stata format (it will be a .dta file). To get a scatterplot matrix of the variables, follow the menu path `Graphics -> Easy Graphs -> Scatterplot Matrix`. This will bring up a dialog box in which you enter variable names. Note that there is a Variables window - you can just click variable names there and they will be entered in the dialog box. Also note when you click Submit on the dialog box, commands are recorded in the Review window. If you click on one of those commands, it is entered in the Stata Command window where you can edit it and submit it (by hitting the Enter key) without all the clicking through menus. The command line from the scatterplot matrix is `graph matrix weight time loss`.

We also want to calculate the correlations and p-values, which we can do by following the menu path

```
Statistics -> Summaries, tables & tests -> Summary statistics -> Pairwise
correlations
```

and fill in the dialog box (all you need to do here is to click the box to request significance levels). The command line generated by this is pwcorr, sig. The results of these two operations are on the next page.




```
             |   weight     time     loss
-------------+------------------------------
      weight |   1.0000
             |
             |
        time |  -0.0663   1.0000
             |   0.8761
             |
        loss |  -0.7725  -0.1073   1.0000
             |   0.0247   0.8003
             |
```

In order to get the Spearman rank correlation coefficients in the same form, follow the menu path

`Statistics -> Summaries, tables & tests -> Nonparametric tests of hypotheses -> Spearman's rank correlation`

or use the command line `spearman, stats(rho p)` to get the following

```
 +-----------------+
 |  Key            |
 |-----------------|
 |   rho           |
 |   Sig. level    |
 +-----------------+

             |   weight     time     loss
-------------+------------------------------
      weight |   1.0000
             |
             |
        time |   0.2857   1.0000
             |   0.4927
             |
        loss |  -0.8743  -0.1557   1.0000
             |   0.0045   0.7128
             |
```

Comments:

1. (Pearson correlations). Blood loss tends to decrease linearly as weight increases, so r should be negative. The output gives $r = -.77$. There is not much of a linear relationship between blood loss and time, so r should be close to 0. The output gives $r = -.11$. Similarly, weight and time have a weak negative correlation, $r = -.07$.

2. The Pearson and Spearman correlations are fairly consistent here. Only the correlation between blood loss and weight is significant at the $\alpha = 0.05$ level (the p-values are given in the rightmost column).

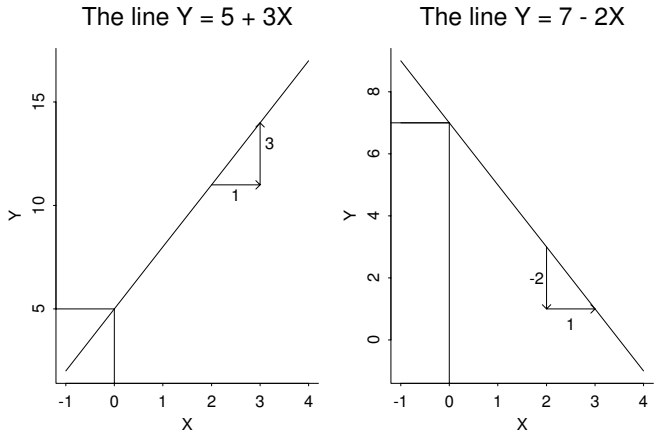
3. Another measure of association available in Stata is Kendall's τ (tau, not given here).

Simple Linear Regression

In linear regression, we are interested in developing a linear equation that best summarizes the relationship in a sample between the **response variable** Y and the **predictor variable** (or **independent variable**) X. The equation is also used to predict Y from X. The variables are not treated symmetrically in regression, but the appropriate choice for the response and predictor is usually apparent.

Linear Equation

If there is a perfect linear relationship between Y and X then $Y = \beta_0 + \beta_1 X$ for some β_0 and β_1, where β_0 is the Y-intercept and β_1 is the slope of the line. Two plots of linear relationships are given below. The left plot has $\beta_0 = 5$ and $\beta_1 = 3$. The slope is positive, which indicates that Y increases linearly when X increases. The right plot has $\beta_0 = 7$ and $\beta_1 = -2$. The slope is negative, which indicates that Y decreases linearly when X increases.



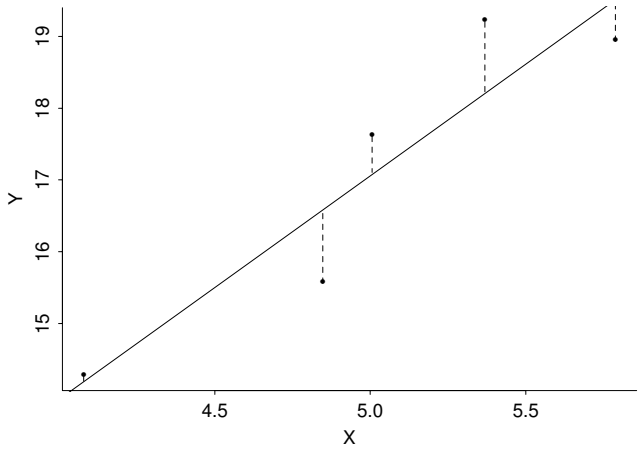
Least Squares

Data rarely, if ever, fall on a straight line. However, a straight line will often describe the **trend** for a set of data. Given a data set (X_i, Y_i), $i = 1, ..., n$ with a **linear trend**, what linear equation "best" summarizes the observed relationship between Y and X? There is no universally accepted definition of "best", but many researchers accept the **Least Squares** line (LS line) as a reasonable summary.

Mathematically, the LS line chooses the values of β_0 and β_1 that minimize

$$\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

over all possible choices of β_0 and β_1. These values can be obtained using calculus. Rather than worry about this calculation, note that the LS line makes the sum of squared deviations between the responses Y_i and the line as small as possible, over all possible lines. The LS line typically goes through "the heart" of the data, and is often closely approximated by an eye-ball fit to the data.



The equation of the LS line is

$$\hat{Y} = b_0 + b_1 X$$

where the intercept b_0 satisfies

$$b_0 = \bar{Y} - b_1 \bar{X}$$

and the slope is

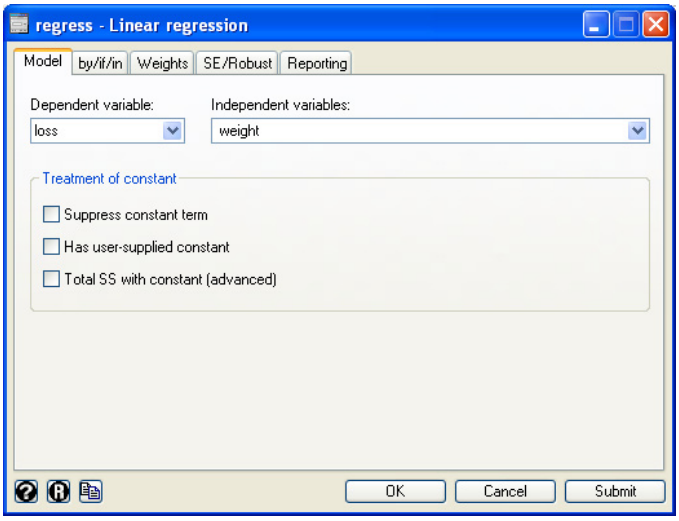
$$b_1 = \frac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} = r \frac{S_Y}{S_X}.$$

As before, r is the Pearson correlation between Y and X, whereas S_Y and S_X are the sample standard deviations for the Y and X samples, respectively. The **sign of the slope** and the **sign of the correlation** are **identical** (i.e. + correlation implies + slope).

Special symbols b_0 and b_1 identify the LS intercept and slope to distinguish the LS line from the generic line $Y = \beta_0 + \beta_1 X$. You should think of $\hat{Y}$ as the **fitted value** at X, or the value of the LS line at X.

Stata Implementation

A least squares fit of a line is carried out in Stata using the menu path `Statistics -> Linear regression and related -> Linear regression` (or the regress command). For the thyroid operation data with Y = Blood loss in ml and X = Weight in kg, we regress blood loss on the patients' weight by filling in the dialog box as below (or using the command regress loss weight), and obtain the following output




```
. regress loss weight
      Source |       SS       df       MS              Number of obs =       8
-------------+------------------------------           F(  1,     6) =    8.88
       Model |  1207.45125     1  1207.45125            Prob > F      =  0.0247
    Residual |  816.048753     6  136.008125            R-squared     =  0.5967
-------------+------------------------------           Adj R-squared =  0.5295
       Total |      2023.5     7  289.071429            Root MSE      =  11.662

------------------------------------------------------------------------------
        loss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -1.300327   .4364156    -2.98   0.025    -2.368198   -.2324567
       _cons |    552.442   21.44088    25.77   0.000     499.9781     604.906
------------------------------------------------------------------------------
```

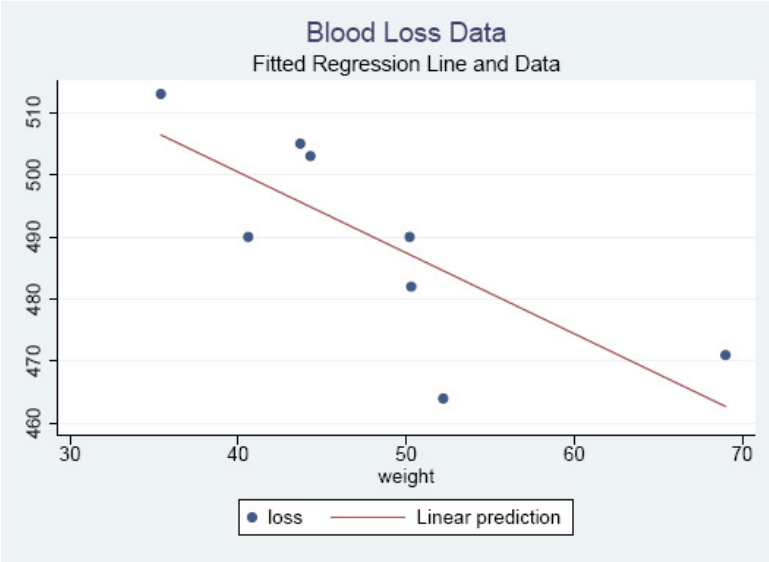
For the **thyroid operation data** with Y = Blood loss in ml and X = Weight in kg, the LS line is $\hat{Y} = 552.44 - 1.30X$, or Predicted Blood Loss = $552.44 - 1.30$ Weight. For an $86kg$ individual, the Predicted Blood Loss = $552.44 - 1.30 * 86 = 440.64ml$. The LS regression coefficients for this model are interpreted as follows. The intercept b_0 is the predicted blood loss for a 0 kg individual. The intercept has no meaning here. The slope b_1 is the predicted increase in blood loss for each additional kg of weight. The slope is -1.30, so the predicted *decrease* in blood loss is 1.30 ml for each increase of 1 kg in weight.

Any fitted linear relationship holds only approximately and does not necessarily extend outside the range of the data. In particular, nonsensical predicted blood losses of less than zero are obtained at very large weights outside the range of data.

To obtain a plot of the line superimposed on the data, a general- purpose approach is as follows:

```
. predict yhat,xb
. twoway (scatter loss weight) (line yhat weight,sort), title(Blood Loss Data)
> subtitle(Fitted Regression Line and Data)
```

The first command puts the predicted values in a new variable named yhat. Everything can be accomplished through dialog boxes if you forget some of the syntax.



ANOVA Table for Regression

The LS line minimizes

$$\sum_{i=1}^{n}\{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

over all choices for β_0 and β_1. Inserting the LS estimates b_0 and b_1 into this expression gives

$$\text{Residual Sums of Squares} = \sum_{i=1}^{n}\{Y_i - (b_0 + b_1 X_i)\}^2.$$

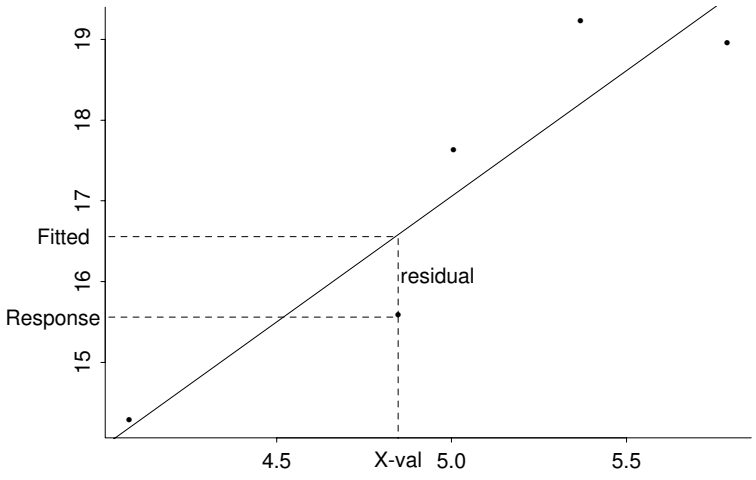
Several bits of notation are needed. Let

$$\hat{Y}_i = b_0 + b_1 X_i$$

be the **predicted** or fitted $Y-$value for an $X-$value of X_i and let $e_i = Y_i - \hat{Y}_i$. The fitted value $\hat{Y}_i$ is the value of the LS line at X_i whereas the **residual** e_i is the distance that the observed response Y_i is from the LS line. Given this notation,

$$\text{Residual Sums of Squares} = \text{Res SS } = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

Here is a picture to clarify matters:



The Residual SS, or sum of squared residuals, is *small* if each $\hat{Y}_i$ is *close to* Y_i (i.e. the line closely fits the data). It can be shown that

$$\text{Total SS in Y} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \geq \text{Res SS} \geq 0.$$

Also define

$$\text{Regression SS} = \text{Reg SS} = \text{Total SS} \ - \ \text{Res SS} = b_1 \sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}).$$

The Total SS measures the variability in the $Y-$sample. Note that

$$0 \leq \text{Regression SS} \leq \text{Total SS}.$$

The percentage of the variability in the $Y-$ sample that is **explained by the linear relationship** between Y and X is

$$R^2 = \text{coefficient of determination} = \frac{\text{Reg SS}}{\text{Total SS}}.$$

Given the definitions of the Sums of Squares, we can show $0 \leq R^2 \leq 1$ and

$$R^2 = \text{square of Pearson correlation coefficient } = r^2.$$

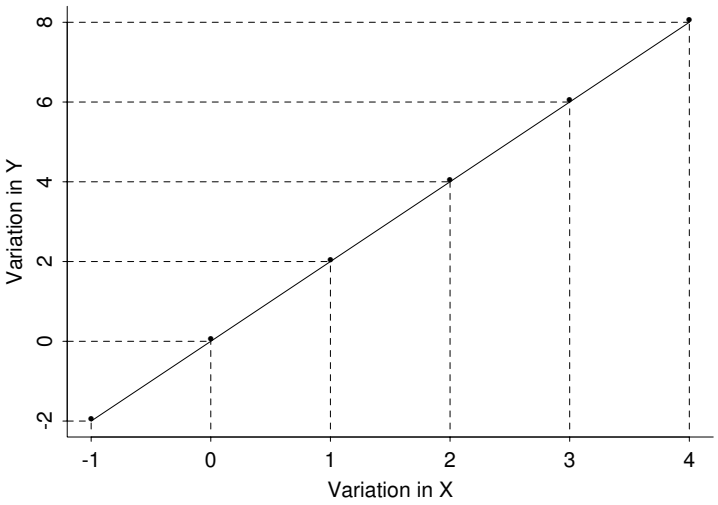
To understand the interpretation of R^2, at least in two extreme cases, note that

$$\text{Reg SS} = \text{Total SS} \quad \Leftrightarrow \quad \text{Res SS} = 0$$

$\Leftrightarrow$ all the data points fall on a straight line

$\Leftrightarrow$ all the variability in Y is explained by the linear relationship with X (which has variation)

$\Leftrightarrow$ $R^2 = 1$. (see the picture below)



Furthermore,

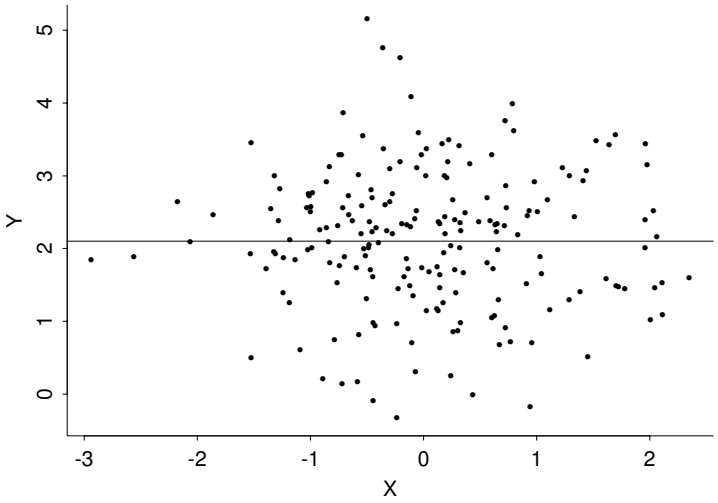
Reg SS $= 0$ $\Leftrightarrow$ Total SS $=$ Res SS

$\Leftrightarrow$ $b_1 = 0$

$\Leftrightarrow$ LS line is $\hat{Y} = \bar{Y}$

$\Leftrightarrow$ none of the variability in Y is explained by a linear relationship

$\Leftrightarrow$ $R^2 = 0$.



LS line with slope zero and intercept of average Y.

Each Sum of Squares has a corresponding *df* (degrees of freedom). The Sums of Squares and *df* are arranged in an analysis of variance (ANOVA) table:

Source	*df*	SS	MS
Regression	1		
Residual	$n-2$		
Total	$n-1$		

The Total *df* is $n-1$. The Residual *df* is n minus the number of parameters (2) estimated by the LS line. The Regression *df* is the number of predictor variables (1) in the model. A Mean Square is always equal to the Sum of Squares divided by the *df*. SW use the following notation for the Residual MS: $s^2_{Y|X} = Resid(SS)/(n-2)$.

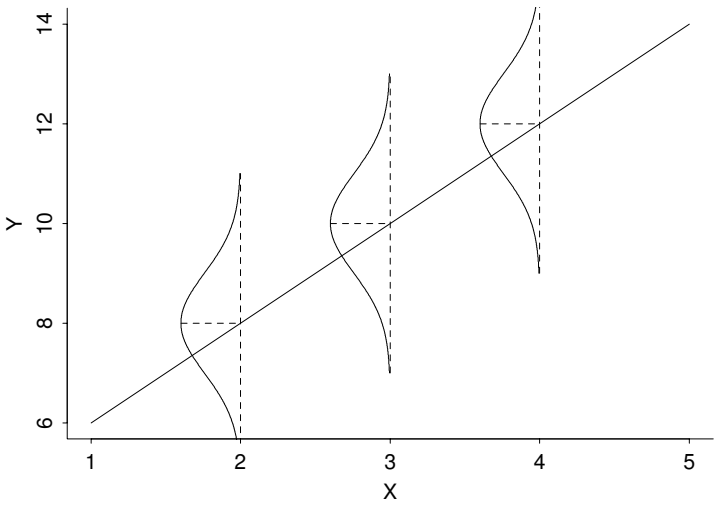
Brief Discussion of Stata Output for Blood Loss Problem

1. Identify the fitted line: Blood Loss = 552.44 - 1.30 Weight (i.e. $b_0 = 552.44$ and $b_1 = -1.30$).
2. Locate the Analysis of Variance Table. In Stata, the Regression SS is called the Model SS. More on this later.
3. Locate Parameter Estimates Table. More on this later.
4. Note that $R^2 = .5967 = (-.77247)^2 = r^2$.

2 The Linear Regression Model

The following statistical model is assumed as a means to provide error estimates for the LS line, regression coefficients, and predictions. Assume that the data (X_i, Y_i), $i = 1, ..., n$ are a sample of (X, Y) values from the population of interest, and

Visual representation of regression model with population regression line



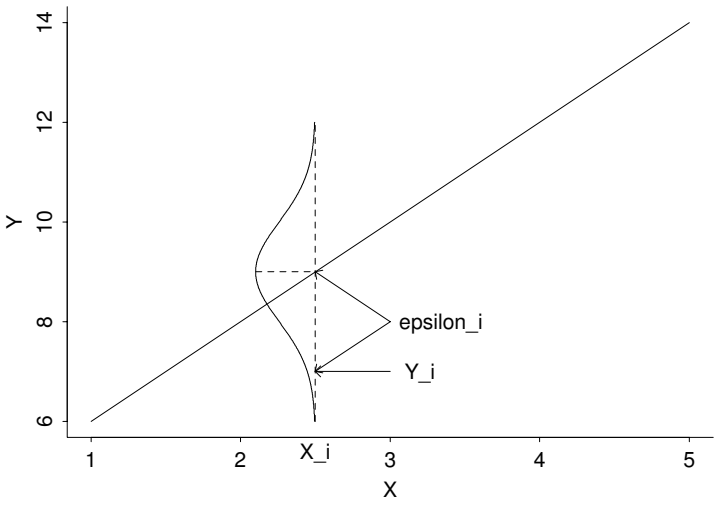
1. The mean in the population of all responses Y at a given X value (called $\mu_{Y|X}$ by SW) falls on a straight line, $\beta_0 + \beta_1 X$, called the population regression line.
2. The variation among responses Y at a given X value is the same for each X, and is denoted by $\sigma^2_{Y|X}$.
3. The population of responses Y at a given X is normally distributed.
4. The pairs (X_i, Y_i) are a random sample from the population. Alternatively, we can think that the X_is were fixed by the experimenter, and that the Y_i are random responses at the selected predictor values.

The model is usually written in the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

(i.e. Response = Mean Response + Residual), where the ϵ_is are, by virtue of assumptions 2, 3 and 4, independent normal random variables with mean 0 and variance $\sigma^2_{Y|X}$. The picture below might help you visualize this. Note that the population regression line is unknown, and is estimated from the data using the LS line.

Visual representation of population regression model notation



Back to the Data

There are three unknown population parameters in the model: β_0, β_1 and $\sigma^2_{Y|X}$. Given the data, the LS line

$$\hat{Y} = b_0 + b_1 X$$

estimates the population regression line $\beta_0 + \beta_1 X$. The LS line is our best guess about the unknown population regression line. Here b_0 estimates the intercept β_0 of the population regression line and b_1 estimates the slope β_1 of the population regression line.

The i^{th} **observed residual** $e_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = b_0 + b_1 X_i$ is the i^{th} **fitted value**, estimates the **unobservable residual** ϵ_i. (ϵ_i is unobservable because β_0 and β_1 are unknown.) The Residual MS from the ANOVA table is used to estimate $\sigma^2_{Y|X}$:

$$s^2_{Y|X} = \text{Res MS} = \frac{\text{Res SS}}{\text{Res df}} = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2}.$$

CI and Tests for β_1

A CI for β_1 is given $b_1 \pm t_{crit} SE_{b_1}$, where the standard error of b_1 under the model is

$$SE_{b_1} = \frac{s_{Y|X}}{\sqrt{\sum_i (X_i - \bar{X})^2}},$$

and where t_{crit} is the appropriate critical value for the desired CI level from a $t-$distribution with df =Res df.

To test $H_0 : \beta_1 = \beta_{1,0}$ (a given value) against $H_A : \beta_1 \neq \beta_{1,0}$, reject H_0 if $|t_s| \geq t_{crit}$, where

$$t_s = \frac{b_1 - \beta_{1,0}}{SE_{b_1}},$$

and t_{crit} is the $t-$critical value for a two-sided test, with the desired size and df =Res df. Alternatively, you can evaluate a p-value in the usual manner to make a decision about H_0.

The parameter estimates table in **Stata** gives the standard error, $t-$statistic, p-value for testing $H_0 : \beta_1 = 0$, and a 95% CI for β_1. Analogous summaries are given for the intercept, but these are typically of less interest.

Testing $\beta_1 = 0$

Assuming the mean relationship is linear, consider testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$. This test can be conducted using a t-statistic, as outlined above, or with an ANOVA $F-$test, as outlined below.

For the analysis of variance (ANOVA) F-test, compute

$$F_s = \frac{\text{Reg MS}}{\text{Res MS}}$$

and reject H_0 when F_s exceeds the critical value (for the desired size test) from an $F-$table with numerator $df = 1$ and denominator $df = n - 2$; see SW, page 654. The hypothesis of zero slope (or no relationship) is rejected when F_s is large, which happens when a significant portion of the variation in Y is explained by the linear relationship with X. **Stata** gives the $F-$statistic and p-value with the ANOVA table output.

The p-values from the $t-$test and the $F-$test are always equal. Furthermore this p-value is equal to the p-value for testing no correlation between Y and X, using the $t-$test described earlier. Is this important, obvious, or disconcerting?

A CI for the Population Regression Line

I can not overemphasize the **power** of the regression model. The model allows you to estimate the mean response at any X value in the range for which the model is reasonable, even if little or no data is observed at that location.

We estimate the mean population response among individuals with $X = X_p$

$$\mu_p = \beta_0 + \beta_1 X_p,$$

with the fitted value, or the value of the least squares line at X_p:

$$\hat{Y}_p = b_0 + b_1 X_p.$$

X_p is not necessarily one of the observed X_is in the data. To get a CI for μ_p, use $\hat{Y}_p \pm t_{crit} SE(\hat{Y}_p)$, where the standard error of $\hat{Y}_p$ is

$$SE(\hat{Y}_p) = s_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}}.$$

The $t-$critical value is identical to that used in the subsection on CI for β_1.

CI for Predictions

Suppose a future individual (i.e. someone not used to compute the LS line) has $X = X_p$. The best prediction for the response Y of this individual is the value of the least squares line at X_p:

$$\hat{Y}_p = b_0 + b_1 X_p.$$

To get a CI (prediction interval) for an individual response, use $\hat{Y}_p \pm t_{crit} SE_{pred}(\hat{Y}_p)$, where

$$SE_{pred}(\hat{Y}_p) = s_{Y|X}\sqrt{1 + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}},$$

and t_{crit} is identical to the critical value used for a CI on β_1.

For example, in the blood loss problem you may want to estimates the blood loss for an $50kg$ individual, and to get a CI for this prediction. This problem is different from computing a CI for the mean blood loss of all $50kg$ individuals!

Comments

1. The prediction interval is wider than the CI for the mean response. This is reasonable because you are less confident in predicting an individual response than the mean response for all individuals.

2. The CI for the mean response and the prediction interval for an individual response become wider as X_p moves away from $\bar{X}$. That is, you get a more sensitive CI and prediction interval for X_ps near the center of the data.

A Further Look at the Blood Loss Data using Stata

We obtain a prediction interval for an individual and confidence intervals for mean blood loss in **Stata** as follows (but note that there are a lot of ways to do this). In a separate **Stata** data set we create a variable that contains the weight values at which we would like to predict blood loss. This is done either with the `input` command or (preferably) using the data editor, an Excel-like spreadsheet utility. We illustrate the use of `input`. We desire predictions at weights of 30, 35, 40, 45, 50, 55, 60, 65, 70, and 75 kg. Examine the following **Stata** code.

```
clear
input weight
30
35
40
45
50
55
60
65
70
75
end
save weight.dta
use bloodloss
append using weight
regress loss weight
predict loss_hat,xb
predict se_line, stdp
predict se_pred, stdf
```

```
generate lci=loss_hat-invttail(6,0.025)*se_line
generate uci=loss_hat+invttail(6,0.025)*se_line
generate lpi=loss_hat-invttail(6,0.025)*se_pred
generate upi=loss_hat+invttail(6,0.025)*se_pred
graph twoway (scatter loss weight)  (line loss_hat weight) ///
            (line lci weight,sort)(line uci weight,sort)  ///
            (line lpi weight,sort)(line upi weight, sort) ///
            , title(Blood Loss Data) subtitle(CI for Line and Prediction Int.)
```

The above commands create a new data set called `weight`, append those weight values to the bloodloss data set (leaving values of weight and time missing) perform regression using only the original data set (cases with missing values of X or Y are discarded), and then save the predicted values $\hat{Y}_p$ (fitted values on the regression line) for each value of the variable `weight` as well as the standard errors for the fitted line, $SE(\hat{Y}_p)$, and standard errors for prediction, $SE_{pred}(\hat{Y}_p)$. The confidence interval for the line and prediction interval is computed and plotted. After this program is run (from a do-file) the data set looks as follows.

```
. list,clean
```

	weight	time	loss	loss_hat	se_line	se_pred	lci	uci	lpi	upi
1.	44.3	105	503	494.8375	4.46279	12.48698	483.9175	505.7576	464.283	525.392
2.	40.6	80	490	499.6487	5.295104	12.80805	486.6921	512.6054	468.3086	530.9889
3.	69	86	471	462.7195	9.965039	15.33982	438.3359	487.103	425.1843	500.2546
4.	43.7	112	505	495.6177	4.569383	12.52547	484.4369	506.7986	464.969	526.2665
5.	50.3	109	482	487.0356	4.222673	12.40319	476.703	497.3681	456.686	517.3851
6.	50.2	100	490	487.1656	4.213473	12.40006	476.8556	497.4756	456.8237	517.5074
7.	35.4	96	513	506.4104	6.947425	13.57479	489.4107	523.4102	473.1941	539.6267
8.	52.2	120	464	484.5649	4.475415	12.4915	473.614	495.5159	453.9994	515.1306
9.	30	.	.	513.4322	8.954061	14.70317	491.5224	535.342	477.4548	549.4095
10.	35	.	.	506.9306	7.088681	13.64762	489.5852	524.2759	473.536	540.3251
11.	40	.	.	500.4289	5.463197	12.87846	487.061	513.7969	468.9165	531.9413
12.	45	.	.	493.9273	4.355063	12.44888	483.2708	504.5838	463.466	524.3886
13.	50	.	.	487.4257	4.196375	12.39426	477.1575	497.6938	457.098	517.7533
14.	55	.	.	480.924	5.076955	12.71942	468.5012	493.3469	449.8008	512.0474
15.	60	.	.	474.4224	6.592747	13.39673	458.2905	490.5543	441.6418	507.203
16.	65	.	.	467.9207	8.406906	14.37652	447.3498	488.4917	432.7427	503.0988
17.	70	.	.	461.4191	10.36392	15.60189	436.0595	486.7787	423.2427	499.5956
18.	75	.	.	454.9175	12.39631	17.01989	424.5848	485.2502	413.2713	496.5636

Given the model Blood Loss = $\beta_0 + \beta_1$ Weight $+\epsilon$:

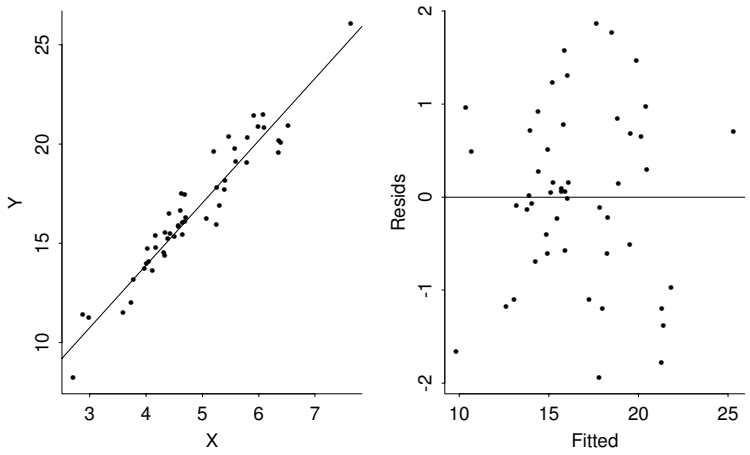
- The LS line is: Predicted Blood Loss = 552.442 - 1.30 Weight.
- The R^2 is .597 (i.e. 59.7%); see Lecture 1.
- The $F-$statistic for testing $H_0 : \beta_1 = 0$ is $F_{obs} = 8.88$ with a $p - value = .0247$. The Error MS is $s^2_{Y|X} = 136.008$; see ANOVA table.
- The Parameter Estimates table gives b_0 and b_1, their standard errors, $t-$statistics and p-values for testing $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$. The $t-$test and $F-$ test p-values for testing that the slope is zero are identical.
- Prediction and CI: The estimated average blood loss for all 50kg patients is $552.442 - 1.30033 * 50 = 487.43$. We are 95% confident that the mean blood loss of all 50kg patients is between (approximately) 477 and 498 ml. A 95% prediction interval for the blood loss of a single 50 kg person is less precise (about 457 to 518 ml).

As a summary we might say that weight is important for explaining the variation in blood loss. In particular, the estimated slope of the least squares line (Predicted Blood loss = 552.442 - 1.30 Weight) is significantly different from zero (p-value = .0247), with weight explaining approximately 60% (59.7%) of the variation in blood loss for this sample of 8 thyroid operation patients.

Checking the regression model

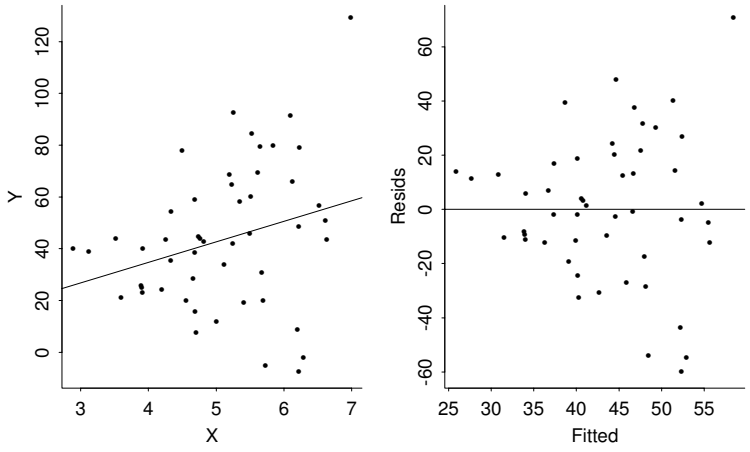
A regression analysis is never complete until the assumptions of the model have been checked. In addition, you need to evaluate whether individual observations, or groups of observations, are unduly influencing the analysis. A first step in any analysis is to plot the data. The plot provides information on the linearity and constant variance assumption. For example, the data plot below shows a linear relationship with roughly constant variance.

In addition to plotting the data, a variety of methods for checking models are based on plots of the residuals, $e_i = Y_i - \hat{Y}_i$ (i.e. Observed $-$ Fitted). The command `rvpplot` in **Stata** plots the e_i against the predictor values X_i. Alternatively (and equivalently for simple linear regression), the command `rvfplot` plots e_i against the fitted values $\hat{Y}_i$, as illustrated in the plots below. Regardless of which you use, the residual plot should exhibit no systematic dependence of the sign or the magnitude of the residuals on the fitted values.

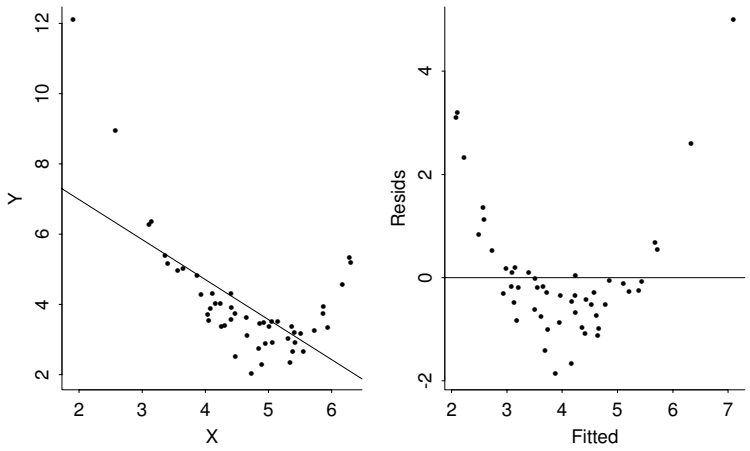


The real power of this plot (e_i against $\hat{Y}_i$) is with multiple predictor problems (multiple regression). For simple linear regression, the information in this plot is similar to the information in the original data plot, except that the residual plot eliminates the effect of the trend on your perceptions of model adequacy.

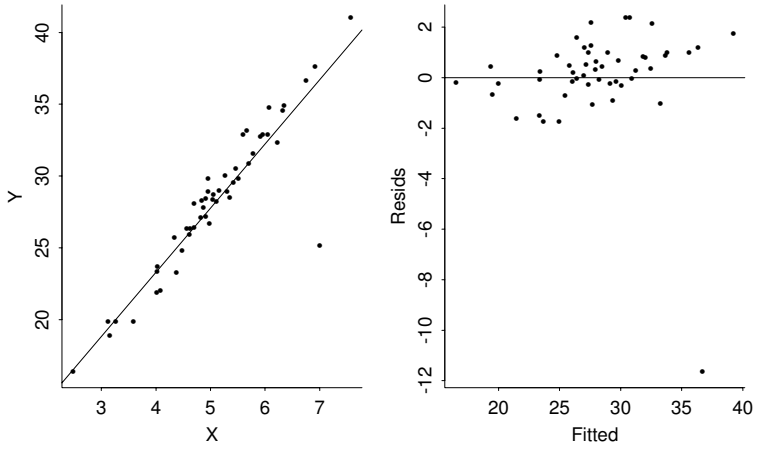
The following plots show how inadequacies in the data plot appear in a residual plot.



The first plot (above) shows a roughly linear relationship between Y and X with non-constant variance. The residual plot shows a megaphone shape rather than the ideal horizontal band. A possible remedy is a **weighted least squares** analysis to handle the non-constant variance, or to transform Y to stabilize the variance. Transforming the data may destroy the linearity.



The next plot (above) shows a nonlinear relationship between Y and X. The residual plot shows a systematic dependence of the sign of the residual on the fitted value. A possible remedy is to transform the data.



The last plot (above) shows an outlier. This point has a large residual. A sensible approach is to refit the model after deleting the case and see if any conclusions change.

Checking Normality

The normality assumption can be evaluated with a boxplot or a normal quantile plot of the residuals (**Stata** command `graph` [residuals], `box` and `qnorm`). A formal test of normality using the residuals

can be based on the Wilk-Shapiro test (discussed in last semester's lab) using the **Stata** command `swilk`.

Checking Independence

Diagnosing dependence among observations usually requires some understanding of the mechanism that generated the data. There are a variety of graphical and inferential tools for checking independence for data collected over time (called a time series). The easiest thing to do is plot the r_i against time index and look for any suggestive patterns.

Outliers

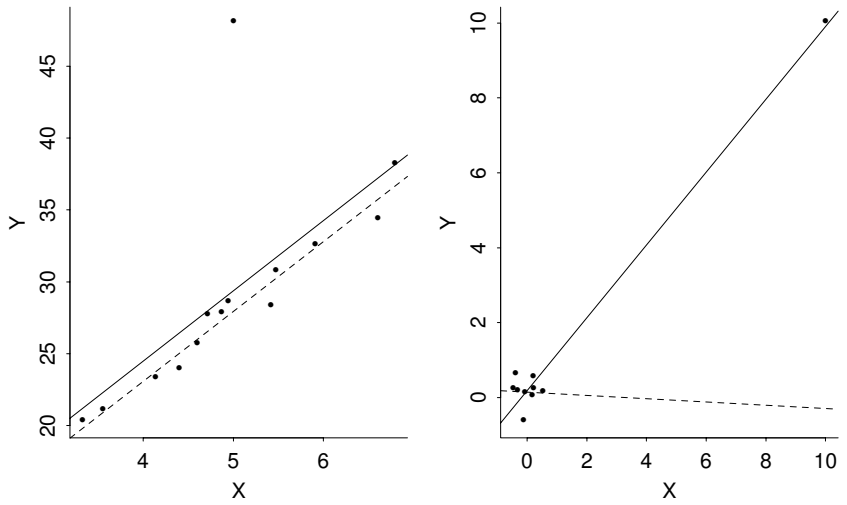
Outliers are observations that are poorly fitted by the regression model. The response for an outlier is far from the fitted line, so outliers have large positive or negative values of the residual e_i.

What do you do with outliers? Outliers may be due to incorrect recordings of the data or failure of the measuring device, or indications or a change in the mean or variance structure for one or more cases. Incorrect recordings should be fixed if possible, but otherwise deleted from the analysis.

Routine deletion of outliers from the analysis is not recommended. This practice can have a dramatic effect on the fit of the model and the perceived precision of parameter estimates and predictions. Analysts who routinely omit outliers without cause tend to overstate the significance of their findings and get a false sense of precision in their estimates and predictions. At the very least, a data analyst should repeat the analysis with and without the outliers to see whether any substantive conclusions are changed.

Influential observations

Certain data points can play a very important role in determining the position of the LS line. These data points may or may not be outliers. For example, the observation with $Y > 45$ in the first plot below is an outlier relative to the LS fit. The extreme observation in the second plot has a very small e_i. Both points are highly **influential observations** - the LS line changes dramatically when these observations are deleted. The influential observation in the second plot is not an outlier because its presence in the analysis determines that the LS line will essentially pass through it! In these plots the solid line is the LS line from the full data set, whereas the dashed line is the LS line after omitting the unusual point.



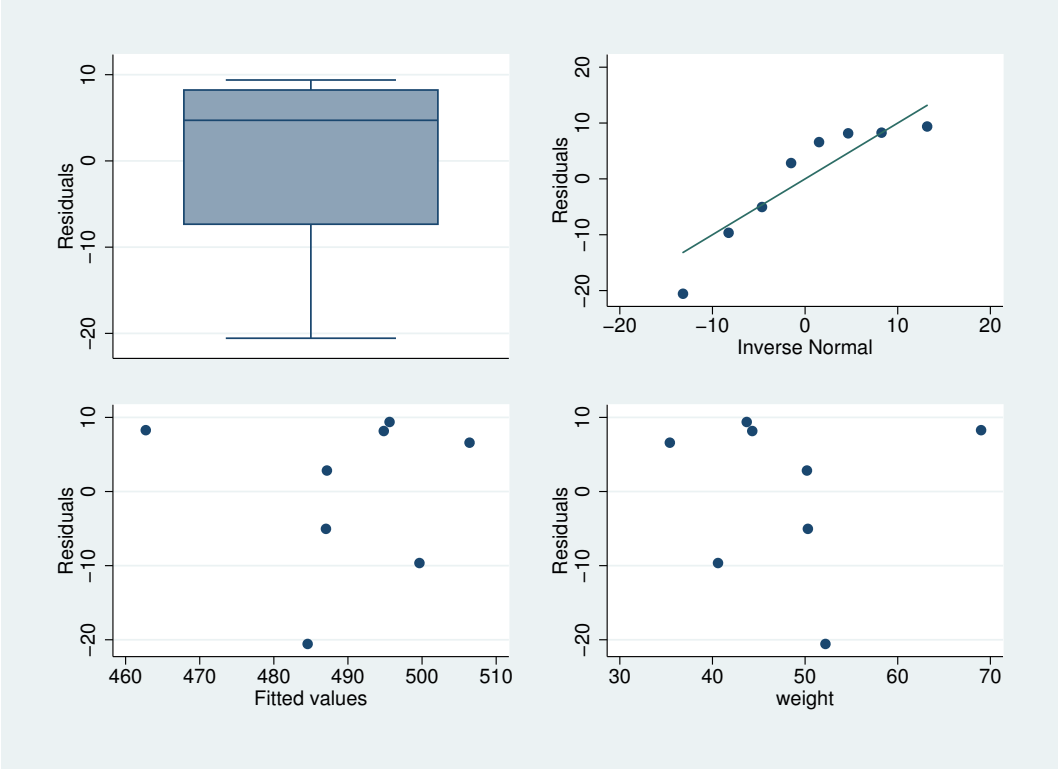
A standard measure of the influence that individual cases have on the LS line is called **Cook's Distance**, which is available as `predict cooksd, cooksd` for example. For simple linear regression most influential cases can be easily spotted by carefully looking at the data plot. If you identify cases that you suspect might be influential, you should hold them out (individually) and see if any important conclusions change. If so, you need to think hard about whether the cases should be included or excluded from the analysis. We will obtain and interpret Cook's distances later.

A Final Look at the Blood Loss Data

We create various diagnostic plots and perform the Shapiro-Wilk test of normality on the residuals using the **Stata** commands

```
use bloodloss
regress loss weight
predict res, r
swilk r
graph box r, saving(boxplot)
qnorm r, saving(probplot)
rvfplot, saving(respredplot)
rvpplot weight, saving(resweightplot)
graph combine boxplot.gph probplot.gph respredplot.gph resweightplot.gph, saving(all)
```

Residual plots for the blood loss problem follow. Do we see any marked problems with influential cases, outliers, or non-normality? Also, go back in the notes and look at the data plot.



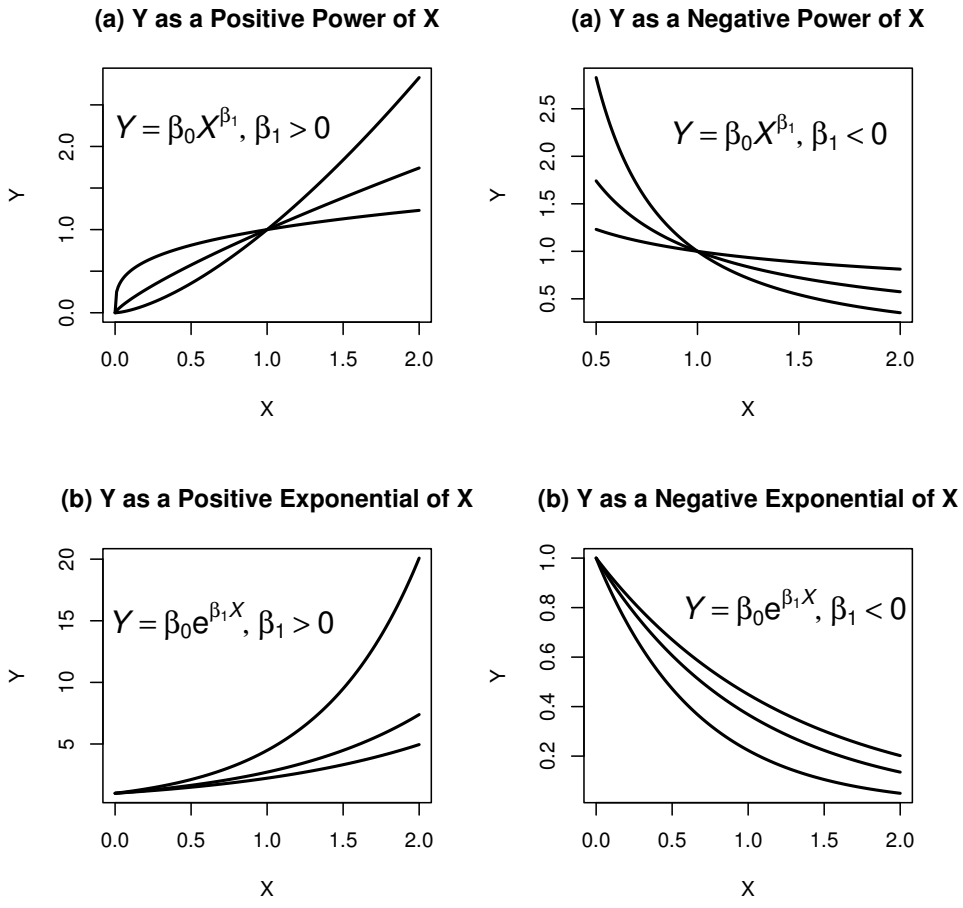
The results of the Shapiro-Wilk normality test on the residuals:

```
                  Shapiro-Wilk W test for normal data
    Variable |     Obs        W           V          z      Prob>z
-------------+--------------------------------------------------
         res |       8    0.84852       2.110      1.328   0.09204
```

3 Transformations in Regression

Simple linear regression is appropriate when the scatterplot of Y against X show a linear trend. In many problems, non-linear relationships are evident in data plots. Linear regression techniques can still be used to model the dependence between Y and X, provided the data can be transformed to a scale where the relationship is roughly linear. In the ideal world, theory will suggest an appropriate transformation. In the absence of theory one usually resorts to empirical model building. Polynomial models are another method for handling nonlinear relationships.

I will suggest transformations that you can try if the **trend** in your scatterplot has one of the following functional forms. The responses are assumed to be non-negative (in some cases strictly positive) in all cases.



The functional relationship between Y and X in (a) is given by $Y = \beta_0 X^{\beta_1}$, that is Y is related to a power of X, where the power is typically unknown. For the left plot, $\beta_1 > 0$ whereas $\beta_1 < 0$ for the plot on the right. For either situation, the logarithm of Y is linearly related to the logarithm of X (regardless of the base):

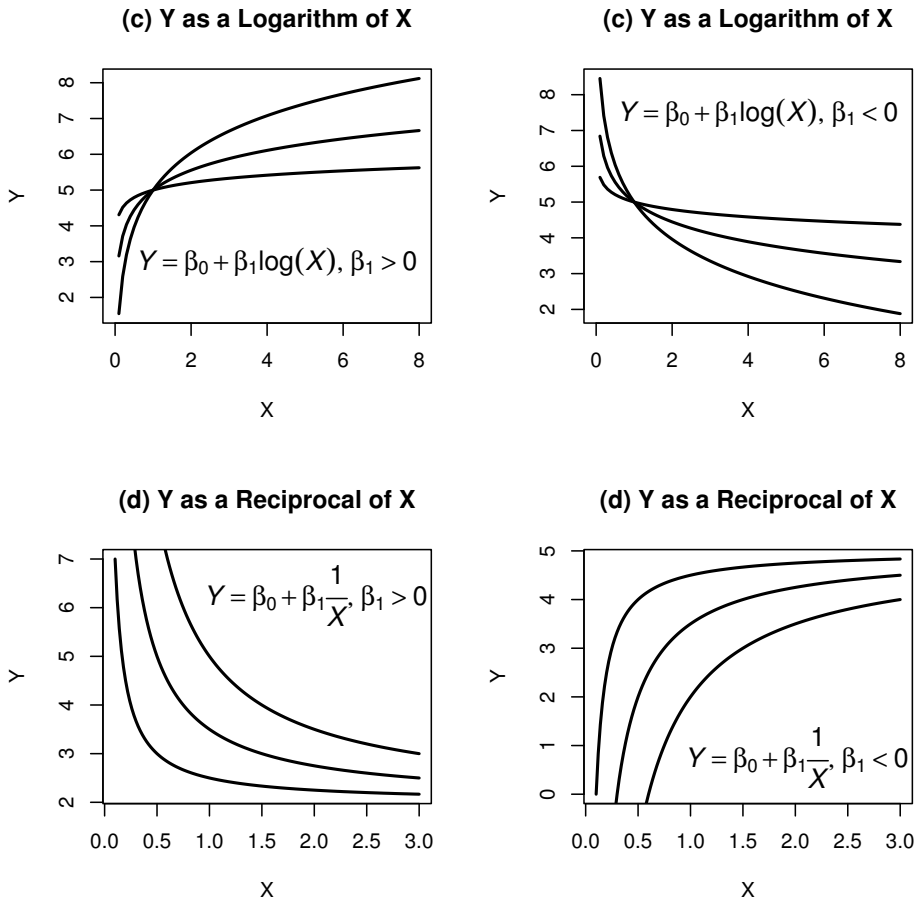
$$\log(Y) = \log(\beta_0) + \beta_1 \log(X).$$

You should consider a simple linear regression of $Y' = \log(Y)$ on $X' = \log(X)$.

The functional relationship between Y and X in (b) is given by $Y = \beta_0 \exp(\beta_1 X)$, that is Y is an exponential function of X. For the plot on the left, $\beta_1 > 0$ whereas $\beta_1 < 0$ for the plot on the right. In either situation, the natural logarithm of Y is linearly related to X:

$$\log_e(Y) = \log_e(\beta_0) + \beta_1 X.$$

You should consider a simple linear regression of $Y' = \log_e(Y)$ on X. Actually, the base of the logarithm is not important here either.



The functional relationship between Y and X in (c) is given by $Y = \beta_0 + \beta_1 \log(X)$, that is Y is an logarithmic function of X. For the plot on the left, $\beta_1 > 0$ whereas $\beta_1 < 0$ for the plot on the right. In each situation, consider a simple linear regression of Y on $X' = \log(X)$.

The functional relationship between Y and X in (d) is

$$Y = \beta_0 + \beta_1 \frac{1}{X}.$$

Hence, consider a simple linear regression of Y on $X' = 1/X$. Note that each plot in (d) has a horizontal asymptote of β_0.

In most problems, the trend or signal will be buried in a considerable amount of noise, or variability, so the best transformation may not be apparent. If two or more transformations are suggested try all of them and see which is best - look at diagnostics from the various fits rather than (meaningless) summaries such as R^2. In situations where a logarithmic transformation is suggested, you might try a square root transformation as well. It often does make a considerable difference in the quality of the fit whether you transform Y only, X only, or both. There are more organized schemes for choosing transformations, but this sort of trial and error is the most common practice. Note that the functional forms (a) - (d), while probably the most frequently encountered, are not at all the only ones used.

The need to transform is sometimes much more apparent in a plot of the residuals against the predicted values from a "linear fit" of the **original data** because you tend not to perceive subtle deviations from linearity. The *Wind Speed* example below illustrates this.

Transformations also can help to control influential values and outliers (recall that an outlying X-value can cause that point to exert undue influence on the fit). Functions such as *log* have the effect of bringing outlying values much closer to the rest of the data. The *Brain Weights vs. Body Weights* example below illustrates this. When I see a variable with a highly skewed distribution, I usually try transforming it to make it more symmetric. This can work both ways, of course - you can make a nice symmetrically distributed variable skewed by transforming it.

Computing Predictions

Transforming the response to a new scale causes no difficulties if you wish to make predictions on the original scale. For example, suppose you fit a linear regression of $\log_e(Y)$ on X. The fitted values satisfy

$$\widehat{\log_e(Y)} = b_0 + b_1 X.$$

The predicted response Y_p for an individual with $X = X_p$ is obtained by first getting the predicted value for $\log_e(Y_p)$:

$$\widehat{\log_e(Y_p)} = b_0 + b_1 X_p.$$

Our best guess for Y_p is obtained by exponentiating our prediction for $\log_e(Y_p)$:

$$\hat{Y}_p = \exp(\widehat{\log_e(Y_p)}) = \exp(b_0 + b_1 X_p).$$

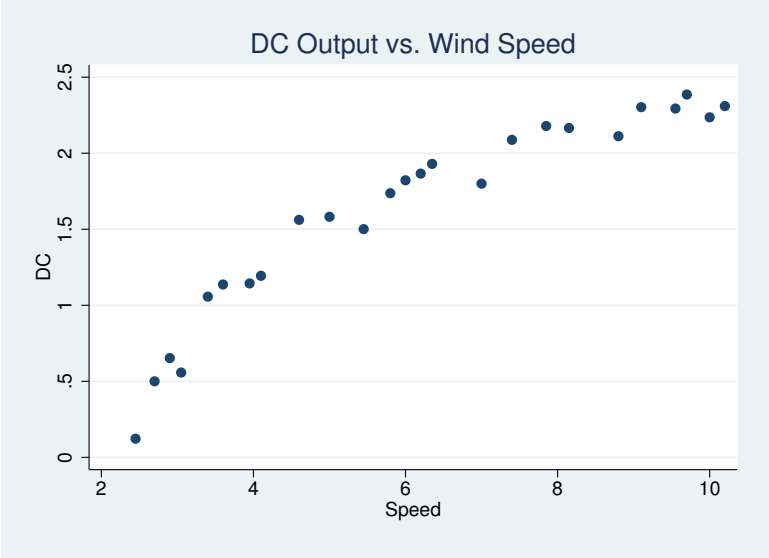
The same idea can be used to get prediction intervals for Y_p from a prediction interval for $\log_e(Y_p)$ (just transform the lower and upper confidence limits).

Other transformations on Y are handled analogously. For example, how do you predict Y using a simple linear regression with $1/Y$ as the selected response?

Example of Transformations: Wind Speed Data

A research engineer is investigating the use of a windmill to generate electricity. She has collected data on the DC output from the windmill and the corresponding wind velocity. She wants to develop a model that explains the dependence of the DC output on wind velocity. The data were read into **Stata** and plotted.

```
. list speed dc,clean
       speed      dc
  1.       5   1.582
  2.       6   1.822
  3.     3.4   1.057
  4.     2.7      .5
  5.      10   2.236
  6.     9.7   2.386
  7.    9.55   2.294
  8.    3.05    .558
  9.    8.15   2.166
 10.     6.2   1.866
 11.     2.9    .653
 12.    6.35    1.93
 13.     4.6   1.562
 14.     5.8   1.737
 15.     7.4   2.088
 16.     3.6   1.137
 17.    7.85   2.179
 18.     8.8   2.112
 19.       7     1.8
 20.    5.45   1.501
 21.     9.1   2.303
 22.    10.2    2.31
 23.     4.1   1.194
 24.    3.95   1.144
 25.    2.45    .123
```

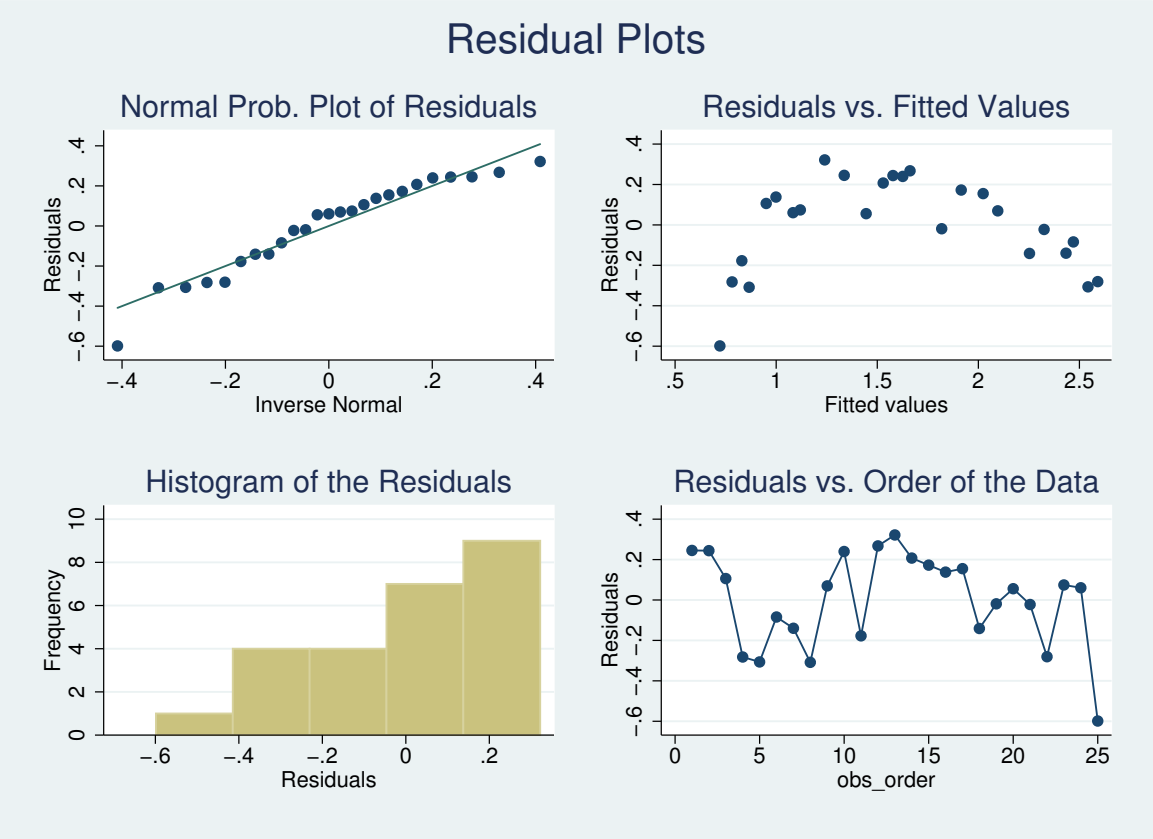



```
. regress dc speed
      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) =  160.26
       Model |  8.92961408     1  8.92961408           Prob > F      =  0.0000
    Residual |  1.28157328    23  .055720577           R-squared     =  0.8745
-------------+------------------------------           Adj R-squared =  0.8690
       Total |  10.2111874    24   .42546614           Root MSE      =  .23605
------------------------------------------------------------------------------
          dc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       speed |   .2411489   .0190492    12.66   0.000     .2017426    .2805551
       _cons |   .1308752   .1259894     1.04   0.310    -.1297537    .3915041
------------------------------------------------------------------------------
```

The data plot shows a strong linear trend, but the relationship is nonlinear. If I ignore the nonlinearity and fit a simple linear regression model, I get

$$\text{Predicted DC Output} = .1309 + .2411 \text{ Wind Speed}.$$

Although the R^2 from this fit is high, $R^2 = .875$, I am unhappy with the fit of the model. The plot of the residuals against the fitted values clearly points out the inadequacy:



The `rvfplot` shows that the linear regression systematically underestimates the DC output for wind speeds in the middle, and overestimates the DC output for low and high wind speeds. This model is not acceptable for making predictions - one can and should do better!

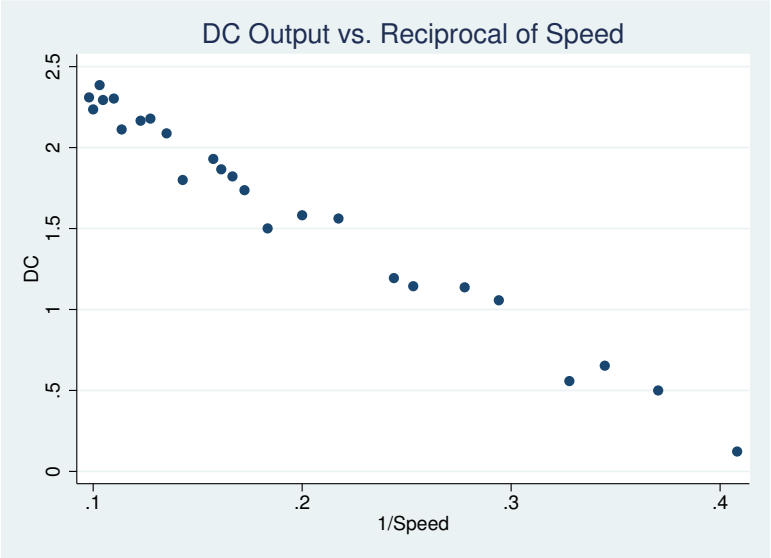
The original data plot indicates that DC output approaches an upper limit of about 2.5 amps as the wind speed increases. Given this fact, and the trend in the plot, I decided to use the inverse of wind speed as a predictor of DC output. Another reasonable first step would be a logarithmic transformation of wind speed but this function steadily increases without approaching a finite limit.

Aside: The above plot is not the same as in the previous notes or in the lab. I decided to illustrate further the flexibility of Stata and the power of `do files`. We obtained exactly those four plots in Minitab if we requested the 4-in-1 plots in regression. You might want to replace the histogram with a boxplot – the modification is simple. The `do file` statements to produce the plot after running the regression command are:

```
predict residual, r
quietly qnorm residual, saving(probplot, replace) nodraw ///
    title(Normal Prob. Plot of Residuals)
quietly rvfplot, saving(respredplot, replace) nodraw ///
    title(Residuals vs. Fitted Values)
quietly hist residual, freq saving(hist, replace) nodraw ///
    title(Histogram of the Residuals)
generate obs_order = _n
quietly twoway connect residual obs_order, saving(obs_order, replace) ///
    nodraw title(Residuals vs. Order of the Data)
drop obs_order
graph combine probplot.gph respredplot.gph hist.gph obs_order.gph, ///
      title(Residual Plots)
```

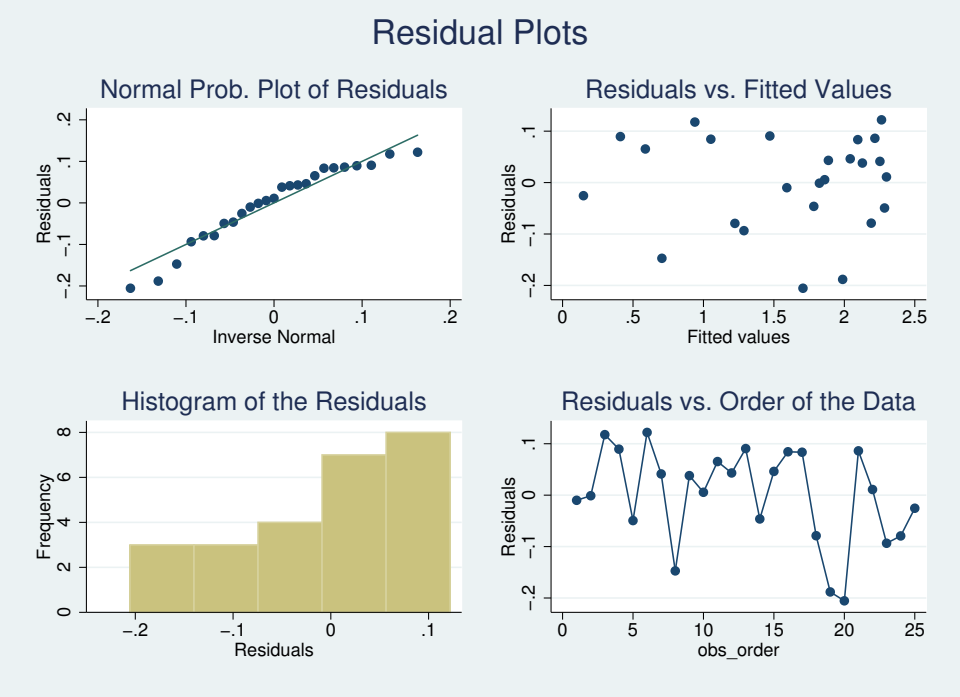
This program will fail if the variable `residual` exists before you run it (that can be fixed).

A plot of DC output against one over the wind speed is fairly linear:



This suggests that a simple linear regression fit on this scale is appropriate. Note that DC output is a decreasing function of one over the wind speed.

```
. regress dc speed_inv
      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) = 1128.43
       Model |  10.0072178     1  10.0072178           Prob > F      =  0.0000
    Residual |  .203969527    23   .00886824           R-squared     =  0.9800
-------------+------------------------------           Adj R-squared =  0.9792
       Total |  10.2111874    24   .42546614           Root MSE      =  .09417
------------------------------------------------------------------------------
          dc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   speed_inv |  -6.934547   .2064335   -33.59   0.000    -7.361588   -6.507507
       _cons |    2.97886   .0449023    66.34   0.000     2.885973    3.071748
------------------------------------------------------------------------------
```

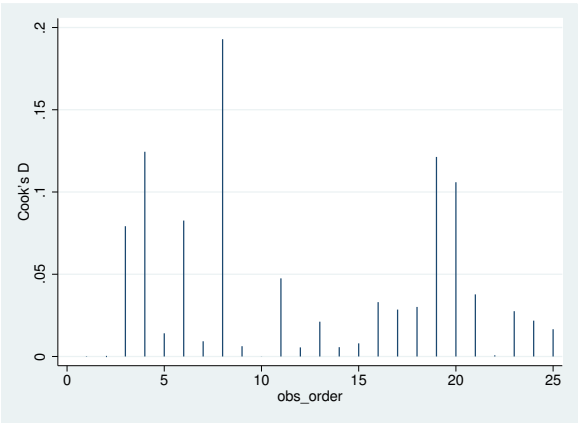


The LS regression line is

$$\text{Predicted DC output} = 2.9789 - 6.9345\frac{1}{\text{Wind speed}}.$$

The residual plots show left skewness, but no serious outliers. The Shapiro-Wilk test has a p-value of 0.08. The transformation appears to work well, although if I tried harder I might be able to symmetrize the residuals a little better (I would start by transforming Y instead of X). I don't think it is worth the trouble here, though. It is fairly clear by examining the scatter plot (the one corresponding to the actual regression we did!) that there are no highly influential points here. Still, we really should check the Cook's D values as a routine matter. Since 1 is a common cutoff for Cook's D, and no values stand out much, we have little to be concerned over.

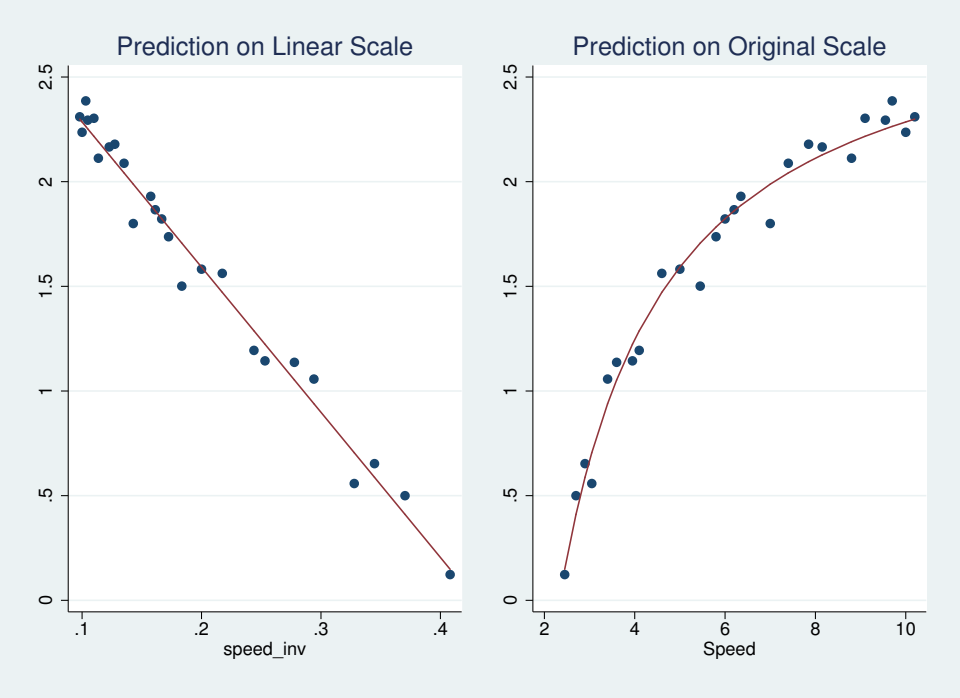
```
. predict cooksd,cooksd
. gene obs_order = _n
. twoway spike cooksd obs_order
```



All our theory and modelling applies in the linear scale (the transformed problem where we fit output to 1/speed). We really want to see how well things appear to work in the original scale, though. The following statements accomplish that.

```
. regress dc speed_inv
. predict pred_dc,xb
. twoway (scatter dc speed_inv) (line pred_dc speed_inv,sort),legend(off)
> title(Prediction on Linear Scale) saving(l,replace)
. twoway (scatter dc speed) (line pred_dc speed,sort),legend(off)
> title(Prediction on Original Scale) saving(o,replace)
. graph combine l.gph o.gph
```

We would put confidence and prediction bands on the plot in a similar manner. How would we predict output (with a prediction interval) for a wind speed of 15?



Brain Weights and Body Weights of Mammals

The data below are the average brain weight (g) and body weights (kg) for 62 species of mammals. We are interested in developing a model for predicting brain weight from body weight.

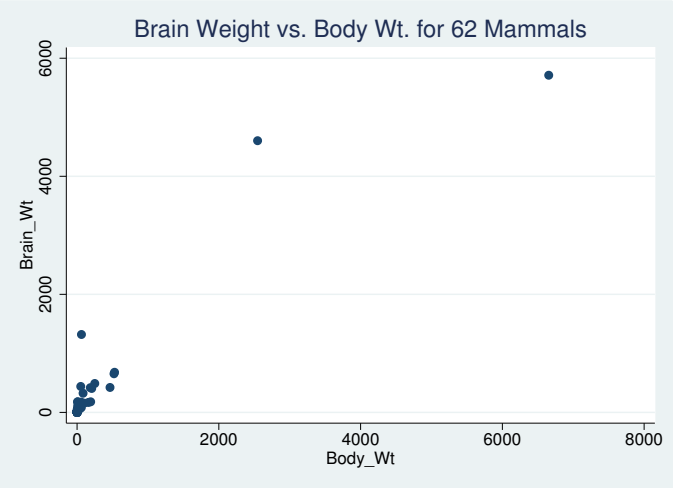
```
. list,clean
```

	species	body_wt	brain_wt
1.	Arctic fox	3.385	44.5
2.	Owl monkey	.48	15.499
3.	Mountain beaver	1.35	8.1
4.	Cow	465	423
5.	Gray wolf	36.33	119.5
6.	Goat	27.66	115
7.	Roe deer	14.83	98.2
8.	Guinea pig	1.04	5.5
9.	Vervet	4.19	58
10.	Chinchilla	.425	6.4
11.	Ground squirrel	.101	4
12.	Arctic ground squirrel	.92	5.7
13.	Africa giant poached rat	1	6.6
14.	Lesser short-tailed shrew	.005	.14
15.	Star-nosed mole	.06	1
16.	Nine-banded armadillo	3.5	10.8
17.	Tree hyrax	2	12.3
18.	N. American opussum	1.7	6.3
19.	Asian elephant	2547	4603
20.	Big brown bat	.023	.3
21.	Donkey	187.1	419
22.	Horse	521	655
23.	European hedgehog	.785	3.5
24.	Patas monkey	10	115
25.	Cat	3.3	25.6
26.	Galago	.2	5
27.	Genet	1.41	17.5
28.	Giraffe	529	680
29.	Gorilla	207	406
30.	Gray seal	85	325
31.	Rock hyrax	.75	12.3
32.	Human	62	1320
33.	African elephant	6654	5712
34.	Water opussum	3.5	3.9
35.	Rhesus monkey	6.8	179
36.	Kangaroo	35	56
37.	Yellow-bellied marmot	4.05	17
38.	Golden hamster	.12	1
39.	Mouse	.023	.4
40.	Little brown bat	.01	.25
41.	Slow loris	1.4	12.5
42.	Okapi	250.01	490
43.	Rabbit	2.5	12.1
44.	Sheep	55.5	175
45.	Jaguar	100	157
46.	Chimpanzee	52.16	440
47.	Baboon	10.55	179.5
48.	Desert hedgehog	.55	2.4
49.	Giant armadillo	60	81
50.	Rock hyrax	3.6	21
51.	Raccoon	4.288	39.2
52.	Rat	.28	1.9
53.	Eastern American mole	.075	1.2
54.	Mole rat	.122	3
55.	Musk shrew	.048	.33
56.	Pig	192	180
57.	Echidna	3	25
58.	Brazilian tapir	160	169
59.	Tenrec	.9	2.6
60.	Phalanger	1.62	11.4

```
61.                 Tree shrew      .104        2.5
62.                    Red fox    4.235       50.4
```

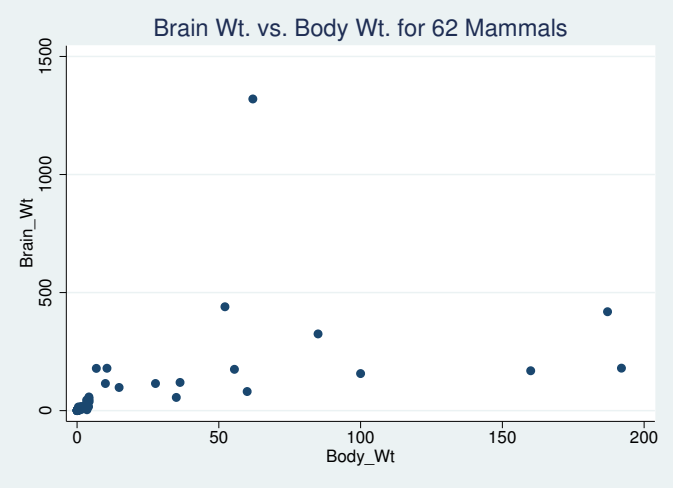
A plot of the brain weights against the body weights is non-informative because many species have very small brain weights and body weights compared to the elephants:

```
. scatter br bo,tit(Brain Weight vs. Body Wt. for 62 Mammals)
```



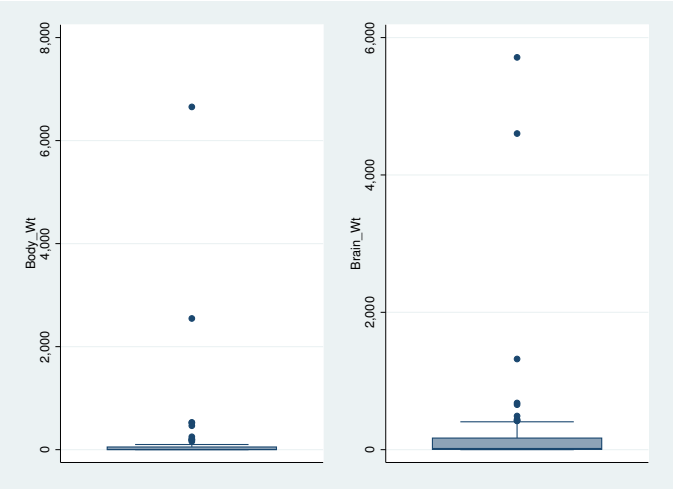
If we momentarily hold out the species with body weights exceeding $200kg$ or brain weights exceeding $200g$, and replot the data, we see that the brain weight of mammals typically increases with the body weight, but the relationship is nonlinear:

```
. scatter br bo if(bo<=200),tit(Brain Wt vs. Body Wt. for 62 Mammals)
```



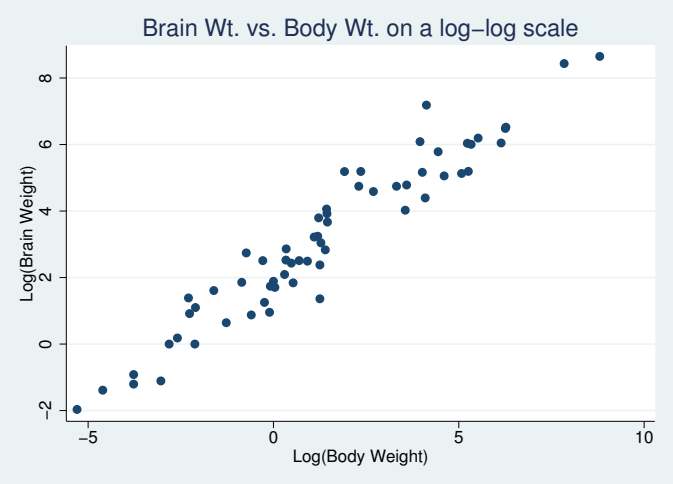
The trend suggests transforming both variables to a logarithmic scale to linearize the relationship between brain weight and body weight. It does not matter which base logarithm you choose. The relationship is no more linear with one base than another. I will use natural logarithms. What is even more compelling about the log transform here is the extreme *right* skewness of both variables – logs pull extremely large values down much more than more modest values, so they tend to symmetrize such data (and regression works much better when both variables have reasonably symmetric distributions).

```
. graph box bod,name(bodbox)
. graph box br,name(brbox)
. graph combine bodbox brbox
```



The plot of $\log_e$(brain weight) against $\log_e$(body weight) is fairly linear:

```
. gene lbod=log(body_wt)
. gene lbr = log(brain_wt)
. scatter lbr lbod,title(Brain Wt. vs. Body Wt. on a log-log scale) xti(Log(Bod
> y Weight)) yti(Log(Brain Weight))
```



At this point I considered fitting the model:

$$\log_e(\text{brain weight}) = \beta_0 + \beta_1 \log_e(\text{body weight}) + \epsilon.$$

Summary information from fitting this model:

```
. regre lbr lbo

      Source |       SS       df       MS              Number of obs =      62
-------------+------------------------------           F(  1,    60) =  697.42
       Model |  336.188164     1  336.188164            Prob > F      =  0.0000
    Residual |  28.9225677    60  .482042795            R-squared     =  0.9208
-------------+------------------------------           Adj R-squared =  0.9195
       Total |  365.110732    61  5.98542184            Root MSE      =  .69429

------------------------------------------------------------------------------
         lbr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lbod |   .7516859   .0284635    26.41   0.000     .6947505    .8086214
       _cons |   2.134787   .0960432    22.23   0.000     1.942672    2.326902
------------------------------------------------------------------------------
```

The fitted relationship:

$$\text{Predicted } \log_e(\text{brain weight}) = 2.135 + 0.752 \log_e(\text{body weight}),$$

explains about 92% of the variation in $\log_e$(brain weight). The $t-$test for $H_0 : \beta_1 = 0$ is highly significant ($p - value = 0$ to three decimal places). This summary information combined with the data plot indicates that there is a strong linear relationship between $\log_e$(brain weight) and $\log_e$(body weight), with the average $\log_e$(brain weight) increasing as $\log_e$(body weight) increases.

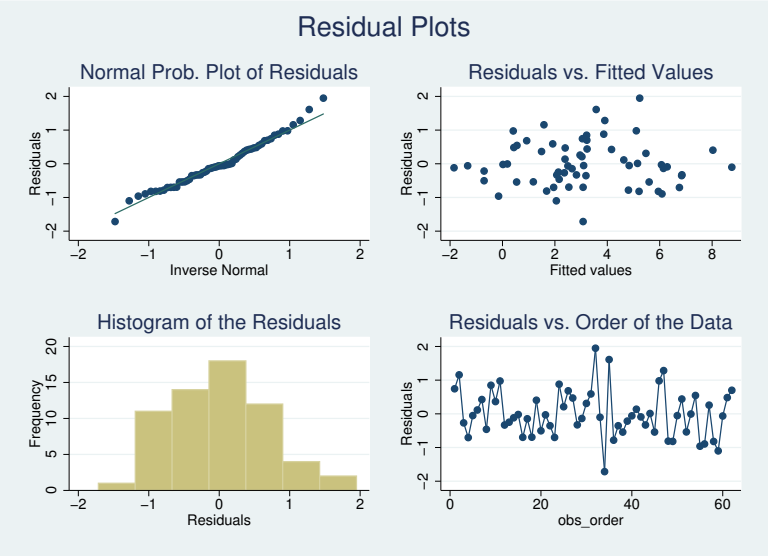
To predict brain weights, use the inverse transformation

$$\text{Predicted brain weight} = \exp\{\text{Predicted } \log_e(\text{brain weight})\}$$

or

$$\begin{aligned} \text{Predicted brain weight} &= \exp\{2.135 + 0.752 \log_e(\text{body weight})\} \\ &= \exp(2.135) * \text{body weight}^{0.752} \\ &= 8.457 * \text{body weight}^{0.752}. \end{aligned}$$

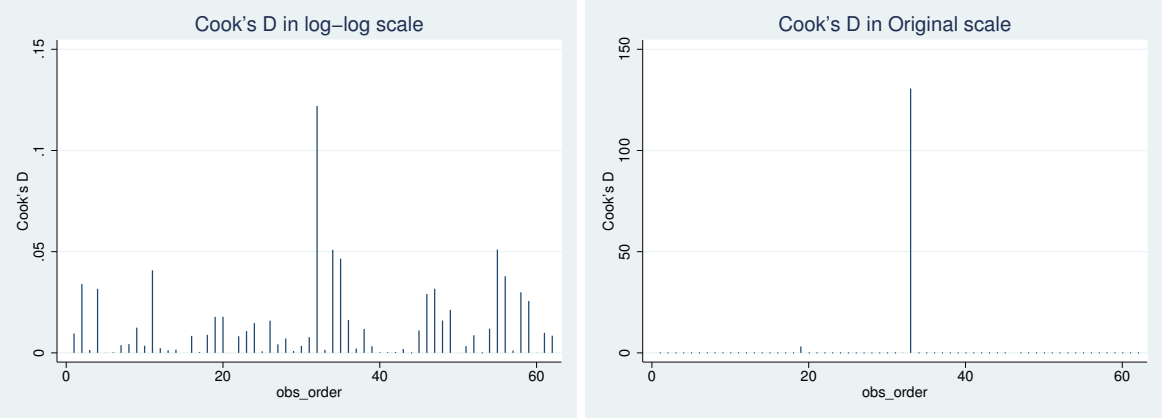
These conclusions are tentative, subject to a careful residual analysis. Residual plots do not suggest any serious deficiencies with the model, but do highlight one or more poorly fitted species:



Can anyone guess what species these may be, and what further analyses might be reasonable? The largest and smallest residuals belong to observations 32 and 34 respectively (obtained from simply entering the data editor). Note that a normal probability (or $Q-Q$) plot of the residuals is reasonably straight and the Shapiro-Wilk test of normality indicates no gross departures from normality:

```
                    Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+-------------------------------------------------
         res |     62    0.98268      0.967     -0.073  0.52927
```

Cook's D does not show any particular problems (until the value approaches 1, most data analysts do not worry much about it). Compare it to the value in the original scale where the distribution of both variables was so skewed.



Usually it is worth plotting the fitted values back on the original scale as we did for the wind speed data. That would not be very useful here since the original scale obscures most of the data.

4 Introduction to Multiple Linear Regression

In **multiple linear regression**, a linear combination of two or more predictor variables is used to explain the variation in a response. In essence, the additional predictors are used to explain the variation in the response not explained by a simple linear regression fit.

As an illustration, I will consider the following problem. The data set is from the statistics package **Minitab**, where it is described thus:

> *Anthropologists wanted to determine the long-term effects of altitude change on human blood pressure. They measured the blood pressures of a number of Peruvians native to the high Andes mountains who had since migrated to lower climes. Previous research suggested that migration of this kind might cause higher blood pressure at first, but over time blood pressure would decrease. The subjects were all males over 21, born at high altitudes, with parents born at high altitudes. The measurements included a number of characteristics to help measure obesity: skin-fold and other physical characteristics. Systolic and diastolic blood pressure are recorded separately; systolic is often a more sensitive indicator. Note that this is only a portion of the data collected.*

The data set is on the web site. Variables in the data set are

Name	**Description**
Age	Age in years
Years	Years since migration
Weight	Weight in kilograms
Height	Height in mm
Chin	Chin skin fold in mm
Forearm	Forearm skin fold in mm
Calf	Calf skin fold in mm
Pulse	Pulse in beats per minute
Systol	Systolic blood pressure
Diastol	Diastolic blood pressure

A question we consider concerns the long term effects of an environmental change on the systolic blood pressure. In particular, is there a relationship between the systolic blood pressure and how long the Indians lived in their new environment as measured by the fraction of their life spent in the new environment? (fraction = years since migration/age - you need to `generate` fraction).

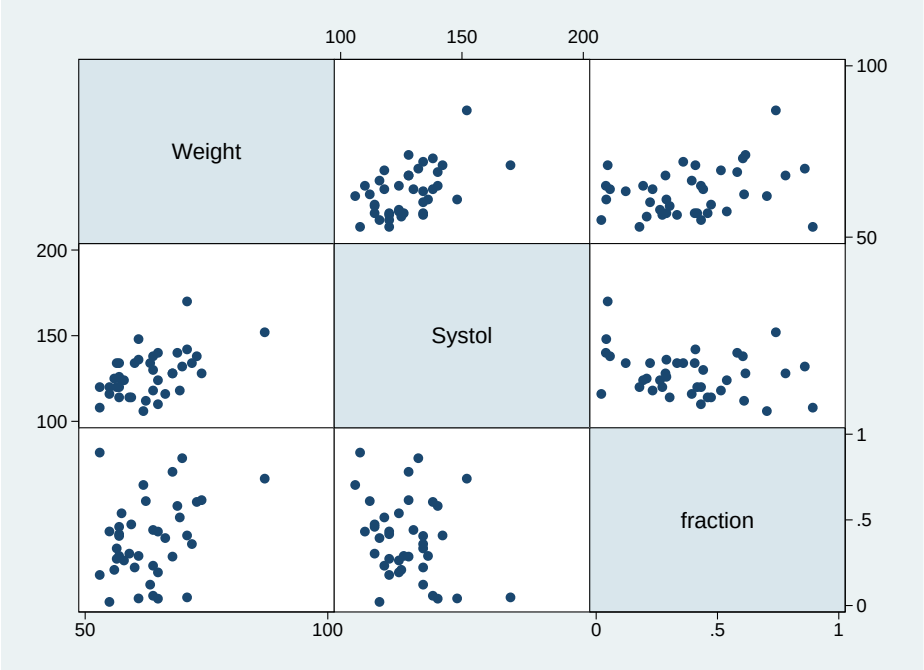
A plot of systolic blood pressure against fraction suggests a weak linear relationship (from `graph matrix weight systol fraction`). Nonetheless, consider fitting the regression model

$$sys\ bp = \beta_0 + \beta_1\ fraction + \epsilon.$$

The least squares line is given by

$$\widehat{sys\ bp} = 133.50 - 15.75\ fraction,$$

and suggests that average systolic blood pressure decreases as the fraction of life spent in modern society increases. However, the $t-$test of $H_0 : \beta_1 = 0$ is not significant at the 5% level (p-value=.089). That is, the weak linear relationship observed in the data is not atypical of a population where there is no linear relationship between systolic blood pressure and the fraction of life spent in a modern society.



Stata output:

```
. regress systol fraction

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  1,    37) =    3.05
       Model |  498.063981     1  498.063981            Prob > F      =  0.0888
    Residual |  6033.37192    37  163.064106            R-squared     =  0.0763
-------------+------------------------------           Adj R-squared =  0.0513
       Total |   6531.4359    38  171.879892            Root MSE      =   12.77

------------------------------------------------------------------------------
      systol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fraction |  -15.75183   9.012962    -1.75   0.089    -34.01382    2.510169
       _cons |   133.4957   4.038011    33.06   0.000     125.3139    141.6775
------------------------------------------------------------------------------
```

Even if this test were significant, the small value of $R^2 = .076$ suggests that fraction does not explain a **substantial** amount of the variation in the systolic blood pressures. If we omit the individual with the highest blood pressure (see the plot) then the relationship would be weaker.

Taking Weight into Consideration

At best, there is a weak relationship between systolic blood pressure and fraction. However, it is usually accepted that systolic blood pressure and weight are related; see the scatterplot matrix for confirmation. A natural way to take weight into consideration is to include weight and fraction as predictors of systolic blood pressure in the multiple regression model:

$$sys\ bp = \beta_0 + \beta_1\ fraction + \beta_2\ weight + \epsilon.$$

As in simple linear regression, the model is written in the form:

$$\text{Response} = \text{Mean of Response} + \text{Residual},$$

so the model implies that average systolic blood pressure is a linear combination of fraction and weight. As in simple linear regression, the standard multiple regression analysis assumes that the responses are normally distributed with a constant variance $\sigma^2_{Y|X}$. The parameters of the regression model $\beta_0,\ \beta_1,\ \beta_2$ and $\sigma^2_{Y|X}$ are estimated by LS.

Stata output for fitting the multiple regression model follows.

```
. regress systol fraction weight

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  2,    36) =   16.16
       Model |  3090.07324     2  1545.03662            Prob > F      =  0.0000
    Residual |  3441.36266    36  95.5934072            R-squared     =  0.4731
-------------+------------------------------           Adj R-squared =  0.4438
       Total |   6531.4359    38  171.879892            Root MSE      =  9.7772

------------------------------------------------------------------------------
      systol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fraction |  -26.76722   7.217801    -3.71   0.001    -41.40559   -12.12884
      weight |   1.216857   .2336873     5.21   0.000     .7429168    1.690796
       _cons |   60.89592   14.28088     4.26   0.000     31.93295    89.85889
------------------------------------------------------------------------------
```

Important Points to Notice About the Regression Output

1. The LS estimates of the intercept and the regression coefficient for fraction, and their standard errors, change from the simple linear model to the multiple regression model. For the simple linear regression

$$\widehat{sys\ bp} = 133.50 - 15.75\ fraction.$$

   For the multiple regression model

$$\widehat{sys\ bp} = 60.89 - 26.76\ fraction + 1.21\ weight.$$

   There is frequently a big difference between coefficients from simple linear regression and those from multiple linear regression (for the *same* predictor variables).

2. Comparing the simple linear regression and the multiple regression models we see that the Model (Regression) *df* has increased to 2 from 1 (2=number of predictor variables) and the Residual (error) *df* has decreased from 37 to 36 ($= n - 1-$ number of predictors). Adding a predictor *increases* the Model (Regression) *df* by 1 and *decreases* the Residual *df* by 1.

3. The Residual SS decreases by 6033.37 - 3441.36 = 2592.01 upon adding the weight term. The Model (Regression) SS increased by 2592.01 upon adding the weight term to the model. The Total SS does not depend on the number of predictors so it stays the same. The Residual SS, or the part of the variation in the response unexplained by the regression model never increases when new predictors are added. After all, you are not going to do any *worse* modelling the data if you use more predictors - the smaller model (simple linear regression) is a special case of the larger model (multiple linear regression). Anything you can fit using the simple one-variable model you also can fit using the two-variable model, but you can do a lot more with the two-variable model.

4. The proportion of variation in the response explained by the regression model:

$$R^2 = \text{Model (or Regression) SS / Total SS}$$

never decreases when new predictors are added to a model. The R^2 for the simple linear regression was .076, whereas $R^2 = .473$ for the multiple regression model. Adding the weight variable to the model increases R^2 by 40%. That is, weight and fraction together explain 40% more of the variation in systolic blood pressure than explained by fraction alone. I am not showing you the output, but if you predict systolic blood pressure using only weight, the R^2 is .27; adding fraction to *that* model increases the R^2 once again to .47. How well two predictors work together is not predictable from how well each works alone.

Stata also reports an *adjusted* R^2. That has a penalty for fitting too many variables built into it, and can decrease when variables are added. If the number of variables is a lot less than n (it should be) there is not much difference between the two R^2s.

5. The estimated variability about the regression line

$$\text{Residual MS} = s^2_{Y|X}$$

decreased dramatically after adding the weight effect. For the simple linear regression model (fitting fraction as the only predictor), $s^2_{Y|X} = 163.06$, whereas $s^2_{Y|X} = 95.59$ for the multiple regression model. This suggests that an important predictor has been added to model. Note that *Stata* also reports Root MSE = $\sqrt{ResidualMS} = \sqrt{s^2_{Y|X}}$, an estimate of the standard deviation rather than the variance about the regression line.

6. The F-statistic for the multiple regression model

$$F_{obs} = \text{Regression MS / Residual MS} = 16.16$$

(which is compared to a F-table with 2 and 36 *df*) tests $H_0 : \beta_1 = \beta_2 = 0$ against $H_A :$ not H_0. This is a test of no relationship between the average systolic blood pressure and fraction and weight, assuming the relationship is linear. If this test is significant then either fraction or weight, or both, are important for explaining the variation in systolic blood pressure. Unlike simple linear regression, this test statistic is not simply the square of a t-statistic. It is a whole new test for us, and simply addresses the question "is anything going on anywhere in this model?"

7. Given the model

$$sys\ bp = \beta_0 + \beta_1\ fraction + \beta_2\ weight + \epsilon,$$

one interest is testing $H_0 : \beta_2 = 0$ against $H_A : \beta_2 \neq 0$. The t-statistic for this test

$$t_{obs} = \frac{b_2 - 0}{SE(b_2)} = \frac{1.217}{.234} = 5.21$$

is compared to a $t-$critical value with Residual $df = 36$. **Stata** gives a p-value of .000, which suggests $\beta_2 \neq 0$. The t-test of $H_0 : \beta_2 = 0$ in the multiple regression model tests whether adding weight to the simple linear regression model explains a significant part of the variation in systolic blood pressure not explained by fraction. In some sense, the $t-$test of $H_0 : \beta_2 = 0$ will be significant if the increase in R^2 (or decrease in Residual SS) obtained by adding weight to this simple linear regression model is substantial. We saw a big increase in R^2, which is deemed significant by the $t-$test. A similar interpretation is given to the $t-$test for $H_0 : \beta_1 = 0$.

8. The $t-$tests for $\beta_0 = 0$ and $\beta_1 = 0$ are conducted, assessed, and interpreted in the same manner. The p-value for testing $H_0 : \beta_0 = 0$ is .000, whereas the p-value for testing $H_0 : \beta_1 = 0$ is .001. This implies that fraction is important in explaining the variation in systolic blood pressure **after** weight is taken into consideration (by including weight in the model as a predictor). These t-tests are tests for the effect of a variable *adjusted* for the effects of all other variables in the model.

9. We compute CIs for the regression parameters β_i in the usual way: $b_i + t_{crit} SE(b_i)$, where t_{crit} is the $t-$critical value for the corresponding CI level with df = Residual df.

Understanding the Model

The $t-$test for $H_0 : \beta_1 = 0$ is highly significant (p-value=.001) in the multiple regression model, which implies that fraction is important in explaining the variation in systolic blood pressure *after weight is taken into consideration* (by including weight in the model as a predictor). Weight is called a **suppressor variable**. Ignoring weight suppresses the relationship between systolic blood pressure and fraction - recall that fraction was not significant as a predictor by itself.

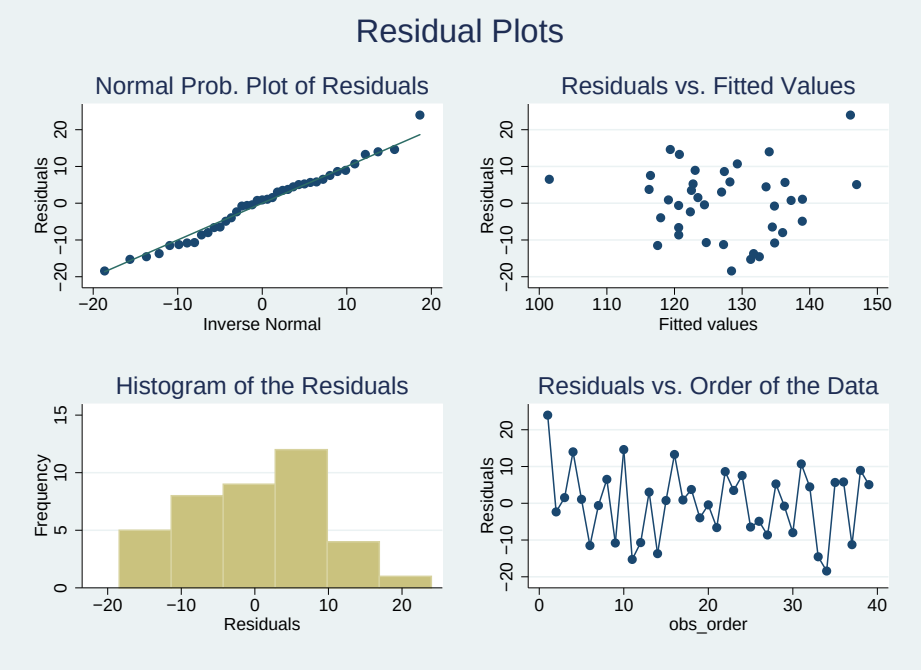
The implications of this analysis are enormous! Essentially, the correlation between a predictor and a response says very little about the importance of the predictor in a regression model with one or more additional predictors. This conclusion also holds in situations where the correlation is high, in the sense that a predictor that is highly correlated with the response may be unimportant in a multiple regression model once other predictors are included in the model.

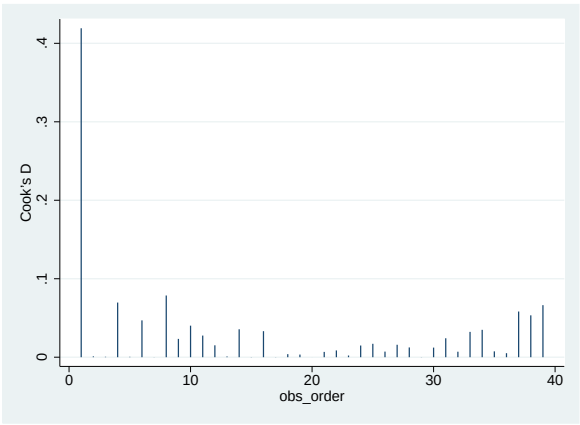
Another issue that I wish to address concerns the interpretation of the regression coefficients in a multiple regression model. For our problem, let us first focus on the fraction coefficient in the fitted model

$$\widehat{sys\ bp} = 60.90 - 26.77\ fraction + 1.22\ weight.$$

The negative coefficient indicates that the predicted systolic blood pressure decreases as fraction increases **holding weight constant**. In particular, the predicted systolic blood pressure decreases by 26.76 for each unit increase in fraction, holding weight constant at any value. Similarly, the predicted systolic blood pressure increases by 1.21 for each unit increase in weight, holding fraction constant at any level.

We should examine residuals. Now the diagnostics are much more important to us, since we cannot see everything in terms of one predictor variable.






```
                  Shapiro-Wilk W test for normal data
    Variable |     Obs        W          V          z     Prob>z
-------------+--------------------------------------------------
    residual |      39     0.98269      0.671     -0.838  0.79910
```

We will discuss these plots in class. Are there any observations we should investigate further? Which ones?

Another Multiple Regression Example

The data below are selected from a larger collection of data referring to candidates for the General Certificate of Education (GCE) who were being considered for a special award. Here, `total` denotes the candidate's TOTAL mark, out of 1000, in the GCE exam, while `comp` is the candidate's score in the compulsory part of the exam, which has a maximum score of 200 of the 1000 points on the exam. `scel` denotes the candidates' score, out of 100, in a School Certificate English Language (SCEL) paper taken on a previous occasion.

```
. list,clean

      total   comp   scel
 1.     476    111     68
 2.     457     92     46
 3.     540     90     50
 4.     551    107     59
 5.     575     98     50
 6.     698    150     66
 7.     545    118     54
 8.     574    110     51
 9.     645    117     59
10.     690    114     80
11.     634    130     57
12.     637    118     51
13.     390     91     44
14.     562    118     61
15.     560    109     66
```

A goal here is to compute a multiple regression of the TOTAL score on COMP and SCEL, and make the necessary tests to enable you to comment intelligently on the extent to which current performance in the compulsory test (COMP) may be used to predict aggregate TOTAL performance on the GCE exam, and on whether previous performance in the School Certificate English Language (SCEL) has any predictive value independently of what has already emerged from the current performance in the compulsory papers.

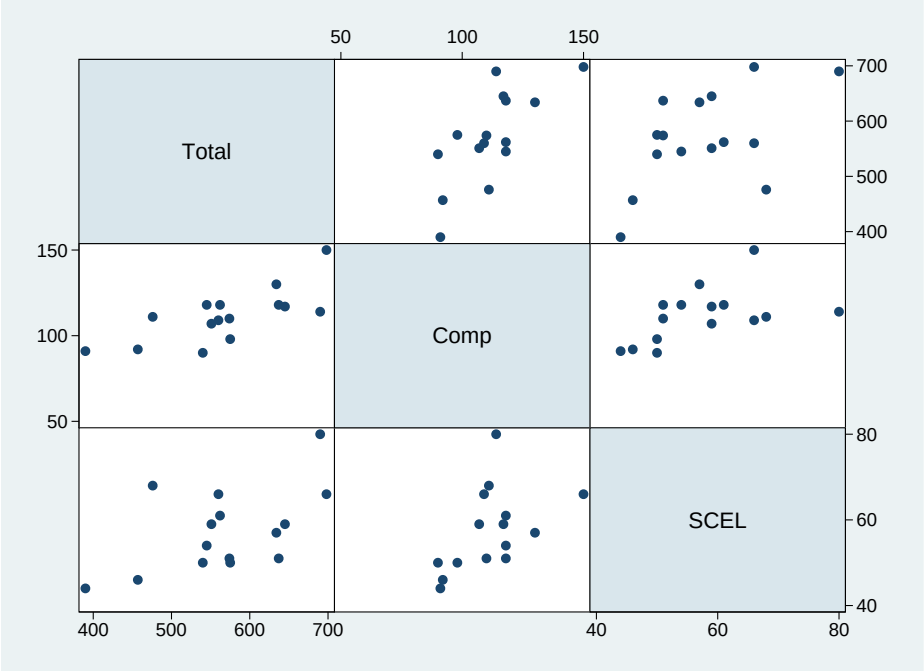
I will lead you through a number of steps to help you answer this question. Let us answer the following straightforward questions based on the **Stata** output.

1. Plot TOTAL against COMP and SCEL individually, and comment on the form (i.e. linear, non-linear, logarithmic, etc.), strength, and direction of the relationships.

2. Plot COMP against SCEL and comment on the form, strength, and direction of the relationship.

3. Compute the correlation between all pairs of variables. Do the correlation values appear reasonable, given the plots?

Stata output: scatterplot matrix and correlations...




```
. pwcorr total comp scel, sig
             |    total     comp     scel
-------------+---------------------------
       total |   1.0000
             |
             |
        comp |   0.7307   1.0000
             |   0.0020
             |
        scel |   0.5477   0.5089   1.0000
             |   0.0346   0.0527
             |
```

In parts 4 through 9, ignore the possibility that TOTAL, COMP or SCEL might ideally need to be transformed.

4. Which of COMP and SCEL explains a larger proportion of the variation in TOTAL? Which would appear to be a better predictor of TOTAL? (Explain).

5. Consider 2 simple linear regression models for predicting TOTAL one with COMP as a predictor, and the other with SCEL as the predictor. Do COMP and SCEL individually appear to be important for explaining the variation in TOTAL (i.e. test that the slopes of the regression lines are zero). Which, if any, of the output, support, or contradicts, your answer to the previous question?

Stata output:

```
. regress total comp

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) =   14.90
       Model |  53969.7272     1  53969.7272            Prob > F      =  0.0020
    Residual |  47103.2062    13  3623.32355            R-squared     =  0.5340
-------------+------------------------------           Adj R-squared =  0.4981
       Total |  101072.933    14  7219.49524            Root MSE      =  60.194

------------------------------------------------------------------------------
       total |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        comp |   3.948465   1.023073     3.86   0.002      1.73825    6.158681
       _cons |   128.5479   115.1604     1.12   0.285     -120.241    377.3367
------------------------------------------------------------------------------

. regress total scel

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) =    5.57
       Model |  30320.6397     1  30320.6397            Prob > F      =  0.0346
    Residual |  70752.2936    13  5442.48412            R-squared     =  0.3000
-------------+------------------------------           Adj R-squared =  0.2461
       Total |  101072.933    14  7219.49524            Root MSE      =  73.773

------------------------------------------------------------------------------
       total |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        scel |   4.826232   2.044738     2.36   0.035     .4088448    9.243619
       _cons |   291.5859   119.0382     2.45   0.029      34.4196    548.7522
------------------------------------------------------------------------------
```

6. Fit the multiple regression model

$$\text{TOTAL} = \beta_0 + \beta_1 \text{COMP} + \beta_2 \text{SCEL} + \epsilon.$$

Test $H_0 : \beta_1 = \beta_2 = 0$ at the 5% level. Describe in words what this test is doing, and what the results mean here.

Stata output:

```
. regress total comp scel

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  2,    12) =    8.14
       Model |  58187.5043     2  29093.7522            Prob > F      =  0.0058
    Residual |   42885.429    12  3573.78575            R-squared     =  0.5757
-------------+------------------------------           Adj R-squared =  0.5050
       Total |  101072.933    14  7219.49524            Root MSE      =  59.781

------------------------------------------------------------------------------
       total |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        comp |   3.295936   1.180318     2.79   0.016      .7242444    5.867628
        scel |    2.09104   1.924796     1.09   0.299     -2.102731    6.284811
       _cons |   81.16147   122.4059     0.66   0.520     -185.5382    347.8611
------------------------------------------------------------------------------
```

7. In the multiple regression model, test $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ individually. Describe in words what these tests are doing, and what the results mean here.

8. How does the R^2 from the multiple regression model compare to the R^2 from the individual simple linear regressions? Is what you are seeing here appear reasonable, given the tests on the individual coefficients?

9. Do your best to answer the question posed above, in the paragraph on page 43 that begins "A goal .... ". Provide an equation (LS) for predicting TOTAL.

Comments on the GCE Analysis

I will give you my thoughts on these data, and how I would attack this problem, keeping the ultimate goal in mind. As a first step, I plot the data and check whether transformations are needed. The plot of TOTAL against COMP is fairly linear, but the trend in the plot of TOTAL against SCEL is less clear. You might see a non-linear trend here, but the relationship is not very strong. When I assess plots I try to not allow a few observations affect my perception of trend, and with this in mind, I do not see any strong evidence at this point to transform any of the variables.

One difficulty that we must face when building a multiple regression model is that these two-dimensional (2D) plots of a response against individual predictors may have little information about the appropriate scales for a multiple regression analysis. In particular, the 2D plots only tell us whether we need to transform the data in a simple linear regression analysis. If a 2D plot shows a strong non-linear trend, I would do an analysis using the suggested transformations, including any other effects that are important. However, it might be that no variables need to be transformed in the multiple regression model.

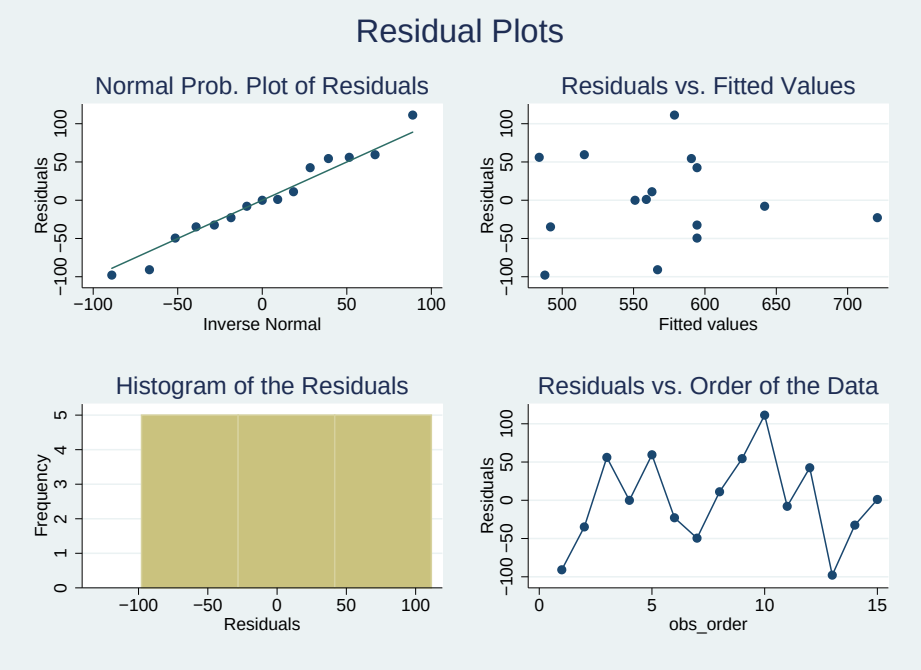
Although SCEL appears to be useful as a predictor of TOTAL on its own, the multiple regression output indicates that SCEL does not explain a significant amount of the variation in TOTAL, once the effect of COMP has been taken into account. In particular, the SCEL effect in the multiple regression model is far from significant (p-value=.30). Hence, previous performance in the SCEL exam has little predictive value independently of what has already emerged from the current performance in the compulsory papers.

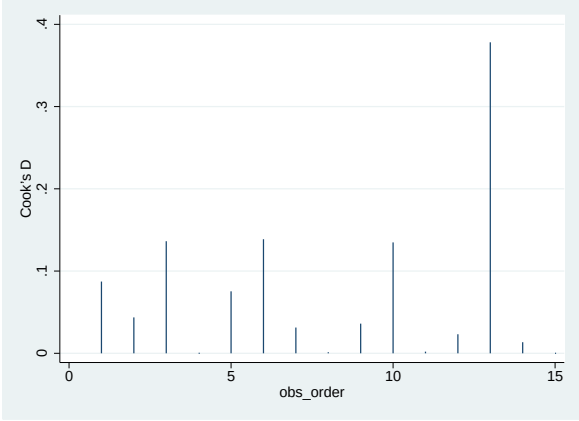
What are my conclusions? Given that SCEL is not a useful predictor in the multiple regression model, I would propose a simple linear regression model to predict TOTAL from COMP:

$$\text{Predicted TOTAL} = 128.55 + 3.95\ \text{COMP}.$$

Output from the fitted model was given earlier. A residual analysis of the model showed no serious deficiencies. In particular, the residuals versus the predicted values looks random and the normal probability plot of the residuals looks reasonably straight. Note that the following summaries are for this one-variable model, not the two-variable model fit earlier.

```
                  Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+-------------------------------------------------
    residual |     15    0.97287      0.526     -1.271  0.89806
```


Residual Plots
Normal Prob. Plot of Residuals
Residuals
Inverse Normal
Residuals vs. Fitted Values
Residuals
Fitted values
Histogram of the Residuals
Frequency
Residuals
Residuals vs. Order of the Data
Residuals
obs_order


Cook's D
obs_order

A Taste of Model Selection for Multiple Regression

Given data on a response variable Y and k predictor variables $X_1,\ X_2,\ ..., X_k$, we wish to develop a regression model to predict Y. Assuming that the collection of variables is measured on the correct scale, and that the candidate list of predictors includes all the important predictors, the most general (linear) model is

$$Y = \beta_0 + \beta_1\ X_1 + \cdots + \beta_k\ X_k + \epsilon.$$

In most problems one or more of the predictors can be eliminated from this general or **full model** without loss of information. We want to identify the important predictors, or equivalently, eliminate the predictors that are not useful for explaining the variation in Y.

We will study several automated methods for model selection. Given a specific criterion for selecting a model, **Stata** gives the best predictors. Before applying any of the methods, you should plot Y against each predictor $X_1,\ X_2, ..., X_k$ to see whether transformations are needed. If a transformation of X_i is suggested, include the transformation along with the original X_i in the candidate list. Note that you can transform the predictors differently, for example, $log(X_1)$ and $\sqrt{X_2}$. However, if several transformations are suggested for the response, then you should consider doing one analysis for each suggested response scale before deciding on the final scale.

At this point, I will only consider the **backward elimination method**. Other approaches can be handled in **Stata**.

Backward Elimination

The backward elimination procedure deletes unimportant variables, one at a time, starting from the full model. The steps in the procedure are:

1. Fit the full model
$$Y = \beta_0 + \beta_1\ X_1 + \cdots + \beta_k\ X_k + \epsilon. \quad (1)$$

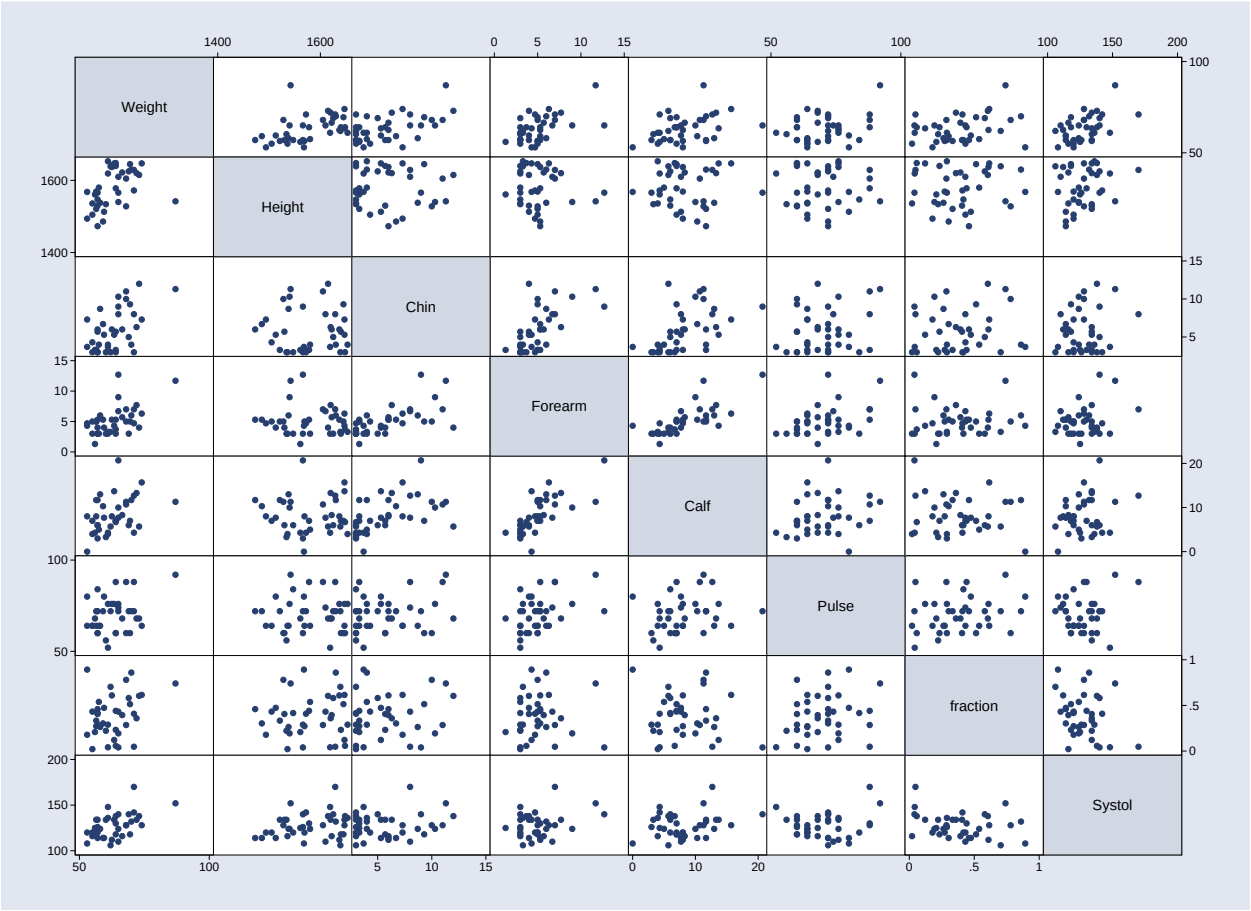
2. Find the variable which when omitted from the full model (1) reduces R^2 the least, or equivalently, increases the Residual SS the least. This is the variable that gives the largest p-value for testing an individual regression coefficient $H_0 : \beta_i = 0$ for $i > 0$. Suppose this variable is X_k. If you reject H_0, stop and conclude that the full model is best. If you do not reject H_0, delete X_k from the full model, giving the new full model
$$Y = \beta_0 + \beta_1\ X_1 + \cdots + \beta_{k-1}\ X_{k-1} + \epsilon.$$
Repeat steps 1 and 2 sequentially until no further predictors can be deleted.

In backward elimination we isolate the least important predictor left in the model, and check whether it is important. If not, delete it and repeat the process. Otherwise, stop. A test level of 0.1 (a very common value to use), for example, on the individual predictors is specified in **Stata** using `pr(0.1)` in the `sw` command.

Epidemiologists use a slightly different approach to building models. They argue strongly for the need to always include **confounding** variables in a model, regardless of their statistical significance. I will discuss this issue more completely for logistic regression, but you should recognize its importance.

Illustration

I will illustrate backward elimination on the Peru Indian data, using systolic blood pressure as the response, and seven candidate predictors: weight in kilos, height in mm, chin skin fold in mm, forearm skin fold in mm, calf skin fold in mm, pulse rate-beats/min, and fraction. A plot of systolic blood pressure against each of the individual potential predictors does not strongly suggest the need to transform either the response or any of the predictors:



The correlation matrix shows that most of the potential predictors are weakly correlated with systolic blood pressure. Based on correlations, the best single variable for predicting blood pressure is weight.

```
. correlate weight height chin forearm calf pulse fraction systol
(obs=39)
             |   weight   height     chin  forearm     calf    pulse fraction   systol
-------------+------------------------------------------------------------------------
      weight |   1.0000
      height |   0.4503   1.0000
        chin |   0.5617  -0.0079   1.0000
     forearm |   0.5437  -0.0689   0.6379   1.0000
        calf |   0.3919  -0.0028   0.5160   0.7355   1.0000
       pulse |   0.3118   0.0078   0.2231   0.4219   0.2087   1.0000
    fraction |   0.2931   0.0512   0.1201   0.0280  -0.1130   0.2142   1.0000
      systol |   0.5214   0.2191   0.1702   0.2723   0.2508   0.1355  -0.2761   1.0000
```

Stata commands for the previous output are (assuming you grabbed the **Stata** data set `peru.dta` from the web and `use`'d it)

```
generate fraction=years/age
graph matrix weight height chin forearm calf pulse fraction systol
correlate weight height chin forearm calf pulse fraction systol
```

Summaries from the full model with 7 predictors follow. The F-test in the full model ANOVA table ($F = 4.92$ with p-value=.0008) tests the hypothesis that the regression coefficient for **each** predictor variable is zero. This test is highly significant, indicating that **one or more of the predictors** is important in the model. Note that $R^2 = .53$ for the full model.

```
. regress systol weight height chin forearm calf pulse fraction
      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  7,    31) =    4.92
       Model |  3436.89993     7  490.985705            Prob > F      =  0.0008
    Residual |  3094.53596    31  99.8237407            R-squared     =  0.5262
-------------+------------------------------           Adj R-squared =  0.4192
       Total |   6531.4359    38  171.879892            Root MSE      =  9.9912
------------------------------------------------------------------------------
      systol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   1.710538   .3864434     4.43   0.000     .9223814    2.498694
      height |  -.0454089   .0394397    -1.15   0.258    -.1258466    .0350289
         chin |  -1.154889    .845932    -1.37   0.182    -2.880179    .5704004
     forearm |  -.7143249   1.350676    -0.53   0.601    -3.469047    2.040397
        calf |   .1058654   .6116778     0.17   0.864     -1.14166     1.35339
       pulse |     .07971   .1959149     0.41   0.687    -.3198611    .4792811
    fraction |  -29.35489    7.86754    -3.73   0.001    -45.40084   -13.30894
       _cons |   106.3085    53.8376     1.97   0.057    -3.494025    216.111
------------------------------------------------------------------------------
```

You can automate stepwise selection of predictors (for which backward elimination is a special case) using the **sw** command. Six model selection procedures are allowed: backward selection, forward selection, backward stepwise, forward stepwise, backward hierarchical selection, and forward hierarchical selection. See the **Stata** manual for descriptions. The command `sw` can also be used with other regression models including logistic (and other binary response model) regression, Poisson regression, and Cox proportional hazards regression. To obtain the stepwise procedure for multiple linear regression in our example, using an cutoff of 0.1, type `sw regress systol weight height chin forearm calf pulse fraction, pr(0.1)`. I cannot seem to get this to work correctly using the pull-down menus, and I'm not sure there is much potential gain anyway. This is a pretty simple command. The **Stata** output follows.

```
. sw regress systol weight height chin forearm calf pulse fraction,pr(0.1)
                     begin with full model
p = 0.8637 >= 0.1000  removing calf
p = 0.6953 >= 0.1000  removing pulse
p = 0.6670 >= 0.1000  removing forearm
p = 0.2745 >= 0.1000  removing height
p = 0.1534 >= 0.1000  removing chin
      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  2,    36) =   16.16
       Model |  3090.07324     2  1545.03662            Prob > F      =  0.0000
    Residual |  3441.36266    36  95.5934072            R-squared     =  0.4731
-------------+------------------------------           Adj R-squared =  0.4438
       Total |   6531.4359    38  171.879892            Root MSE      =  9.7772
------------------------------------------------------------------------------
      systol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   1.216857   .2336873     5.21   0.000     .7429168    1.690796
    fraction |  -26.76722   7.217801    -3.71   0.001    -41.40559   -12.12884
       _cons |   60.89592   14.28088     4.26   0.000     31.93295    89.85889
------------------------------------------------------------------------------
```

The procedure summary tells you that the least important variable in the full model, as judged by the p-value, is calf skin fold. This variable, upon omission, will reduce R^2 the least, or equivalently, increases the Residual SS the least. The p-value of .86 exceeds the specified 0.10 cut-off, so the first step of the backward elimination would be to eliminate calf skin fold from the model. This is the p-value for the t-test on calf in the 7-variable model.

The next variable eliminated is pulse because of the p-value of .70 in the 6-variable model where calf was not fit (**Stata** isn't showing you all of that output). Notice that this is different from the p-value in the 7-variable model. Next **Stata** removes forearm because of the large p-value of .67 in a 5-variable model with calf and pulse removed. Other variables are eliminated similarly. There is a huge amount of computation summarized in this one table.

Looking at the rest of the step history, the backward elimination procedure eliminates five variables from the full model, in the following order: calf skin fold, pulse rate, forearm skin fold, height, and chin skin fold. As we progress from the full model to the selected model, R^2 decreases as follows: $R^2 = .53$ (full model), .53, .52, .52, .50, and .47 (from several regression fits not shown). The decrease is slight across this spectrum of models.

The model summary selected by backward elimination includes two predictors: weight and fraction. The fitted model is given by:

$$\text{Predicted SYS BP} = 60.90 + 1.22\ \text{Weight} - 26.77\ \text{Fraction}.$$

Each predictor is significant at the .001 level. The fitted model explains 47% of the variation in systolic blood pressures. This 2-variable model does as well, for any practical purposes, in predicting systolic blood pressure as a much more complicated 7-variable model. There was no real surprise here, since these two variables were the only ones significant in the 7-variable model, but often you will be left with a model you would not have guessed from a fit of all variables.

Using a mechanical approach, we are led to a model with weight and fraction as predictors of systolic blood pressure. At this point you should closely examine the fitted model.

Stepwise procedures receive a great deal of criticism. When a large number of variables are screened this way, the resulting relationships tend to be exaggerated. There is a big multiple comparisons problem here as well. This technique should be regarded as exploratory and the resulting p-values and coefficients assessed from independent data, although common practice is just to report final results. It is likely that the strength of relationships discovered in stepwise procedures will be hard to replicate in later studies, however. This is, nonetheless, an invaluable screening device when one has a lot of predictor variables.

5 One-Way ANOVA (Review) and Experimental Design

Samuels and Witmer Chapter 11 - all sections except 6.

The one-way analysis of variance (**ANOVA**) is a generalization of the two sample $t-$test to $k \geq 2$ groups. Assume that the populations of interest have the following (unknown) population means and standard deviations:

	population 1	population 2	$\cdots$	population k
mean	μ_1	μ_2	$\cdots$	μ_k
std dev	σ_1	σ_2	$\cdots$	σ_k

A usual interest in ANOVA is whether $\mu_1 = \mu_2 = \cdots = \mu_k$. If not, then we wish to know which means differ, and by how much. To answer these questions we select samples from each of the k populations, leading to the following data summary:

	sample 1	sample 2	$\cdots$	sample k
size	n_1	n_2	$\cdots$	n_k
mean	$\overline{Y}_1$	$\overline{Y}_2$	$\cdots$	$\overline{Y}_k$
std dev	s_1	s_2	$\cdots$	s_k

A little more notation is needed for the discussion. Let Y_{ij} denote the j^{th} observation in the i^{th} sample and define the total sample size $n^* = n_1 + n_2 + \cdots + n_k$. Finally, let $\overline{\overline{Y}}$ be the average response over all samples (combined), that is

$$\overline{\overline{Y}} = \frac{\sum_{ij} Y_{ij}}{n^*} = \frac{\sum_i n_i \overline{Y}_i}{n^*}.$$

Note that $\overline{\overline{Y}}$ is *not* the average of the sample means, unless the sample sizes n_i are equal.

An $F-$statistic is used to test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ against $H_A :$ not H_0. The assumptions needed for the standard ANOVA $F-$test are analogous to the independent two-sample $t-$test assumptions: (1) Independent random samples from each population. (2) The population frequency curves are normal. (3) The populations have equal standard deviations, $\sigma_1 = \sigma_2 = \cdots = \sigma_k$.

The $F-$test is computed from the ANOVA table, which breaks the spread in the combined data set into two components, or **Sums of Squares** (SS). The **Within SS**, often called the **Residual SS** or the **Error SS**, is the portion of the total spread due to variability *within* samples:

$$\text{SS(Within)} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 = \textstyle\sum_{ij}(Y_{ij} - \overline{Y}_i)^2.$$

The **Between SS**, often called the Model SS, measures the spread between (actually among!) the sample means

$$\text{SS(Between)} = n_1(\overline{Y}_1 - \overline{\overline{Y}})^2 + n_2(\overline{Y}_2 - \overline{\overline{Y}})^2 + \cdots + n_k(\overline{Y}_k - \overline{\overline{Y}})^2 = \textstyle\sum_i n_i(\overline{Y}_i - \overline{\overline{Y}})^2,$$

weighted by the sample sizes. These two SS add to give

$$\text{SS(Total)} = \text{SS(Between)} + \text{SS(Within)} = \textstyle\sum_{ij}(Y_{ij} - \overline{\overline{Y}})^2.$$

Each SS has its own degrees of freedom (*df*). The *df*(Between) is the number of groups minus one, $k - 1$. The *df*(Within) is the total number of observations minus the number of groups: $(n_1-1)+(n_2-1)+\cdots(n_k-1) = n^*-k$. These two *df* add to give *df*(Total) $= (k-1)+(n^*-k) = n^*-1$.

The Sums of Squares and *df* are neatly arranged in a table, called the ANOVA table:

Source	*df*	SS	MS
Between Groups	$k-1$	$\sum_i n_i(\overline{Y}_i - \overline{\overline{Y}})^2$	
Within Groups	n^*-k	$\sum_i (n_i-1)s_i^2$	
Total	n^*-1	$\sum_{ij}(Y_{ij}-\overline{\overline{Y}})^2$.	

The ANOVA table often gives a **Mean Squares** (MS) column, left blank here. The Mean Square for each source of variation is the corresponding SS divided by its *df*. The Mean Squares can be easily interpreted.

The MS(Within)

$$\frac{(n_1-1)s_1^2+(n_2-1)s_2^2+\cdots+(n_k-1)s_k^2}{n^*-k} = s^2_{pooled}$$

is a weighted average of the sample variances. The MS(Within) is known as the pooled estimator of variance, and estimates the assumed common population variance. If all the sample sizes are equal, the MS(Within) is the average sample variance. The MS(Within) is identical to the **pooled variance estimator** in a two-sample problem when $k=2$.

The MS(Between)

$$\frac{\sum_i n_i(\overline{Y}_i-\overline{\overline{Y}})^2}{k-1}$$

is a measure of variability among the sample means. This MS is a multiple of the sample variance of $\overline{Y}_1,\ \overline{Y}_2, ..., \overline{Y}_k$ when all the sample sizes are equal.

The MS(Total)

$$\frac{\sum_{ij}(Y_{ij}-\overline{\overline{Y}})^2}{n^*-1}$$

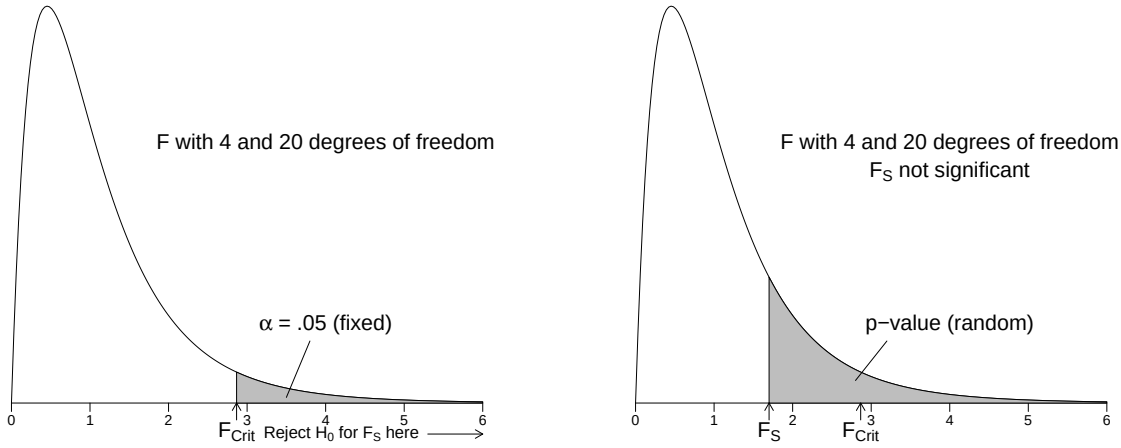
is the variance in the combined data set.

The decision on whether to reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is based on the ratio of the MS(Between) and the MS(Within):

$$F_s = \frac{MS(Between)}{MS(Within)}.$$

Large values of F_s indicate large variability among the sample means $\overline{Y}_1,\ \overline{Y}_2, ..., \overline{Y}_k$ relative to the spread of the data within samples. That is, large values of F_s suggest that H_0 is false.

Formally, for a size α test, reject H_0 if $F_s \geq F_{crit}$, where F_{crit} is the upper-α percentile from an F distribution with numerator degrees of freedom $k-1$ and denominator degrees of freedom n^*-k (i.e. the *df* for the numerators and denominators in the $F-$ratio.). An F distribution table is given on pages 687-696 of SW. The p-value for the test is the area under the $F-$ probability curve to the right of F_s. **Stata** summarizes the ANOVA $F-$test with a p-value. In **Stata**, use the `anova` or `oneway` commands to perform 1-way ANOVA. The data should be in the form of a variable containing the response Y_{ij} and a grouping variable. For $k=2$, the test is equivalent to the pooled two-sample $t-$test.

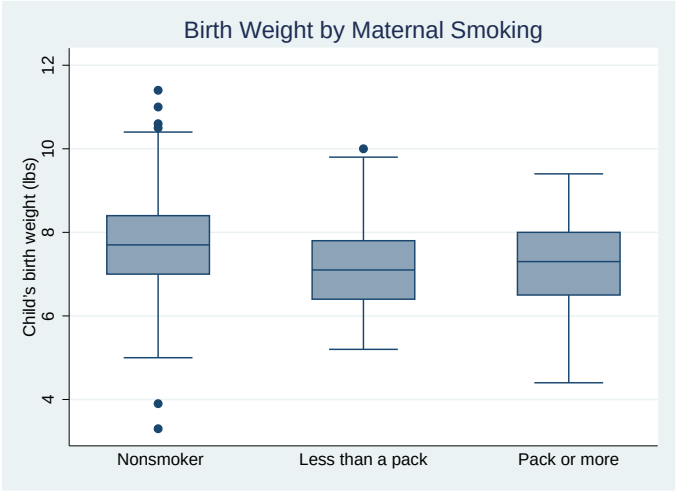


Example from the Child Health and Development Study (CHDS)

We consider data from the birth records of 680 live-born white male infants. The infants were born to mothers who reported for pre-natal care to three clinics of the Kaiser hospitals in northern California. As an initial analysis, we will examine whether maternal smoking has an effect on the birth weights of these children. To answer this question, we define 3 groups based on mother's smoking history: (1) mother does not currently smoke or never smoked (2) mother smoked less than one pack of cigarettes a day during pregnancy (3) mother smoked at least one pack of cigarettes a day during pregnancy.

Let μ_i = pop mean birth weight (in lbs) for children in group i, $(i = 1, 2, 3)$. We wish to test $H_0 : \mu_1 = \mu_2 = \mu_3$ against $H_A :$ not H_0.

The side-by-side boxplots of the data show roughly the same spread among groups and little evidence of skew:



There is no strong evidence against normality here. Furthermore the sample standard deviations are close (see the following output). We may formally test the equality of variances across the three groups (remember - the F-test is not valid if its assumptions are not met) using **Stata**'s `robvar`

command. In this example we obtain a set of three *robust* tests for the hypothesis $H_0 : \sigma_1 = \sigma_2 = \sigma_3$ where σ_i is the population standard deviation of weight in group i, $i = 1, 2, 3$. What *robust* means in this context is that the test still works reasonably well if assumptions are not quite met. The classical test of this hypothesis is Bartlett's test, and that test is well known to be extraordinarily sensitive to the assumption of normality of all the distributions. There are two ways a test may not work well when assumptions are violated - the level may not be correct, or the power may be poor. For Bartlett's test, the problem is the level may not be accurate, which in this case means that you may see a small p-value that does not reflect unequal variances but instead reflects non-normality. A test with this property is known as *liberal* because it rejects H_0 too often (relative to the nominal α).

Stata output follows; we do not reject that the variances are equal across the three groups at any reasonable significance level using any of the three test statistics:

```
. robvar( weight),by(ms_gp_txt)
                  |   Summary of Child's birth weight
                  |                (lbs)
       ms_gp_txt |        Mean   Std. Dev.       Freq.
-----------------+------------------------------------
       Nonsmoker |   7.7328084   1.0523406         381
Less than a pack |   7.2213018   1.0777604         169
    Pack or more |   7.2661539   1.0909461         130
-----------------+------------------------------------
           Total |   7.5164706   1.0923455         680
W0  = .82007944   df(2, 677)     Pr > F = .44083367
W50 = .75912861   df(2, 677)     Pr > F = .46847213
W10 = .77842523   df(2, 677)     Pr > F = .45953896
```

There are multiple ways to get the ANOVA table here, the most common being the command `anova weight ms_gp` or the more specialized `oneway weight ms_gp`. Bartlett's test for equal variances is given when using the latter. In the following output, I also gave the `,b` option to get Bonferroni multiple comparisons, discussed after the Fisher's Method (next section).

The ANOVA table is:

```
. oneway weight ms_gp_txt,b
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      40.7012466      2   20.3506233     17.90     0.0000
Within groups        769.4943    677   1.13662378
------------------------------------------------------------------------
   Total           810.195546    679   1.19321877
Bartlett's test for equal variances:  chi2(2) =   0.3055  Prob>chi2 = 0.858
          Comparison of Child's birth weight (lbs) by ms_gp_txt
                                (Bonferroni)
Row Mean-|
Col Mean |    Nonsmok   Less tha
---------+----------------------
Less tha |   -.511507
         |      0.000
         |
Pack or  |   -.466655    .044852
         |      0.000      1.000
```

The p-value for the $F-$test is less than .0001. We would reject H_0 at any of the usual test levels (i.e. .05 or .01), concluding that the population mean birth weights differ *in some way* across smoking status groups. The data (boxplots) suggest that the mean birth weights are higher for children born to mothers that did not smoke during pregnancy, but that is not a legal conclusion based upon the F-test alone.

The **Stata** commands to obtain this analysis are:

```
infile id head ...et cetera... pheight using c:/chds.txt
generate ms_gp = 1 if msmoke == 0
replace  ms_gp = 2 if msmoke >= 1 & msmoke < 20
replace  ms_gp = 3 if msmoke >= 20
gene ms_gp_txt = " Nonsmoker" if ms_gp ==1
replace ms_gp_txt = "Less than a pack" if ms_gp ==2
replace ms_gp_txt = "Pack or more" if ms_gp ==3
graph box weight, over(ms_gp_txt)
robvar weight, by (ms_gp_txt)
oneway weight ms_gp_txt,b
```

Multiple Comparison Methods: Fisher's Method

The ANOVA $F-$test checks whether all the population means are equal. **Multiple comparisons** are often used as a follow-up to a significant ANOVA $F-$test to determine which population means are different. I will discuss Fisher, Bonferroni, and Tukey methods for comparing all pairs of means. Fisher's and Tukey's approaches are implemented in **Stata** using **Stata**'s `prcomp` command. This command is not automatically installed in **Stata** 8 or 9. You will have to search for "pairwise comparisons" under **Help** > **Search...** and click on the blue `sg101` link. Click on `[Click here to install]` (your computer must be connected to the internet to do this) and you will then have access to this command.

Fisher's Least significant difference method **(LSD or FSD)** is a two-step process:

1. Carry out the ANOVA $F-$test of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ at the α level. If H_0 is not rejected, stop and conclude that there is insufficient evidence to claim differences among population means. If H_0 is rejected, go to step 2.

2. Compare each pair of means using a pooled two sample $t-$test at the α level. Use s_{pooled} from the ANOVA table and $df = df$(Residual). Using this denominator is different from just doing all the possible pair-wise t-tests.

To see where the name LSD originated, consider the $t-$test of $H_0 : \mu_i = \mu_j$ (i.e. populations i and j have same mean). The $t-$statistic is

$$t_s = \frac{\overline{Y}_i - \overline{Y}_j}{s_{pooled}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}.$$

You reject H_0 if $|t_s| \geq t_{crit}$, or equivalently, if

$$|\overline{Y}_i - \overline{Y}_j| \geq t_{crit}\ s_{pooled}\ \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

The minimum absolute difference between $\overline{Y}_i$ and $\overline{Y}_j$ needed to reject H_0 is the LSD, the quantity on the right hand side of this inequality.

Stata gives all possible comparisons between pairs of populations means. The error level (i.e. α) can be set to an arbitrary value using the `level()` subcommand, with 0.05 being the standard. Looking at the CI's in the **Stata** output, we conclude that the mean birth weights for children born to non-smoking mothers (group 1) is significantly different from the mean birth weights for each of the other two groups (2 and 3), since confidence intervals do not contain 0. The **Stata** command `prcomp weight ms_gp` produced the output (it needs group defined numerically); the default output includes CIs for differences in means. Alternatively, one obtains the p-values for testing the hypotheses that the population means are equal using the `test` subcommand. This is illustrated in the section on Tukey's method. Examining the output from the prcomp command, we see the FSD method is called the **t method** by **Stata**.

```
. prcomp weight ms_gp
                  Pairwise Comparisons of Means
Response variable (Y): weight      Child's birth weight (lbs)
Group variable (X):    ms_gp       Maternal Smoking Group
   Group variable (X): ms_gp        Response variable (Y): weight
-------------------------------    -------------------------------
      Level                              n         Mean         S.E.
-------------------------------------------------------------------
          1                            381     7.732808      .053913
          2                            169     7.221302     .0829046
          3                            130     7.266154     .0956823
-------------------------------------------------------------------

Individual confidence level: 95%    (t method)
Homogeneous error SD = 1.066126, degrees of freedom = 677

                                                                    95%
 Level(X)    Mean(Y)   Level(X)    Mean(Y)      Diff Mean    Confidence Limits
-------------------------------------------------------------------------------
       2   7.221302          1   7.732808      -.5115066   -.7049746  -.3180387
       3   7.266154          1   7.732808      -.4666546   -.6792774  -.2540318
                             2   7.221302       .0448521   -.1993527   .2890568
-------------------------------------------------------------------------------
```

Discussion of the FSD Method

With k groups, there are $c = \binom{k}{2} = \frac{k(k-1)}{2}$ pairs of means to compare in the second step of the FSD method. Each comparison is done at the α level, where for a generic comparison of the i^{th} and j^{th} populations

$$\alpha = \text{probability of rejecting } H_0 : \mu_i = \mu_j \text{ when } H_0 \text{ is true.}$$

This probability is called the **comparison error rate** or the **individual error rate**.

The individual error rate is not the only error rate that is important in multiple comparisons. The **family error rate** (FER), or the **experimentwise error rate**, is defined to be the probability of at least one false rejection of a true hypothesis $H_0 : \mu_i = \mu_j$ over all comparisons. When many comparisons are made, you *may* have a large probability of making one or more false rejections of true null hypotheses. In particular, when all c comparisons of two population means are performed, each at the α level, then $\alpha \leq FER \leq c\alpha$.

For example, in the birth weight problem where $k = 3$, there are $c = .5 * 3 * 2 = 3$ possible comparisons of two groups. If each comparison is carried out at the 5% level, then $.05 \leq FER \leq .15$. At the second step of the FSD method, you *could* have up to a 15% chance of claiming one or more pairs of population means are different if no differences existed between population means.

The first step of the FSD method is the ANOVA "screening" test. The multiple comparisons are carried out *only if* the $F-$test suggests that not all population means are equal. This screening test tends to deflate the FER for the two-step FSD procedure. However, the FSD method is commonly criticized for being extremely liberal (too many false rejections of true null hypotheses) when some, but not many, differences exist - especially when the number of comparisons is large. This conclusion is fairly intuitive. When you do a large number of tests, each, say, at the 5% level, then sampling variation alone will suggest differences in 5% of the comparisons where the H_0 is true. The number of false rejections could be enormous with a large number of comparisons. For example, chance variation alone would account for an average of 50 significant differences in 1000 comparisons each at the 5% level.

The Bonferroni Multiple Comparison Method

The Bonferroni method goes directly after the preceding relationship, $\alpha \leq FER \leq c\alpha$. To keep the FER below level α, do the individual tests at level $\frac{\alpha}{c}$, or equivalently multiply each of the reported

p-values by c. This is, in practice, extremely conservative but it does guarantee the FER is below α. If you have, for instance, $c = 3$ comparisons to make, and a reported p-value from a t-test is .02 then the Bonferroni p-value is $3(.02) = .06$ and the difference would not be judged significant. With more comparisons it becomes extremely hard for the Bonferroni method to *find* anything. The FSD method tends to have a too-high FER, the Bonferroni method a too-low FER. Very often they agree.

Earlier (p. 69) we looked at the ANOVA output following the `oneway weight group,b` command. Examining that output we see p-values of 0 for testing $H_0 : \mu_1 = \mu_2$ and $H_0 : \mu_1 = \mu_3$, and a p-value of 1 for testing $H_0 : \mu_2 = \mu_3$ using the Bonferroni method. The Bonferroni tests see group 1 differing from both 2 and 3, and no difference between 2 and 3, in complete agreement with FSD.

Tukey's Multiple Comparison Method

One commonly used alternative to FSD and Bonferroni is **Tukey's** honest significant difference method **(HSD)**. Unlike FSD (but similar to Bonferroni), Tukey's method allows you to prespecify the FER, at the cost of making the individual comparisons more conservative than in FSD (but less conservative than Bonferroni).

To implement Tukey's method with a FER of α, reject $H_0 : \mu_i = \mu_j$ when

$$|\overline{Y}_i - \overline{Y}_j| \geq \frac{q_{crit}}{\sqrt{2}} s_{pooled} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}},$$

where q_{crit} is the α level critical value of the studentized range distribution (tables not in SW). The right hand side of this equation is called the HSD. For the birth weight data, the groupings based on the Tukey and Fisher methods are identical. We obtain Tukey's groupings via the **Stata** command `prcomp weight group, tukey test`. The differences with an asterisk next to them are significant (the |numerator| is larger than the denominator):

```
. prcomp weight ms_gp,tukey test
                  Pairwise Comparisons of Means
Response variable (Y): weight      Child's birth weight (lbs)
Group variable (X):    ms_gp       Maternal Smoking Group
   Group variable (X): ms_gp        Response variable (Y): weight
------------------------------    ------------------------------
     Level                              n         Mean         S.E.
-------------------------------------------------------------------
         1                            381     7.732808      .053913
         2                            169     7.221302     .0829046
         3                            130     7.266154     .0956823
-------------------------------------------------------------------
Simultaneous significance level: 5%    (Tukey wsd method)
Homogeneous error SD = 1.066126, degrees of freedom = 677
          (Row Mean - Column Mean) / (Critical Diff)
Mean(Y) |   7.7328    7.2213 Level(X)
        |        1         2
--------+--------------------
  7.2213|  -.51151*
      2|   .23145
       |
  7.2662|  -.46665*   .04485
      3|   .25436    .29214
       |
```

Stata does not provide, as built-in commands or options, very many multiple comparison procedures. The one-way ANOVA problem we have been looking at is relatively simple, and the Tukey method appears as something of an afterthought for it. For more complicated multi-factor

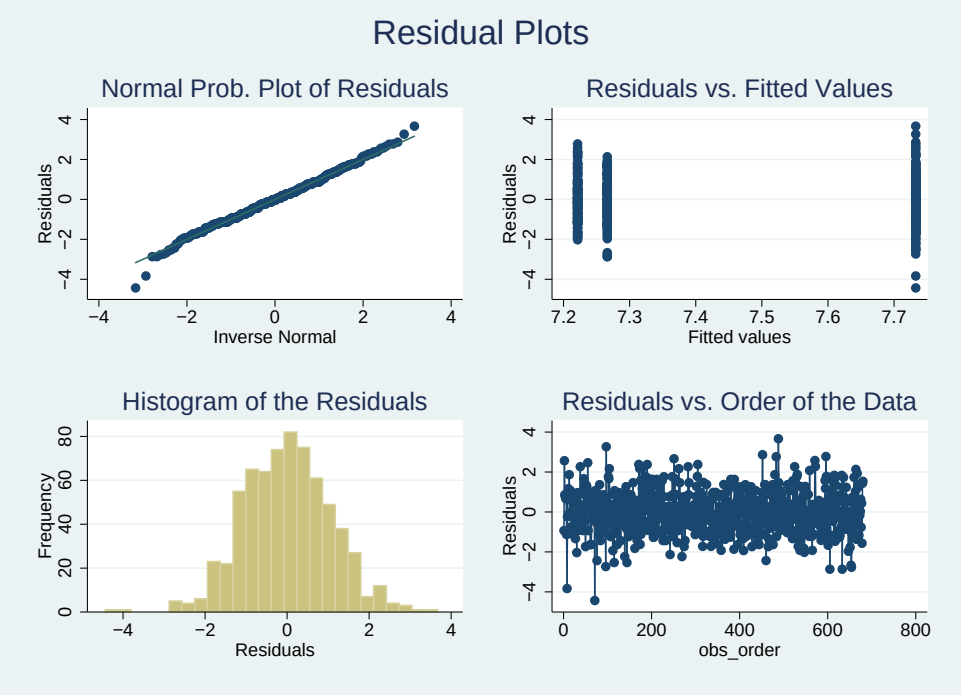
models, about all **Stata** offers is Bonferroni and a two other methods (Holm and Sidak) that adjust the p-value similarly using slightly different principles to control FER, but less conservatively than Bonferroni. The help file on mtest has details. FSD is always available, since that amounts to no adjustment. In response to questions on the www about doing multiple comparisons, **Stata** has pointed out how easy it is to program whatever you want in do files (probably the right answer for experts). Some packages like **SAS** offer a larger number of options. What **Stata** offers is adequate for many areas of research, but for some others it will be necessary to go beyond the built-in offerings of **Stata** (a reviewer on your paper will let you know!)

Checking Assumptions in ANOVA Problems

The classical ANOVA assumes that the populations have normal frequency curves and the populations have equal variances (or spreads). You can test the normality assumption using multiple Wilk-Shapiro tests (i.e. one for each sample). In addition, you can save (to the worksheet) the **centered** data values, which are the observations minus the mean for the group from which each observation comes. These centered values, or **residuals**, should behave as a single sample from a normal population. A boxplot and normal quantile test of the residuals gives an overall assessment of normality. The commands `predict residuals, resid` and then `swilk residuals` indicates that, although not significant at the 5% level, normality may be suspect:

```
. swilk residuals
                    Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+--------------------------------------------------
   residuals |    680    0.99580      1.866      1.520   0.06425
```

Mathematically, this is just a specialized regression problem and we can construct the same diagnostic plots we have been doing for regression. Cook's D is not worth doing in this case, though.



There are several alternative procedures that can be used when either the normality or equal variance assumption are not satisfied. Welch's ANOVA method (available in JMP-In, not directly available in Stata) is appropriate for normal populations with unequal variances. The test is a generalization of Satterthwaite's two-sample test discussed last semester. Most statisticians probably would use weighted least squares or transformations to deal with the unequal variance problem (we will discuss this if time permits this semester). The Wilcoxon or Kruskal-Wallis non-parametric ANOVA is appropriate with non-normal populations with similar spreads.

For the birth weight data, recall that formal tests of equal variances are not significant (p-values $> .4$). Thus, there is insufficient evidence that the population variances differ. Given that the distributions are fairly symmetric, with no extreme values, the standard ANOVA appears to be the method of choice. As an illustration of an alternative method, though, the summary from the Kruskal-Wallis approach follows, leading to the same conclusion as the standard ANOVA. One weakness of **Stata** is that it does not directly provide for non-parametric multiple comparisons. One could do all the pair-wise Mann-Whitney two-sample tests and use a Bonferroni adjustment (the `ranksum` command implements this two sample version of the Kruskal-Wallis test). The Bonferroni adjustment just multiplies all the p-values by 3 (the number of comparisons). If you do this, you find the same conclusions as with the normal-theory procedures: Group 1 differs from the other two, and groups 2 and 3 are not significantly different. Recall from last semester the Kruskal-Wallis and the Mann-Whitney amount to little more than one-way ANOVA and two-sample t-tests, respectively, on ranks in the combined samples (this controls for outliers).

```
. kwallis weight,by(ms_gp_txt)
Test: Equality of populations (Kruskal-Wallis test)
  +-----------------------------------+
  |        ms_gp_txt | Obs |  Rank Sum |
  |------------------+-----+-----------|
  |        Nonsmoker | 381 | 144979.00 |
  | Less than a pack | 169 |  47591.00 |
  |     Pack or more | 130 |  38970.00 |
  +-----------------------------------+
chi-squared =    36.594 with 2 d.f.
probability =     0.0001
chi-squared with ties =    36.637 with 2 d.f.
probability = 0.0001
```

Basics of Experimental Design

This section describes an experimental design to compare the effectiveness of four insecticides to eradicate beetles. The primary interest is determining which treatment is most effective, in the sense of providing the lowest typical survival time.

In a **completely randomized design** (CRD), the scientist might select a sample of genetically identical beetles for the experiment, and then randomly assign a predetermined number of beetles to the treatment groups (insecticides). The sample sizes for the groups need not be equal. A power analysis is often conducted to determine sample sizes for the treatments. For simplicity, assume that 48 beetles will be used in the experiment, with 12 beetles assigned to each group.

After assigning the beetles to the four groups, the insecticide is applied (uniformly to all experimental units or beetles), and the individual survival times recorded. A natural analysis of the data would be to compare the survival times using a one-way ANOVA.

There are several important controls that should be built into this experiment. The same strain of beetles should be used to ensure that the four treatment groups are alike as possible, so that

differences in survival times are attributable to the insecticides, and not due to genetic differences among beetles. Other factors that may influence the survival time, say the concentration of the insecticide or the age of the beetles, would be held constant, or fixed by the experimenter, if possible. Thus, the same concentration would be used with the four insecticides.

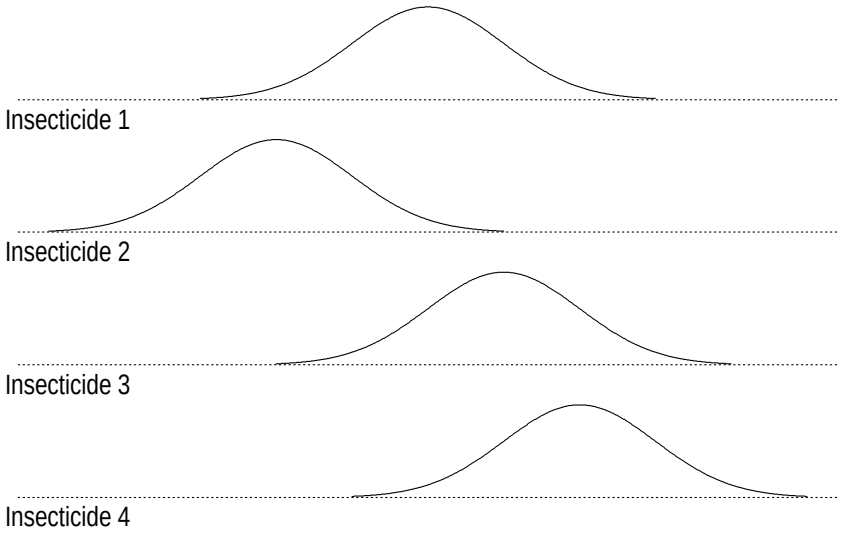
In complex experiments, there are always potential influences that are not realized or thought to be unimportant that you do not or can not control. The **randomization** of beetles to groups ensures that there is no systematic dependence of the observed treatment differences on the uncontrolled influences. This is extremely important in studies where genetic and environmental influences can not be easily controlled (as in humans, more so than in bugs or mice). The randomization of beetles to insecticides tends to diffuse or greatly reduce the effect of the uncontrolled influences on the comparison of insecticides, in the sense that these effects become part of the uncontrolled or error variation of the experiment.

Suppose y_{ij} is the response for the j^{th} experimental unit in the i^{th} treatment group, where $i = 1, 2, ..., I$. The statistical model for a completely randomized **one-factor design** that leads to a one-way ANOVA is given by:

$$y_{ij} = \mu_i + e_{ij},$$

where μ_i is the (unknown) population mean for all potential responses to the i^{th} treatment, and e_{ij} is the residual or deviation of the response from the population mean. The responses within and across treatments are assumed to be independent, normal random variables with constant variance.

For the insecticide experiment, y_{ij} is the survival time for the j^{th} beetle given the i^{th} insecticide, where $i = 1, 2, 3, 4$ and $j = 1, 2, .., 12$. The random selection of beetles coupled with the randomization of beetles to groups ensures the independence assumptions. The assumed population distributions of responses for the $I = 4$ insecticides can be represented as follows:



Let $\mu = \frac{1}{I}\sum_i \mu_i$ be the grand mean, or average of the population means. Let $\alpha_i = \mu_i - \mu$ be the i^{th} **treatment group effect**. The treatment effects add to zero, $\alpha_1 + \alpha_2 + \cdots + \alpha_I = 0$, and measure the difference between the treatment population means and the grand mean. Given this notation, the one-way ANOVA model is

$$y_{ij} = \mu + \alpha_i + e_{ij}.$$

The model specifies that the

$$\text{Response} = \text{Grand Mean} + \text{Treatment Effect} + \text{Residual}.$$

An hypothesis of interest is whether the population means are equal: $H_0 : \mu_1 = \cdots = \mu_I$, which is equivalent to the hypothesis of no treatment effects: $H_0 : \alpha_1 = \cdots = \alpha_I = 0$. If H_0 is true, then the one-way model is

$$y_{ij} = \mu + e_{ij},$$

where μ is the common population mean. We know how to test H_0 and do multiple comparisons of the treatments, so I will skip this material.

Most epidemiological studies are **observational studies** where the groups to be compared ideally consist of individuals that are similar on all characteristics that influence the response, except for the feature that defines the groups. In a designed experiment, the groups to be compared are defined by treatments randomly assigned to individuals. If, in an observational study we can not define the groups to be homogeneous on important factors that might influence the response, then we should adjust for these factors in the analysis. I will discuss this more completely in the next 2 weeks. In the analysis we just did on smoking and birth weight, we were not able to randomize with respect to several factors that might influence the response, and will need to adjust for them.

Paired Experiments and Randomized Block Experiment

A **randomized block design** is often used instead of a completely randomized design in studies where there is extraneous variation among the experimental units that may influence the response. A significant amount of the extraneous variation may be removed from the comparison of treatments by partitioning the experimental units into fairly **homogeneous subgroups** or **blocks**.

For example, suppose you are interested in comparing the effectiveness of four antibiotics for a bacterial infection. The recovery time after administering an antibiotic may be influenced by the patients general health, the extent of their infection, or their age. Randomly allocating experimental subjects to the treatments (and then comparing them using a one-way ANOVA) may produce one treatment having a "favorable" sample of patients with features that naturally lead to a speedy recovery. Additionally, if the characteristics that affect the recovery time are spread across treatments, then the variation within samples due to these uncontrolled features can dominate the effects of the treatment, leading to an inconclusive result.

A better way to design this experiment would be to **block** the subjects into groups of four patients who are alike as possible on factors other than the treatment that influence the recovery time. The four treatments are then randomly assigned to the patients (one per patient) within a block, and the recovery time measured. The blocking of patients usually produces a more sensitive comparison of treatments than does a completely randomized design because the variation in recovery times due to the blocks is eliminated from the comparison of treatments.

A randomized block design is a **paired experiment** when two treatments are compared. The usual analysis for a paired experiment is a parametric or non-parametric paired comparison. In certain experiments, each experimental unit receives each treatment. The experimental units are "natural" blocks for the analysis.

Example: Comparison of Treatments to Relieve Itching

Ten male volunteers between 20 and 30 years old were used as a study group to compare seven treatments (5 drugs, a placebo, and no drug) to relieve itching. Each subject was given a different treatment on seven study days. The time ordering of the treatments was randomized across days. Except on the no-drug day, the subjects were given the treatment intravenously, and then itching was induced on their forearms using an effective itch stimulus called cowage. The subjects recorded the duration of itching, in seconds. The data are given in the table below. From left to right the drugs are: papaverine, morphine, aminophylline, pentobarbitol, tripelenamine.

Patient	Nodrug	Placebo	Papv	Morp	Amino	Pento	Tripel
1	174	263	105	199	141	108	141
2	224	213	103	143	168	341	184
3	260	231	145	113	78	159	125
4	255	291	103	225	164	135	227
5	165	168	144	176	127	239	194
6	237	121	94	144	114	136	155
7	191	137	35	87	96	140	121
8	100	102	133	120	222	134	129
9	115	89	83	100	165	185	79
10	189	433	237	173	168	188	317

The volunteers in the study were treated as blocks in the analysis. At best, the volunteers might be considered a representative sample of males between the ages of 20 and 30. This limits the extent of inferences from the experiment. The scientists can not, without sound medical justification, extrapolate the results to children or to senior citizens.

The Analysis of a Randomized Block Design

Assume that you designed a randomized block experiment with I blocks and J treatments, where each treatment occurs once in each block. Let y_{ij} be the response for the j^{th} treatment within the i^{th} block. The model for the experiment is

$$y_{ij} = \mu_{ij} + e_{ij},$$

where μ_{ij} is the population mean response for the j^{th} treatment in the i^{th} block and e_{ij} is the deviation of the response from the mean. The population means are assumed to satisfy the additive model

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

where μ is a grand mean, α_i is the effect for the i^{th} block, and β_j is the effect for the j^{th} treatment. The responses are assumed to be independent across blocks, normally distributed and with constant variance. The randomized block model does not require the observations within a block to be independent, but does assume that the correlation between responses within a block is identical for each pair of treatments. This is a reasonable working assumption in many analyses. In this case you really need to be sure the order in which treatments are administered to subjects is randomized in order to assume equal correlation.

The model is sometimes written as

$$\text{Response} = \text{Grand Mean} + \text{Treatment Effect} + \text{Block Effect} + \text{Residual}.$$

Given the data, let $\bar{y}_{i\cdot}$ be the i^{th} block sample mean (the average of the responses in the i^{th} block), $\bar{y}_{\cdot j}$ be the j^{th} treatment sample mean (the average of the responses on the j^{th} treatment), and $\bar{y}_{\cdot\cdot}$ be the average response of all IJ observations in the experiment.

An ANOVA table for the randomized block experiment partitions the Model SS into SS for Blocks and Treatments.

Source	*df*	SS	MS
Blocks	$I-1$	$J\sum_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$	
Treats	$J-1$	$I\sum_j(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2$	
Error	$(I-1)(J-1)$	$\sum_{ij}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2$	
Total	$IJ-1$	$\sum_{ij}(y_{ij}-\bar{y}_{\cdot\cdot})^2$.	

A primary interest is testing whether the treatment effects are zero: $H_0 : \beta_1 = \cdots = \beta_J = 0$. The treatment effects are zero if the population mean responses are identical for each treatment. A formal test of no treatment effects is based on the p-value from the F-statistic F_{obs} = MS Treat/MS Error. The p-value is evaluated in the usual way (i.e. as an upper tail area from an F-distribution with $J-1$ and $(I-1)(J-1)$ df.) This H_0 is rejected when the treatment averages $\bar{y}_{\cdot j}$ vary significantly relative to the error variation.

A test for no block effects ($H_0 : \alpha_1 = \cdots = \alpha_I = 0$) is often a secondary interest, because, if the experiment is designed well, the blocks will be, by construction, noticeably different. There are no block effects if the block population means are identical. A formal test of no block effects is based on the p-value from the F-statistic F_{obs} = MS Blocks/MS Error. This H_0 is rejected when the block averages $\bar{y}_{i\cdot}$ vary significantly relative to the error variation.

A Randomized Block Analysis of the Itching Data

The `anova` command is used to get the randomized block analysis. You will be shown the steps in Thursday's Lab, but I will mention a few important points.

- The data are comprised of three variables: **itchtime**, **person** (ranges from 1-10), and **treatment** (ranges from 1-7). A data file called itch.txt was created with these three variables to be read into **Stata**.
- In the **anova** table, persons play the role of Blocks in this analysis. Using the commands `infile itchtime person treatment using c:/itch.txt` and then `anova itchtime person treatment` we obtain the following output:

```
                  Number of obs =      70     R-squared     =  0.4832
                  Root MSE      = 55.6327     Adj R-squared =  0.3397
    Source |  Partial SS    df       MS           F     Prob > F
-----------+----------------------------------------------------
     Model |    156292.6    15  10419.5067       3.37     0.0005
           |
    person |  103279.714     9  11475.5238       3.71     0.0011
 treatment |  53012.8857     6  8835.48095       2.85     0.0173
           |
  Residual |  167129.686    54  3094.99418
-----------+----------------------------------------------------
     Total |  323422.286    69   4687.2795
```

- The Model SS is the Sum of the **person** SS and **treatment** SS; check that they add up. The F-test on the Whole-Model test ANOVA checks for whether Treatments or Persons, or both, are significant, i.e. provides an overall test of all effects in the model.

- Next comes the SS for Persons and Treatments, and the corresponding F-statistics and p-values.
- It is possible in Minitab, JMP-IN and SAS (but not directly in **Stata**) to obtain Tukey multiple comparisons of the treatments. These are options in the analysis of the individual effects.
- In **Stata**, we obtain the results of testing differences in the treatments (averaged over persons) using Fisher's method from the `test` command. You will cover this in more detail in Thursday's lab. To obtain Bonferroni's adjusted p-values, simply multiply the p-value for each of Fisher's tests by the number of comparisons you are making; in the itching time example this is $\begin{pmatrix} 7 \\ 2 \end{pmatrix} = 21$ paired comparisons. We obtain, for example the results of Fisher's method for comparing treatment 1 with treatment 2 (no drug versus placebo) and treatment 1 with treatment 3 (no drug versus papaverine) with the following commands:

```
test _b[treatment[1]] = _b[treatment[2]]
test _b[treatment[1]] = _b[treatment[3]]
```

We obtain the output:

```
 ( 1)  treatment[1] - treatment[2] = 0
       F(  1,    54) =    0.31
            Prob > F =    0.5814
 ( 1)  treatment[1] - treatment[3] = 0
       F(  1,    54) =    8.56
            Prob > F =    0.0050
```

We see that using Fisher's method, treatments 1 and 2 do not significantly differ, but treatments 1 and 3 do significantly differ at the 5% level. The corresponding Bonferroni p-values are $0.58(2) > 1$ and $0.005(2) = 0.01$ for only the *two* comparisons. They are $0.58(21) > 1$ and $0.005(21) = 0.105$ if *all* 21 paired comparisons are to be made. We would accept that there is no significant difference in mean itching time between either pairs of treatments when all 21 comparisons are to be made. The tabulated p-values resulting from Fisher's method are:

Treatment	1	2	3	4	5	6
2	0.58					
3	0.01	0.00				
4	0.09	0.03	0.23			
5	0.07	0.02	0.30	0.88		
6	0.56	0.26	0.02	0.26	0.20	
7	0.34	0.13	0.05	0.44	0.36	0.71

We have the following groupings:

```
3 5 4 7 6 1 2
-------
  ---------       Fisher's
      -------


-----------
  -----------     Bonferroni's [and Tukey's obtained in JMP-IN]
```

Looking at the means for each treatment averaged over persons, we see that each of the five drugs appears to have an effect, compared to the placebo and to no drug, which have similar means. Papaverine appears to be the most effective drug, whereas placebo is the least effective treatment. A formal F-test shows significant differences among the treatments (p-value=0.017), and among patients (p-value=0.001). The only significant pairwise difference in treatments is between papaverine and placebo using Bonferroni (or Tukey) adjustments.

This all looks more difficult than it needs to be in practice. The usual strategy to start grouping population means this way is first to get the ordering of sample means. Examining the following produces the order `3 5 4 7 6 1 2` above.

```
. mean itchtime,over(treatment) noheader
------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
itchtime     |
           1 |        191   17.34871      156.3903    225.6097
           2 |      204.8   33.43278      138.1034    271.4966
           3 |      118.2   16.69983      84.88474    151.5153
           4 |        148   14.14763      119.7762    176.2238
           5 |      144.3   13.30585      117.7556    170.8444
           6 |      176.5   21.77422      133.0616    219.9384
           7 |      167.2   21.34521      124.6175    209.7825
------------------------------------------------------------
```

The target p-value for the Fisher method probably will be .05, and for the Bonferroni method is obtained by simple calculation:

```
. disp .05/21
.00238095
```

We really want to avoid running all 21 tests, and we can skip most. Once the comparisons between 3 and 7 turn out not significant, it is unnecessary to compare 3 to 5 and 4. Once the comparison between 3 and 6 turns out significant, it is not necessary to compare 3 to 1 and 2. Careful examination of patterns can make this fairly quick.

For this particular problem, there are a few outliers and possible problems with the normality assumption. The data set is on the web site - do the residual analysis, and try transforming itchtime with something like the square root to handle outliers a little better. Boxplots can be very valuable here. Redo the comparisons to see if anything has changed in the transformed scale.

A final note: An analysis that ignored person (the blocking factor), i.e. a simple one-way ANOVA, would be incorrect here since it would be assuming all observations are independent. In fact, that analysis finds no differences because the MSE is too large when ignoring blocks (you still should not treat that p-value as valid).

6 Two-factor Experiments

Last week we considered a CRD (completely randomized design) for comparing insecticides where the levels of one factor (insecticide) vary while controlling other factors that influence survival time. The inferences from the one-way ANOVA apply to beetles with a given age from the selected strain that might be given the selected concentration of the insecticides. Any generalization of the conclusions to other situations must be justified scientifically, typically through further experimentation.

Recall the way we set up the model: y_{ij} is the response for the j^{th} experimental unit (replicate) in the i^{th} treatment group, where $i = 1, 2, ..., I$;

$$y_{ij} = \mu_i + \epsilon_{ij},$$

where μ_i is the (unknown) population mean for all potential responses to the i^{th} treatment, and ϵ_{ij} is the residual or deviation of the response from the population mean. The responses within and across treatments are assumed to be independent, normal random variables with constant variance. We further decomposed μ_i as $\mu_i = \mu + \alpha_i$.

There are several ways to broaden the scope of the study. For example, several strains of beetles or several concentrations of the insecticide might be used. For simplicity, consider a simple two-factor experiment where three concentrations (Low, Medium, and High) are applied with each of the four insecticides. This is a completely crossed **two-factor experiment** where each of the $4 \times 3 = 12$ combinations of the two factors (insecticide and dose) are included in the comparison of survival times. With this experiment, the scientist can compare insecticides, compare concentrations, and check for an interaction between dose and insecticide.

Assuming that 48 beetles are available, the scientist would randomly assign them to the 12 experimental groups, giving prespecified numbers of beetles to the 12 groups. For simplicity, assume that the experiment is **balanced**, that is, the same number of beetles (4) is assigned to each group ($12 \times 4 = 48$). This is a CRD with two factors.

A Balanced Two-Factor Model

We will analyze survival times of groups of four beetles randomly allocated to twelve treatment groups obtained by crossing the levels of four insecticides (1,2,3,4) at each of three concentrations of the insecticides (1=Low, 2=Medium, 3=High). This is a balanced 4-by-3 **factorial** design (two-factor design) that is replicated four times. Three variables are needed to uniquely represent each response in the spreadsheet: dose (1-3, nominal), insecticide (1-4, nominal), and the survival time (called time). The unit of measure for the survival times is 10 hours. That is, .3 is a survival time of 3 hours. The data are given below, collected into 12 **cells** (4 rows and 3 columns):

	Dose		
Insecticide	1	2	3
1	.31, .45, .46, .43	.36, .29, .40, .23	.22, .21, .18, .23
2	.82, 1.10, .88, .72	.92, .61, .49, 1.24	.30, .37, .38, .29
3	.43, .45, .63, .76	.44, .35, .31, .40	.23, .25, .24, .22
4	.45, .71, .66, .62	.56, 1.02, .71, .38	.30, .36, .31, .33

We model this in terms of population means, just as we did in the one-way ANOVA. Now, though, population means are indexed two ways, by insecticide and by dose, so we write

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}; \;\; i = 1, 2, \ldots, I; \;\; j = 1, 2, \ldots, J; \;\; k = 1, 2, \ldots, K$$

where i refers to insecticide (I=4), j refers to dose (J=3), and k refers to replicate (K=4). For instance $y_{314} = .76$. Since we have the same number (4) of replicates in each cell this is called balanced. More generally it happens that $k = 1, 2, \ldots, K_{ij}$, i.e. there can be different numbers of replicates (usually not designed that way, but things happen!), and the analysis gets somewhat more complicated. We will consider an unbalanced problem later.

In the one-way problem, the basic test of hypothesis is that all the means are equal. That is not very useful here. What we want to do is compare insecticides, compare doses, and see if the effect of dose varies with insecticide. We need to define some additional population averages to attack all these hypotheses. The population marginal mean for Insecticide i is $\bar{\mu}_{i.} = \frac{1}{J}\sum_{j=1}^{J}\mu_{ij} = \frac{1}{3}\sum_{j=1}^{3}\mu_{ij}$, the average of Insecticide i across Dose levels. The population marginal mean for Dose j is $\bar{\mu}_{.j} = \frac{1}{I}\sum_{i=1}^{I}\mu_{ij} = \frac{1}{4}\sum_{i=1}^{4}\mu_{ij}$, the average of Dose j across Insecticide levels. There also is an overall population average, $\bar{\mu}_{..} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij} = \frac{1}{12}\sum_{i=1}^{4}\sum_{j=1}^{3}\mu_{ij}$, the average of all 12 population means. What we are interested in is the structure in the following table of population mean values:

	Dose			
Insecticide	1	2	3	Insecticide marginal
1	μ_{11}	μ_{12}	μ_{13}	$\bar{\mu}_{1.}$
2	μ_{21}	μ_{22}	μ_{23}	$\bar{\mu}_{2.}$
3	μ_{31}	μ_{32}	μ_{33}	$\bar{\mu}_{3.}$
4	μ_{41}	μ_{42}	μ_{43}	$\bar{\mu}_{4.}$
Dose marginal	$\bar{\mu}_{.1}$	$\bar{\mu}_{.2}$	$\bar{\mu}_{.3}$	$\bar{\mu}_{..}$

The basic unit of analysis is **sample cell means**, which are the direct estimators of the above population averages. We have a sample of K observations in cell ij - the natural estimator of μ_{ij} is $\bar{y}_{ij.} = \frac{1}{K}\sum_{k=1}^{K} y_{ijk} = \frac{1}{4}\sum_{k=1}^{4} y_{ijk}$. We define sample marginal means as we did for population values above (row averages and column averages), $\bar{y}_{i..} = \frac{1}{J}\sum_{j=1}^{J}\bar{y}_{ij.}$, $\bar{y}_{.j.} = \frac{1}{I}\sum_{i=1}^{I}\bar{y}_{ij.}$, and $\bar{y}_{...} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\bar{y}_{ij.}$. this gives us natural estimators of the above population values as the sample values:

	Dose			
Insecticide	1	2	3	Insecticide marginal
1	$\bar{y}_{11.}$	$\bar{y}_{12.}$	$\bar{y}_{13.}$	$\bar{y}_{1..}$
2	$\bar{y}_{21.}$	$\bar{y}_{22.}$	$\bar{y}_{23.}$	$\bar{y}_{2..}$
3	$\bar{y}_{31.}$	$\bar{y}_{32.}$	$\bar{y}_{33.}$	$\bar{y}_{3..}$
4	$\bar{y}_{41.}$	$\bar{y}_{42.}$	$\bar{y}_{43.}$	$\bar{y}_{4..}$
Dose marginal	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	$\bar{y}_{.3.}$	$\bar{y}_{...}$

Values calculated with these data are as follows:

	Dose			
Insecticide	1	2	3	Insect marg
1	.413	.320	.210	.314
2	.880	.815	.335	.677
3	.568	.375	.235	.393
4	.610	.667	.325	.534
Dose marg	.618	.544	.276	.479

Because the experiment is balanced, a marginal mean also is the average of all observations that receive a given treatment. For example, the marginal mean for insecticide 1 is the average survival time for the 12 beetles given insecticide 1.

The basic model for the two-factor design, as applied to this experiment is that the

$$\text{Response} = \text{Grand Mean} + \text{Insect Effect} + \text{Dose Effect} + \text{Insect} * \text{Dose Interaction} + \text{Residual}.$$

or

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

i.e. $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$. The assumptions for the analysis of the model are identical to those for a one-way ANOVA on the 12 treatment combinations (insecticide and dose), i.e. all 48 residual effects ϵ_{ijk} are independent and variances are all the same, $\sigma^2_{ij} = \sigma^2$. There are constraints put on the terms above, since there are too many, but those depend upon the software used. We will discuss this at some length.

The ANOVA table for this experimental design decomposes the total variation in the data, as measured by the Total SS, into components that measure the variation of marginal sample means of insecticide and dose individually (the Insecticide SS and Dose SS), a component that measures the degree to which the factors interact (the insecticide by dose SS), and a component that pools the sample variances across the 12 samples (the Error SS). Each SS has a df, given in the following ANOVA table. As usual, the MS for each source of variation is the corresponding SS divided by the df. The MS Error estimates the common population variance for the 12 treatments.

Source	df	SS	MS = SS/df
Insecticide	$I-1=4-1=3$	$JK\sum_{i=i}^{I}(\bar{y}_{i..}-\bar{y}_{...})^2$	
Dose	$J-1=3-1=2$	$IK\sum_{j=i}^{J}(\bar{y}_{.j.}-\bar{y}_{...})^2$	
Interaction	$(I-1)(J-1)=(3)(2)=6$	$K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij.}-\bar{y}_{i.}-\bar{y}_{.j.}+\bar{y}_{...})^2$	
Error	$IJ(K-1)=(4)(3)(3)=36$	$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{ij.})^2$	
Total	$IJK-1=48-1=47$	$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{...})^2$	

Believe it or not, these formulas actually make sense! SS for Insecticide just compares the Insecticide marginal means to each other by computing their variance (up to a constant), and similarly with the SS for Dose. SS Total is the usual sum of all squared deviations from the overall mean. MS Error is just $\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}s^2_{ij} = s^2_{pooled}$, the average of the sample variances from each of the cells, a natural way to estimate σ^2 (this is the within cell variability). The only term that is not easily understood is SS for Interaction. We will turn to that in a little while (after we have decided what interaction is).

There are three usual tests of interest.

1. The test of no insecticide effect. The absence of an insecticide effect implies that each level of insecticide has the same **population mean** response **when the means are averaged over levels of** dose. The test for no insecticide effect is based on the p-value for the $F-$statistic: F_{obs} = MS Insecticide/MS Error. This hypothesis is rejected when the insecticide marginal means vary significantly relative to the within cell variation. Formally, this is a test of $H_0 : \bar{\mu}_{1.} = \bar{\mu}_{2.} = \bar{\mu}_{3.} = \bar{\mu}_{4.}(= \bar{\mu}_{..})$. The form of the SS certainly matches this hypothesis.

2. The test of no dose effect. The absence of a dose effect implies that each dose level has the same **population mean** response **when the means are averaged over levels of** insecticide. The test for no dose effect is based on the p-value for the $F-$statistic: F_{obs} = MS Dose/MS Error. This hypothesis is rejected when the marginal means for dose vary significantly relative to the within cell variation. Formally, this is a test of $H_0 : \bar{\mu}_{.1} = \bar{\mu}_{.2} = \bar{\mu}_{.3}(= \bar{\mu}_{..})$. The form of the SS certainly matches this hypothesis.

3. The test of no interaction between dose and insecticide is based on the p-value for the $F-$statistic: F_{obs} = MS Interaction/MS Error. This is a test of a hypothesis that the structure is simple. Let's explore what that means.

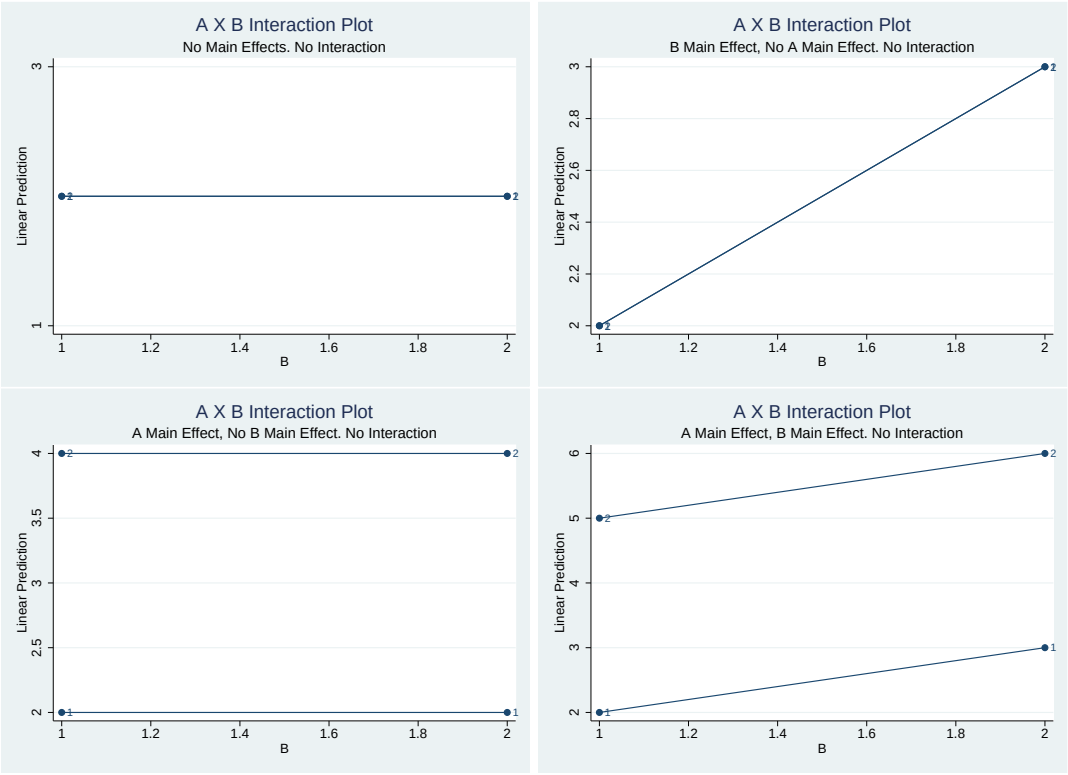
Interpretation of Interaction

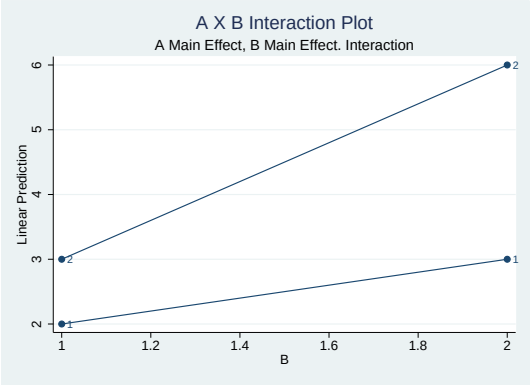
The idea of **no** interaction is that the margins of the table tell you what the structure is. If row 2, for instance, is "large" for one column, it is similarly large for all the other columns. This gets awkward to quantify, though, and we need a better approach. We have already seen the decomposition $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$. This imposes no restrictions on the cell means. We *can* impose a restriction if we force $(\alpha\beta)_{ij} \equiv 0$ so that $\mu_{ij} = \mu + \alpha_i + \beta_j$. This **additive** model is how we force the margins of the table to tell us everything. The hypothesis of no interaction is thus formally $H_0 : (\alpha\beta)_{ij} = 0$ for all i and j.

There are serious implications of this hypothesis. One is that $\mu_{ij} = \bar{\mu}_{i.} + \bar{\mu}_{.j} - \bar{\mu}_{..}$. If you look back at the SS for Interaction, this is exactly what is being tested. That is not nearly as useful or interesting, though, as this: If i, i', j, j' are legal indexes, then $\mu_{ij} - \mu_{ij'} = \mu_{i'j} - \mu_{i'j'}$, which is to say the difference between doses j and j' is the same for insecticide i as for insecticide i'; and $\mu_{ij} - \mu_{i'j} = \mu_{ij'} - \mu_{i'j'}$, which is to say the difference between insecticides i and i' is the same for dose j as it is for dose j'. These differences in cell means are slopes of line segments in interaction plots. What no interaction tells you is that the slopes of the line segments (connecting) sample cell means should be approximately parallel, and the formal test for no interaction is a check on whether the profile plots of the population means are perfectly parallel.

Prototype Interaction Plots

These profile plots are extremely important tools for understanding our analysis, so let us examine various possible patterns. Consider the simplest example with two factors A and B each at two levels, and let the population cell means μ_{ij} be broken down as $\mu_{ij} = \mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}$. Following are plots of those population cell means (those based on data have noise in them so will not be so perfect) for various combinations of effects present. Make sure you understand why each appears as it does and match it with the appropriate model.





Stata Analysis of Insecticide Data

The data need to go into three columns in the spreadsheet with full information for each observation as follows (note that I folded output to save space – this should go on for 48 rows):

poison	dose	time	poison	dose	time	poison	dose	time
1	1	.43	2	2	1.24	3	3	.23
1	1	.46	2	2	.61	3	3	.25
1	1	.31	2	2	.49	3	3	.24
1	1	.45	2	2	.92	3	3	.22
1	2	.4	2	3	.29	4	1	.66
1	2	.29	2	3	.3	4	1	.45
1	2	.36	2	3	.37	4	1	.62
1	2	.23	2	3	.38	4	1	.71
1	3	.23	3	1	.63	4	2	.56
1	3	.21	3	1	.43	4	2	.71
1	3	.18	3	1	.76	4	2	1.02
1	3	.22	3	1	.45	4	2	.38
2	1	.88	3	2	.4	4	3	.3
2	1	1.1	3	2	.35	4	3	.31
2	1	.72	3	2	.31	4	3	.36
2	1	.82	3	2	.44	4	3	.33

The table of cell and marginal means

	Dose			
Insecticide	1	2	3	Insect marg
1	.413	.320	.210	.314
2	.880	.815	.335	.677
3	.568	.375	.235	.393
4	.610	.667	.325	.534
Dose marg	.618	.544	.276	.479

is produced within **Stata** thus

```
. tabulate poison dose,summarize(time) means
                           Means of time

           |                  dose
    poison |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |     .4125        .32        .21 | .31416667
         2 | .88000001  .81500001       .335 | .67666667
         3 |     .5675       .375       .235 |     .3925
         4 |       .61  .66749999  .32500001 | .53416667
-----------+---------------------------------+----------
     Total |     .6175   .544375    .27625 |   .479375
```

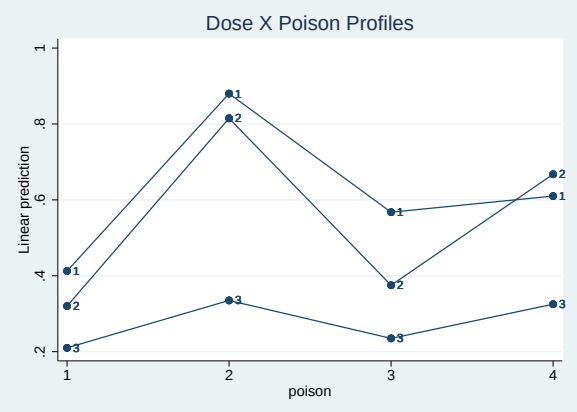
The ANOVA table is produced using the `anova` command forcing the software to fit the model $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$

```
. anova time poison dose poison*dose
                          Number of obs =      48     R-squared     =  0.7335
                          Root MSE      = .149139     Adj R-squared =  0.6521
                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  2.20435628    11  .200396025       9.01     0.0000
                         |
                  poison |  .921206282     3  .307068761      13.81     0.0000
                    dose |  1.03301249     2  .516506246      23.22     0.0000
             poison*dose |  .250137502     6  .041689584       1.87     0.1123
                         |
                Residual |  .800724989    36  .022242361
              -----------+----------------------------------------------------
                   Total |  3.00508126    47  .063937899
```

To examine interaction, consider the Dose*Insecticide profile plot given below. For each dose, we have a plot of the mean survival times across insecticides, giving 3 profiles. There is no interaction in the data if these profiles are parallel. The formal test for no interaction is a check on whether the profile plots of the population means are perfectly parallel. Every statistical package requires some special means to obtain these plots. In earlier versions of Stata we downloaded a command named `cmeans`, but that appears no longer available. The method now is easy enough if a little obscure looking:

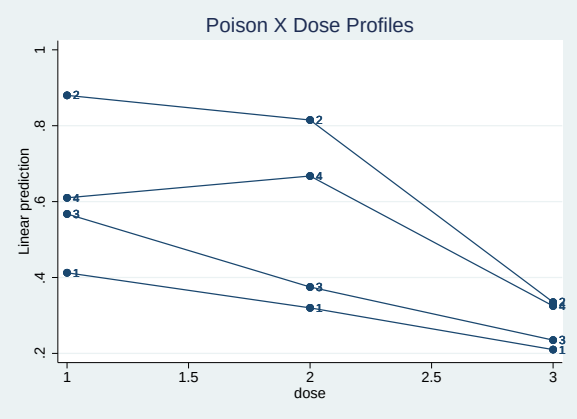
```
anova time poison dose poison*dose
predict yhat,xb
sort dose poison
scatter yhat poison,c(L) mlabel(dose) title(Dose X Poison Profiles)
```

The points being plotted by this method are those actually fit by the `anova` model – with the `poison*dose` term we are imposing no restrictions so these are the averages we calculated earlier. Had we left that term off we would have forced the profiles to be parallel since we would have imposed an additive model. The order of the sort is very important here. The `c(L)` option is a very special connected line version suited to this application (`c(l)` does not work). `mlabel` lets us label dose levels.



The last two lines above could be modified to produce four profiles, one for each poison. The plots are equally useful – sometimes both are worth examining.

```
sort poison dose
scatter yhat dose,c(L) mlabel(poison) title(Poison X Dose Profiles)
```



These lines are not parallel, but they have the same general trend. If we accept that the interaction is not significant, we may fit the additive model instead and base inferences on additive structure. To fit the model with the interaction term we type `anova time poison dose poison*dose`; the additive model is fit using `anova time poison dose`.

Interpretation of the ANOVA

The "Model" row in the ANOVA table gives a p-value for testing no differences among the population mean survival times for the 12 dose and insecticide combinations (or whatever model we fit – in this case we allow all 12 means to vary with no restriction because we fit the interaction term). The p-value of .0000 strongly suggests that the population mean survival times are not equal across all 12 groups.

The ANOVA table gives a breakdown of the Model SS into the SS for insecticide, dose, and the insecticide by dose interaction. The Mean Squares, F-statistics and p-values for testing these effects are given. The p-values indicate that the dose and insecticide effects are significant at the .0001 level. However, the F-test for no dose by insecticide interaction is not significant at the .10 level (p-value=.1123).

The cell means give us an idea about the nature of the differences among doses and insecticides (the F-tests only tell us if *some* difference appears to be there). In particular, the insecticides have noticeably different mean survival times averaged over doses, with insecticide 1 having the lowest mean survival time averaged over doses. Similarly, higher doses tend to produce lower survival times. More formal comparisons of the doses and insecticides are possible using the output from the Tukey comparisons of LS MEANS in **JMP-IN** or **SAS**, or from Fisher comparisons in **Stata**. For example, using **JMP-IN** output for the Tukey comparisons (not shown here), the high dose is significantly different from the low and medium doses, which are not significantly different from each other. We obtain Fisher's method in **Stata** using the commands

```
test _b[dose[2]]=_b[dose[1]]
test _b[dose[3]]=_b[dose[1]]
test _b[dose[3]]=_b[dose[2]]
```

These test the hypotheses $H_0 : \beta_1 = \beta_2$, $H_0 : \beta_1 = \beta_3$, and $H_0 : \beta_2 = \beta_3$ respectively. This is the same as testing $H_0 : \bar{\mu}_{.1} = \bar{\mu}_{.2}$, $H_0 : \bar{\mu}_{.1} = \bar{\mu}_{.3}$, and $H_0 : \bar{\mu}_{.2} = \bar{\mu}_{.3}$ We obtain:

```
 ( 1) - dose[1] + dose[2] = 0
       F(  1,    36) =    0.30
            Prob > F =    0.5889
 ( 1) - dose[1] + dose[3] = 0
       F(  1,    36) =    7.30
```

```
           Prob > F =    0.0104
 ( 1) - dose[2] + dose[3] = 0
       F(  1,    36) =   10.55
            Prob > F =    0.0025
```

We see that doses 1 and 2 are not significantly different from each other, but dose 3 is significantly different from doses 1 or 2, *averaged over the poison effects*. Bonferroni comparisons simply multiply the above p-values by 3 (the number of comparisons), so Bonferroni, Tukey, and Fisher all agree here.

Assuming the interaction is not important, we can obtain the three estimated pairwise differences in the three doses *using any poison* with the commands

```
lincom _b[dose[2]]-_b[dose[1]]
lincom _b[dose[3]]-_b[dose[1]]
lincom _b[dose[3]]-_b[dose[2]]
```

we obtain

```
. lincom _b[dose[2]]-_b[dose[1]]
------------------------------------------------------------------------------
        time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |      .0575    .105457     0.55   0.589    -.1563767    .2713767
------------------------------------------------------------------------------
. lincom _b[dose[3]]-_b[dose[1]]
------------------------------------------------------------------------------
        time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |      -.285    .105457    -2.70   0.010    -.4988767   -.0711233
------------------------------------------------------------------------------
. lincom _b[dose[3]]-_b[dose[2]]
------------------------------------------------------------------------------
        time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |     -.3425    .105457    -3.25   0.003    -.5563767   -.1286233
------------------------------------------------------------------------------
```

We see, for instance, that we are 95% confident that the mean difference in survival time between dose 2 and dose 3 is between 0.13 and 0.56. Put another way, beetles given dose 2 last between 1.3 and 5.6 hours longer *on average* than those given dose 3, *regardless of the insecticide used*. The last part of this statement would not hold if we had an important interaction and a more detailed analysis of *how* the difference changed with the insecticide used would be warranted.

Results from fitting the *additive model* are similar, although the confidence intervals are tighter:

```
. lincom _b[dose[2]]-_b[dose[1]]
------------------------------------------------------------------------------
        time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   -.073125   .0559247    -1.31   0.198    -.1859855    .0397355
------------------------------------------------------------------------------
. lincom _b[dose[3]]-_b[dose[1]]
------------------------------------------------------------------------------
        time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    -.34125   .0559247    -6.10   0.000    -.4541105   -.2283895
------------------------------------------------------------------------------
. lincom _b[dose[3]]-_b[dose[2]]
------------------------------------------------------------------------------
        time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   -.268125   .0559247    -4.79   0.000    -.3809855   -.1552645
------------------------------------------------------------------------------
```

More Interpretation of the Dose Effect

The interpretation of the dose and insecticide effects (called the **main-effects**) depends on whether interaction is present. The distinction is important, so I will give both interpretations to emphasize the differences. Given that the test for interaction was not significant, I would likely summarize the main effects assuming no interaction. For simplicity, I will restrict attention to the dose effect.

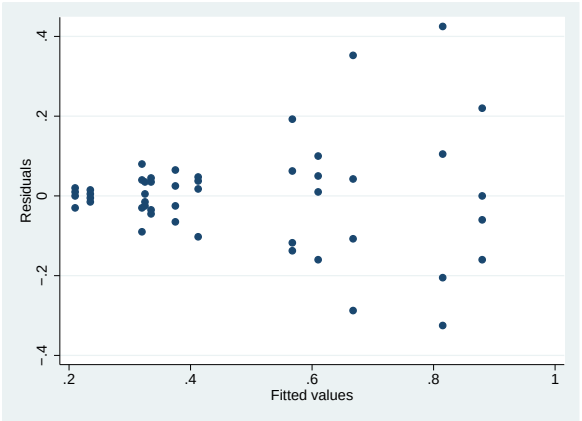
The average survival time decreases as the dose increases, with estimated mean survival times of .618, .544, and .276, respectively. If dose and insecticide **interact**, you can conclude that beetles given a high dose of the insecticide typically survive for shorter periods of time **averaged over insecticides**. You can not, in general, conclude that the highest dose yields the lowest survival time **regardless** of insecticide. For example, the difference in the medium and high dose marginal means of .544 - .276 = .268 estimates the typical decrease in survival time achieved by using the high dose instead of the medium dose, averaged over insecticides.

If the two factors interact, then the difference in mean times between the medium and high doses on a given insecticide may be significantly greater than .268, significantly less than .268, or even negative. In the latter case the medium dose would be **better** than the high dose for the given insecticide, even though the high dose gives better performance averaged over insecticides. An interaction forces you to use the cell means to decide which combination of dose and insecticide gives the best results.

If dose and insecticide **do not interact**, then the difference in marginal dose means averaged over insecticides also estimates the difference in population mean survival times between two doses, **regardless of the insecticide**. This follows from the parallel profiles definition of no interaction. Thus, the difference in the medium and high dose marginal means (.544 - .276 = .268) estimates the expected decrease in survival time anticipated from using the high dose instead of the medium dose, **regardless of the insecticide** (and hence also when averaged over insecticides).

A practical implication of no interaction is that you can conclude that the high dose is best, regardless of the insecticide used. The difference in marginal means for two doses estimates the difference in average survival expected, regardless of the insecticide.

As a final note, I will mention that the residual plot suggests that the variability in the survival times increases with increasing mean (obtained using `rvfplot`). A transformation to the reciprocal scale (which turns the response into a rate) is often suggested with these data. You should repeat the analysis on that scale to see the improvement.



An Unbalanced Two-Factor Experiment and Analysis

The sample sizes are rarely equal for the different treatments in an experiment. This has no consequence on the specification of a model, and we proceed as in the balanced case.

Example: Insulin Levels in Rats

The data below are the insulin levels in rats a certain length of time after a fixed dose of insulin was injected into their jugular or portal veins. This is a two-factor study with two vein types (jugular=1, portal=2) and three time levels (0 minutes = 1, 30 minutes = 2, and 60 minutes = 3). A feature of this experiment is that the rats used in the six vein and time combinations are distinct. I will fit a two-factor interaction model, which assumes that the responses are independent within and across treatments. The design is unbalanced, with sample sizes varying from 3 to 12.

Vein	Time	Insulin Levels											
jugular	0	18	36	12	24	43							
jugular	30	61	116	63	132	68	37						
jugular	60	18	133	33									
portal	0	96	72	34	41	98	77	120	49	92	111	99	94
portal	30	146	193	78	127	136	144	115	199	253	338		
portal	60	132	110	141	204	69	152	196	195	84	105	71	83

An alternative experimental design might randomly assign rats to the two vein groups, and then measure the insulin levels of each rat at the three time points. Depending on the questions of interest, you could compare veins using a one-way MANOVA, or a repeated measures design that allows correlated responses within rats.

The model written abstractly is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}.$$

Here, $i = 1, 2$ denote the two vein types, $j = 1, 2, 3$ denote the three times, and $k = 1, 2, \ldots, K_{ij}$ denotes the k^{th} rat out of K_{ij} in the group with vein i and time j. You should verify that $K_{11} = 5$, $K_{12} = 6$, $K_{13} = 3$, $K_{21} = 12$, $K_{22} = 10$, and $K_{13} = 12$. The `tabulate` command makes that fairly easy

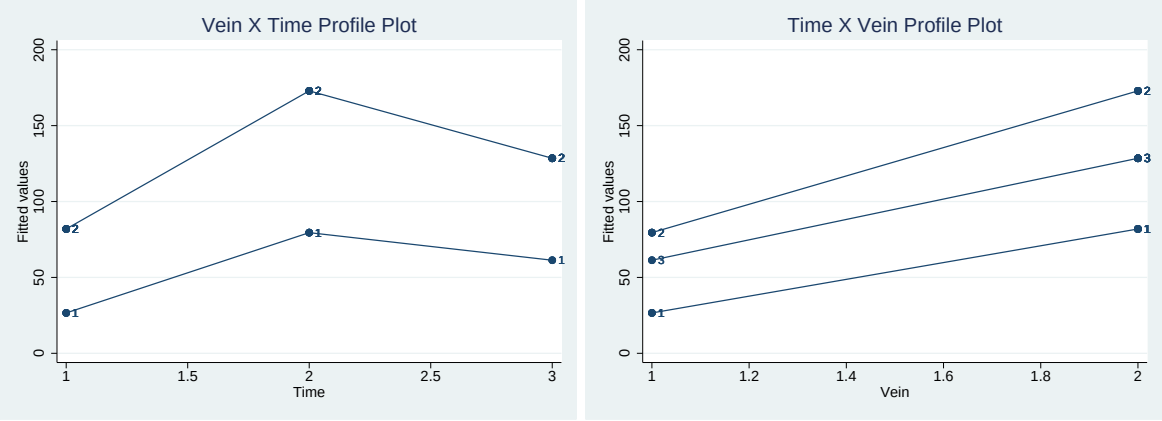
```
. tabulate vein time,summarize(insulin)
           Means, Standard Deviations and Frequencies of Insulin
           |              Time
      Vein |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |      26.6       79.5  61.333333 | 56.714286
           |  12.75931   36.44585  62.516664 | 41.899933
           |         5          6          3 |        14
-----------+---------------------------------+----------
         2 | 81.916667      172.9      128.5 | 125.11765
           | 27.747099  76.117526  49.718297 | 63.525115
           |        12         10         12 |        34
-----------+---------------------------------+----------
     Total | 65.647059    137.875  115.06667 | 105.16667
           | 35.284453   78.10239  57.218212 | 65.621848
           |        17         16         15 |        48
```

In order to get the interaction (profile) plots, we need to fit the ANOVA with interaction present.

```
. anova insulin vein time vein*time
                          Number of obs =      48     R-squared     =  0.4915
                          Root MSE      = 49.5009     Adj R-squared =  0.4310
                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  99478.4833     5  19895.6967       8.12     0.0000
                         |
                    vein |  48212.7037     1  48212.7037      19.68     0.0001
                    time |   37734.188     2   18867.094       7.70     0.0014
                vein*time |   2745.9139     2  1372.95695       0.56     0.5752
                         |
                Residual |  102914.183    42   2450.3377
              -----------+----------------------------------------------------
                   Total |  202392.667    47  4306.22695
```

It probably makes sense to look at both profile plots:

```
predict yhat
sort vein time
scatter yhat time, c(L) ml(vein) title(Vein X Time Profile Plot)
sort time vein
scatter yhat time, c(L) ml(vein) title(Vein X Time Profile Plot)
```



The profile or *interaction* plots show roughly parallel profiles indicating that the interaction term may not be important. The profile plots, along with the table of means , indicate that the insulin level is at its highest (of the three times considered) at 30 minutes for either vein considered alone, or averaged over veins. The portal vein yields a higher insulin level at any of the three time periods and averaged over the three time periods.

The ANOVA table indicates that the vein and time effects are significant, with p-values of .0001 and .0014, respectively, but that the interaction is not significant (p-value=.575). Recall that the profiles are reasonably parallel, which is consistent with a lack of interaction.

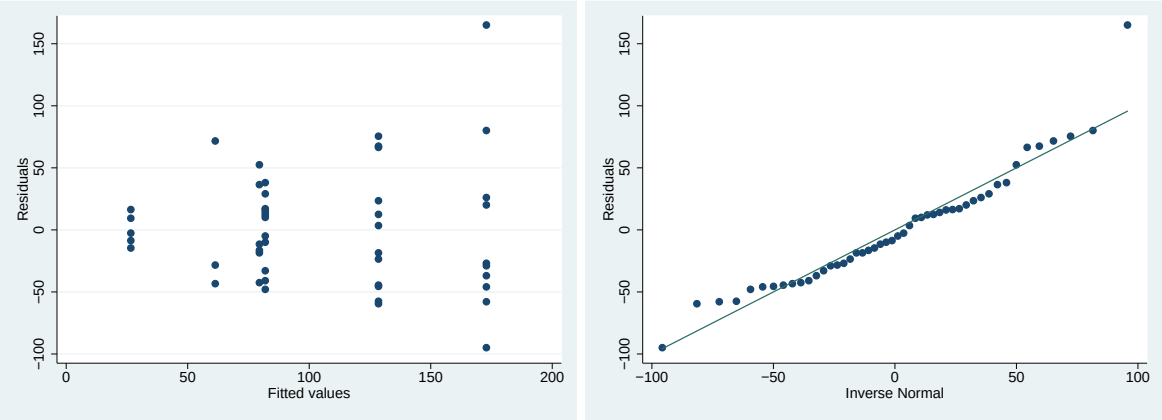
The means table above shows that the mean insulin level in the portal vein is significantly greater than the mean insulin level in the jugular vein. Because of the lack of interaction, the difference in mean levels for the portal veins is reasonably consistent across times.

Since we accept that there is no interaction here, it makes sense to compare the overall main effects *vein* and *time* using pairwise comparisons. A test that there is no difference in vein types (`test _b[vein[2]] = _b[vein[1]]`) yields a p-value of 0.0416; we reject that there is no difference. We estimate the difference in insulin levels from the portal versus the jugular veins in **Stata** using the command `lincom _b[vein[2]]-_b[vein[1]]` and find the estimate to be $\hat{\alpha}_2 - \hat{\alpha}_1 = 67.2$ with a 95% CI of (2.7, 132), *independent of time*. This is close to the estimate obtained from the marginal

means as 125 - 57 = 68, *but need not be in unbalanced designs!!* For this reason, one should always use the model estimates from **Stata** rather than estimates obtained from looking at a raw means table.

Since there is no interaction here, the difference in insulin levels is the same at time = 0, time = 30, and time = 60 minutes. Recall that when no interaction is present we say the model is *additive*. Similarly we may look at *differences* in insulin levels at the three times *independent of vein type*. The p-values for testing that there is no difference between (1) 30 minutes and 0 minutes, (2) 60 minutes and 0 minutes, and (3) 60 minutes and 30 minutes are (1) 0.0001, (2) 0.0262, and (3) 0.0423. These are Fisher values. The corresponding Bonferroni-adjusted values are obtained by multiplying each by $\begin{pmatrix} 3 \\ 2 \end{pmatrix} = 3$: (1) 0.0003 (2) 0.078, and (3) 0.13; the only significant difference is between 30 minutes and 0 minutes. Using `lincom` we would estimate this difference in insulin levels to be about 72, *independent of vein type*.

What are your thoughts on the residual plots?



Finally, note that we accept $H_0 : (\alpha\beta)_{ij} = 0$ at any reasonable significance level. We can thus fit and base inferences on the additive model if we choose to do so. The ANOVA table is

```
      Number of obs =      48     R-squared     =  0.4779
      Root MSE      = 49.0037     Adj R-squared =  0.4424
------------------------------------------------------------------------------
      Source |  Partial SS    df       MS           F     Prob > F
-------------+----------------------------------------------------------------
       Model |  96732.5694     3  32244.1898      13.43     0.0000
             |
        vein |  51594.4685     1  51594.4685      21.49     0.0000
        time |  50332.2893     2  25166.1447      10.48     0.0002
             |
    Residual |  105660.097    44  2401.36585
-------------+----------------------------------------------------------------
       Total |  202392.667    47  4306.22695
```

Now, using estimates from the additive model, we obtain the estimated mean difference in vein effects to be $\hat{\alpha}_2 - \hat{\alpha}_1 = 73.0$ with a 95% CI of (41, 105). The CI is now *smaller* than when calculated with a model that includes an interaction term!! This is good news as it provides a tighter range of plausible values and thus more powerful inference.

7 Two-factor Experiments, Continued

In the last lecture and in lab we dealt with the parameters **Stata** (and most software packages) use to fit the additive and the interaction models for two-way ANOVA. The `lincom` command was one way we learned to deal with the parameters. That probably is not the easiest approach in many problems, however. If we learn a little more about the parameters, some information is fairly immediate.

Let's continue with the insecticide problem, where we have 4 poisons and 3 doses. Remember the pattern of population cell means as follows (ignoring marginal means for now):

	Dose		
Insecticide	1	2	3
1	μ_{11}	μ_{12}	μ_{13}
2	μ_{21}	μ_{22}	μ_{23}
3	μ_{31}	μ_{32}	μ_{33}
4	μ_{41}	μ_{42}	μ_{43}

Consider the full interaction model first. The parameterization for this is $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},\ i = 1, \ldots, 4;\ j = 1, \ldots, 3$. We dodged the issue of constraints last time, but recall the problem: There are 12 real parameters (the μ_{ij}), but 20 new parameters ($1 + 4 + 3 + 12$). We need to put 8 constraints (restrictions) on these new parameters to bring us back down to 12. An old standard textbook solution to this, and one that makes the math look a lot simpler (for marginal means at least) is

$$0 = \sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij}$$

(that looks like 9 constraints but one is redundant so it is 8). Software packages like **Stata** and **SAS** use an algorithm called the sweep algorithm that makes a completely different and much more useful set of constraints more natural, though. Effectively, they start adding parameters in the model and as soon as they hit a redundant one, they set the new parameter to 0. The new constraints become

$$0 = \alpha_4 = \beta_3 = (\alpha\beta)_{41} = (\alpha\beta)_{42} = (\alpha\beta)_{43} = (\alpha\beta)_{13} = (\alpha\beta)_{23} = (\alpha\beta)_{33}$$

i.e. if there are I levels of i and J levels of j then any time i reaches level I or j reaches level J then the parameter becomes 0. It is a little easier to see 8 constraints here.

If we plug in these constraints and rewrite all 12 cell means, we see the following pattern:

	Dose		
Insecticide	1	2	3
1	$\mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$	$\mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12}$	$\mu + \alpha_1$
2	$\mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21}$	$\mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22}$	$\mu + \alpha_2$
3	$\mu + \alpha_3 + \beta_1 + (\alpha\beta)_{31}$	$\mu + \alpha_3 + \beta_2 + (\alpha\beta)_{32}$	$\mu + \alpha_3$
4	$\mu + \beta_1$	$\mu + \beta_2$	μ

At first this may not seem like much simplification, but let's examine it a bit more carefully. μ is often refereed to as the *grand mean* (this comes from the old textbook parameterization) but here we see $\mu = \mu_{43}$. The last cell of the table has become the *reference group* with all other parameters being deviations from that reference group. β_2 is the difference between doses 2 and 3 for the 4^{th} poison, β_1 is the difference between doses 1 and 3 for the 4^{th} poison. α_3 is the difference between

poisons 3 and 4 for the 3^{rd} dose, α_2 is the difference between poisons 2 and 4 for the 3^{rd} dose, and α_1 is the difference between poisons 1 and 4 for the 3^{rd} dose. The difference between poisons 3 and 4 for the 2^{nd} dose is $\alpha_3 + (\alpha\beta)_{32}$ rather than just α_3 (if the difference does not depend on poison then there is no interaction).

With these constraints then $(\mu_{32} - \mu_{42}) - (\mu_{33} - \mu_{43}) = (\alpha\beta)_{32}$. Recall last lecture we said that **no** interaction (parallel profiles in an interaction plot) was this: If i, i', j, j' are legal indexes, then $\mu_{ij} - \mu_{ij'} = \mu_{i'j} - \mu_{i'j'}$, which is to say the difference between doses j and j' is the same for insecticide i as for insecticide i'; and $\mu_{ij} - \mu_{i'j} = \mu_{ij'} - \mu_{i'j'}$, which is to say the difference between insecticides i and i' is the same for dose j as it is for dose j'. Last week we saw that $[\mu_{ij} - \mu_{ij'}] - [\mu_{i'j} - \mu_{i'j'}] = [(\alpha\beta)_{ij} - (\alpha\beta)_{ij'}] - [(\alpha\beta)_{i'j} - (\alpha\beta)_{i'j'}]$. The constraints that $0 = (\alpha\beta)_{Ij} = (\alpha\beta)_{iJ}$ allow us to simplify greatly the `lincom` command for many such terms. Since $(\mu_{32} - \mu_{42}) - (\mu_{33} - \mu_{43}) = (\alpha\beta)_{32}$ *(only for these constraints, though!)* then a simple `lincom(_b[poison[3]*dose[2]])` gets that term. Better yet, we can have it automatically printed.

Stata Implementation

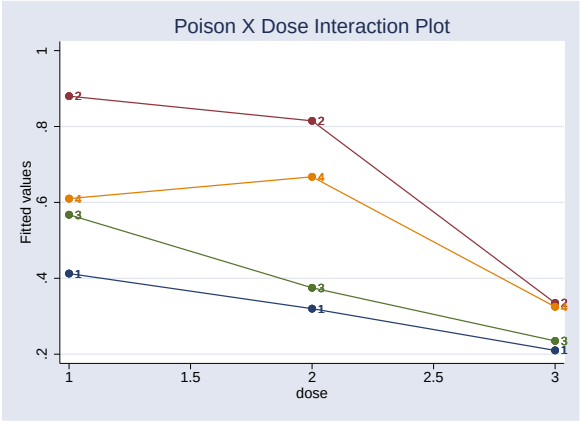
Anova problems actually are specialized regression problems (we will grapple with this idea later). What we want are regression estimates of all the effects $(\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij})$. The `regress` option with `anova` gets that for us in a form that matches the `lincom` syntax. This can be treated as a post-estimation command, i.e. after issuing the `anova time poison dose poison*dose` command (and examining the the ANOVA table, interaction plots, etc.) just type another command `anova,regress` to get the following results

```
. anova, regress
      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F( 11,    36) =    9.01
       Model |  2.20435628    11  .200396025           Prob > F      =  0.0000
    Residual |   .800724989   36  .022242361           R-squared     =  0.7335
-------------+------------------------------           Adj R-squared =  0.6521
       Total |  3.00508126    47  .063937899           Root MSE      =  .14914
------------------------------------------------------------------------------
        time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |        .325   .0745694     4.36   0.000     .1737663    .4762337
      poison |
          1  |       -.115    .105457    -1.09   0.283    -.3288767    .0988767
          2  |        .01    .105457     0.09   0.925    -.2038767    .2238767
          3  |       -.09    .105457    -0.85   0.399    -.3038767    .1238767
          4  |  (dropped)
      dose |
          1  |       .285    .105457     2.70   0.010     .0711233    .4988767
          2  |      .3425    .105457     3.25   0.003     .1286233    .5563767
          3  |  (dropped)
poison*dose |
        1  1 |     -.0825   .1491387    -0.55   0.584    -.3849674    .2199674
        1  2 |     -.2325   .1491387    -1.56   0.128    -.5349673    .0699674
        1  3 |  (dropped)
        2  1 |       .26   .1491387     1.74   0.090    -.0424673    .5624674
        2  2 |     .1375   .1491387     0.92   0.363    -.1649673    .4399674
        2  3 |  (dropped)
        3  1 |      .0475   .1491387     0.32   0.752    -.2549674    .3499674
        3  2 |     -.2025   .1491387    -1.36   0.183    -.5049673    .0999674
        3  3 |  (dropped)
        4  1 |  (dropped)
        4  2 |  (dropped)
        4  3 |  (dropped)
------------------------------------------------------------------------------
```

Let's try to make sense of these coefficients by relating them both to the table of sample cell means and to an interaction plot from the last lecture. First, try to reconstruct the sample cell means from the coefficients. Recall that the full interaction model imposes no restrictions.

Various questions arise. Why is poison highly significant yet none of the poison coefficients is significant? – Get your answer from the interaction plot and what these coefficients are estimating. One of the large differences appears to be between poisons 1 and 2 at dose level 2. How would you estimate that difference? How would you test for specific interactions (i.e. different slopes in the above plot?). We will spend some time examining all this.

	Dose			
Insecticide	1	2	3	Insect marg
1	.413	.320	.210	.314
2	.880	.815	.335	.677
3	.568	.375	.235	.393
4	.610	.668	.325	.534
Dose marg	.618	.544	.277	.480



The Additive Model

Since interaction was not significant (there is not much data so this *might* just be poor power) we should see how all this looks when we fit the no-interaction (additive) model. This is a highly restricted model and we will not reproduce all the sample cell means from this model. Now we fit $\mu_{ij} = \mu + \alpha_i + \beta_j$ using the command, with ensuing results given below:

```
. anova time poison dose
                          Number of obs =      48     R-squared     =  0.6503
                          Root MSE      = .158179     Adj R-squared =  0.6087
                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  1.95421877     5  .390843755      15.62     0.0000
                         |
                  poison |   .921206282     3  .307068761      12.27     0.0000
                    dose |  1.03301249     2  .516506246      20.64     0.0000
                         |
                Residual |  1.05086249    42  .025020536
              -----------+----------------------------------------------------
                   Total |  3.00508126    47  .063937899
```

All constraints are as previously described, and easily seen from the following:

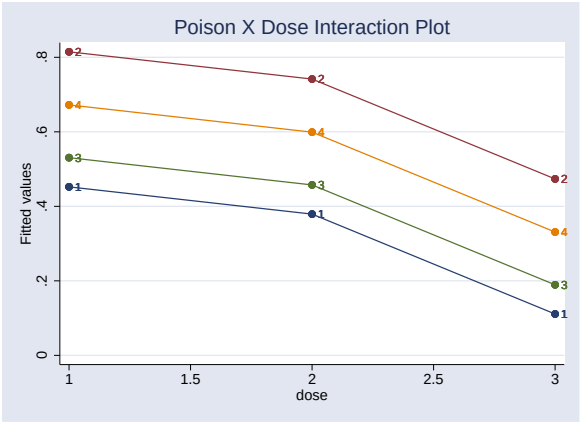
```
. anova,regress
      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  5,    42) =   15.62
       Model |  1.95421877     5  .390843755           Prob > F      =  0.0000
    Residual |  1.05086249    42  .025020536           R-squared     =  0.6503
-------------+------------------------------           Adj R-squared =  0.6087
```

```
       Total |  3.00508126    47  .063937899           Root MSE      =  .15818
------------------------------------------------------------------------------
        time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .3310417   .0559247     5.92   0.000     .2181811    .4439022
      poison |
          1 |        -.22   .0645762    -3.41   0.001    -.3503201   -.0896799
          2 |       .1425   .0645762     2.21   0.033     .0121799    .2728201
          3 |  -.1416667   .0645762    -2.19   0.034    -.2719868   -.0113466
          4 |  (dropped)
       dose |
          1 |     .34125   .0559247     6.10   0.000     .2283895    .4541105
          2 |    .268125   .0559247     4.79   0.000     .1552645    .3809855
          3 |  (dropped)
------------------------------------------------------------------------------
```

What we have fit now is the much simpler structure for population cell means (all the parameters have very easy interpretations – what are they?):

	Dose		
Insecticide	1	2	3
1	$\mu+\alpha_1+\beta_1$	$\mu+\alpha_1+\beta_2$	$\mu+\alpha_1$
2	$\mu+\alpha_2+\beta_1$	$\mu+\alpha_2+\beta_2$	$\mu+\alpha_2$
3	$\mu+\alpha_3+\beta_1$	$\mu+\alpha_3+\beta_2$	$\mu+\alpha_3$
4	$\mu+\beta_1$	$\mu+\beta_2$	μ

You should confirm that you no longer reproduce the sample cell means $\bar{y}_{ij.}$ from the estimated regression coefficients, but you do reproduce the sample marginal means $\bar{y}i..$ and $\bar{y}.j.$. We can look at an interaction plot of predicted cell means, but note that we have *forced* it to look this way.



Estimable Functions

What we have been covering is expanded upon at considerable length in the **SAS** manual and many textbooks under the topic of *estimable functions*. This is a fairly advanced topic, but the gist of it is that only linear combinations of *population cell means* can legally be estimated (things of the form $\sum_i \sum_j c_{ij}\mu_{ij}$ for some constants c_{ij} – we have been using 0, 1 and -1 as constants). Anything we estimate or test has to be a linear combination of population cell means. In particular, μ and α_i, for instance, are not estimable since there is ambiguity about what they are until constraints are put on them. Different constraints give different interpretations. What we have been doing is relating everything back to the μ_{ij} in order to keep it all "legal". I can run an analysis on two different packages (or even the same package with different options) and get considerably different estimates of α_1 reported. As long as I stick to estimable functions, though, I always get the same estimate.

Parameters in the One-Way Problem

A look back at the one-way ANOVA problem shows the constraints can make some things simpler there too. In that case we have a model $y_{ij} = \mu_i + \epsilon_{ij};\ i = 1, \ldots, I;\ j = 1, \ldots, n_i$. Let's do the same type of decomposition of μ_i, i.e. $\mu_i = \mu + \alpha_i$, with the same problem that we have I "real" group means and now $I + 1$ new parameters. We need a constraint, and the one **Stata** and **SAS** impose is $\alpha_I = 0$ (set the first redundant parameter to 0).

What is the implication? Now $\mu = \mu_I$ and $\alpha_i = \mu_i - \mu_I$, i.e. the *last* group has become a reference group and all the parameter estimates (the α_i) are deviations from this reference group. In the CHDS example this means we could get mostly for free the t-tests comparing non-smokers to heavy smokers and light smokers to heavy smokers. It might be more convenient to reorder so that nonsmokers were last, so that the easy tests would compare the two smoking groups to nonsmokers. The `lincom` command can fill in the last comparison in either case ($\mu_1 - \mu_2 = \alpha_1 - \alpha_2$). The idea of estimable functions still applies; we need to make sure we are looking at linear combinations of the means.

Unbalanced Data

Returning to the two-way problem, we wrote a model

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk};\ \ i = 1, 2, \ldots, I;\ \ j = 1, 2, \ldots, J;\ \ k = 1, 2, \ldots, K$$

where the sample size in each cell (i, j) was the same number K. Unbalanced data allow the sample size to vary with cell, so now

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk};\ \ i = 1, 2, \ldots, I;\ \ j = 1, 2, \ldots, J;\ \ k = 1, 2, \ldots, K_{ij}$$

What difference does it make? — In practice, not much. The reason the topic gets mentioned is that there are disagreeable aspects to unbalanced designs. There are multiple ways to formulate SSs and F-tests. **SAS** provides 4, Types I-IV, and **Stata** provides the same as **SAS** Types I and III. With balanced data all the types agree, but for unbalanced data they do not all agree. The only reason this is not much of a practical problem is that most analysts use Type III (**Stata**'s default) and don't anguish much over it. The t-tests on coefficients are not obviously affected (i.e. `lincom` results), although comparing main effects in the presence of interaction is a subtle business in unbalanced designs (the preferred approach being least squares means, and Stata makes those awkward to get).

Let's return to the unbalanced example of rat insulin levels from the last lecture. The ANOVA table indicates that the vein and time effects are significant, with p-values of .0001 and .0014, respectively, but that the interaction is not significant (p-value=.575). Recall that the jugular and portal profiles are reasonably parallel, which is consistent with a lack of interaction. Looking at the estimates below, let us figure out as an in-class exercise how to interpret the various coefficient estimates, and how to test for significance of the important effects. We really ought to see if we can transform to a scale where the residuals look better, too.

```
. anova,regress
      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  5,    42) =    8.12
       Model |  99478.4833     5  19895.6967           Prob > F      =  0.0000
    Residual |  102914.183    42   2450.3377           R-squared     =  0.4915
-------------+------------------------------           Adj R-squared =  0.4310
       Total |  202392.667    47  4306.22695           Root MSE      =  49.501
------------------------------------------------------------------------------
     insulin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |       128.5   14.28967     8.99   0.000     99.66227    157.3377
vein         |
          1  |   -67.16667   31.95268    -2.10   0.042    -131.6498   -2.68354
          2  |   (dropped)
             |
time         |
          1  |   -46.58333   20.20865    -2.31   0.026    -87.36604   -5.800623
          2  |        44.4   21.19501     2.09   0.042     1.626732    87.17327
          3  |   (dropped)
             |
vein*time    |
        1  1 |       11.85   41.41541     0.29   0.776    -71.72969    95.42969
        1  2 |   -26.23333    40.9194    -0.64   0.525     -108.812    56.34536
        1  3 |   (dropped)
        2  1 |   (dropped)
        2  2 |   (dropped)
        2  3 |   (dropped)
------------------------------------------------------------------------------

. anova insulin vein time
                           Number of obs =      48     R-squared     =  0.4779
                           Root MSE      = 49.0037     Adj R-squared =  0.4424
                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  96732.5694     3  32244.1898      13.43     0.0000
                         |
                    vein |  51594.4685     1  51594.4685      21.49     0.0000
                    time |  50332.2893     2  25166.1447      10.48     0.0002
                         |
                Residual |  105660.097    44  2401.36585
              -----------+----------------------------------------------------
                   Total |  202392.667    47  4306.22695

. anova,regress
      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  3,    44) =   13.43
       Model |  96732.5694     3  32244.1898           Prob > F      =  0.0000
    Residual |  105660.097    44  2401.36585           R-squared     =  0.4779
-------------+------------------------------           Adj R-squared =  0.4424
       Total |  202392.667    47  4306.22695           Root MSE      =  49.004
------------------------------------------------------------------------------
     insulin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |    129.6685   13.03897     9.94   0.000     103.3902    155.9468
vein         |
          1  |   -73.00912   15.75087    -4.64   0.000    -104.7529   -41.26531
          2  |   (dropped)
             |
time         |
          1  |   -42.54816   17.42256    -2.44   0.019    -77.66102   -7.435306
          2  |    35.58493   17.82622     2.00   0.052    -.3414601    71.51132
          3  |   (dropped)
------------------------------------------------------------------------------
```

To see one difference with the unbalanced design, consider the following two ANOVA tables; the first is the usual one, the second is an optional one. Note the differences for SS of main effects. If the data were balanced, the default (SAS Type III SS) and the sequential (SAS Type I SS) would be the same. The second form even depends upon the order terms are entered into the model.

```
. anova insulin vein time vein*time
                          Number of obs =      48     R-squared     =  0.4915
                          Root MSE      = 49.5009     Adj R-squared =  0.4310
                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  99478.4833     5  19895.6967       8.12     0.0000
                         |
                    vein |  48212.7037     1  48212.7037      19.68     0.0001
                    time |   37734.188     2   18867.094       7.70     0.0014
                vein*time |   2745.9139     2  1372.95695       0.56     0.5752
                         |
                Residual |  102914.183    42   2450.3377
              -----------+----------------------------------------------------
                   Total |  202392.667    47  4306.22695
. anova insulin vein time vein*time,seq
                          Number of obs =      48     R-squared     =  0.4915
                          Root MSE      = 49.5009     Adj R-squared =  0.4310
                  Source |    Seq. SS     df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  99478.4833     5  19895.6967       8.12     0.0000
                         |
                    vein |  46400.2801     1  46400.2801      18.94     0.0001
                    time |  50332.2893     2  25166.1447      10.27     0.0002
                vein*time |   2745.9139     2  1372.95695       0.56     0.5752
                         |
                Residual |  102914.183    42   2450.3377
              -----------+----------------------------------------------------
                   Total |  202392.667    47  4306.22695
```

Regression on Dummy Variables: Stata's `xi`

The way `anova` works is that it creates special variables called indicator or *dummy* variables for categorical variables and performs regression on them. One dummy variable is created for each level of a categorical variable and has a value of 1 if the observation has that level of the category else the value is 0. We do not need to worry much about this if we can use the `anova` command, but if we want to do all this in logistic regression we have to get explicit about it. All the big statistics packages do this, and if we were using SAS I could hide it all from you, but Stata requires you to learn about it. The `xi` facility in Stata is one we will need for many problems.

Following is the insecticide data analyzed as a regression problem using xi:

```
. xi: regress time i.poison i.dose i.poison*i.dose
i.poison          _Ipoison_1-4        (naturally coded; _Ipoison_1 omitted)
i.dose            _Idose_1-3          (naturally coded; _Idose_1 omitted)
i.poi~n*i.dose    _IpoiXdos_#_#       (coded as above)
      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F( 11,    36) =    9.01
       Model |  2.20435628    11  .200396025           Prob > F      =  0.0000
    Residual |  .800724989    36  .022242361           R-squared     =  0.7335
-------------+------------------------------           Adj R-squared =  0.6521
       Total |  3.00508126    47  .063937899           Root MSE      =  .14914
------------------------------------------------------------------------------
        time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Ipoison_2 |      .4675    .105457     4.43   0.000     .2536233    .6813767
  _Ipoison_3 |       .155    .105457     1.47   0.150    -.0588767    .3688767
  _Ipoison_4 |      .1975    .105457     1.87   0.069    -.0163767    .4113767
    _Idose_2 |     -.0925    .105457    -0.88   0.386    -.3063767    .1213767
    _Idose_3 |     -.2025    .105457    -1.92   0.063    -.4163767    .0113767
  _Ipoison_2 |  (dropped)
  _Ipoison_3 |  (dropped)
  _Ipoison_4 |  (dropped)
```

```
    _Idose_2 |  (dropped)
    _Idose_3 |  (dropped)
_IpoiXdo~2_2 |      .0275   .1491387     0.18   0.855    -.2749674    .3299674
_IpoiXdo~2_3 |     -.3425   .1491387    -2.30   0.028    -.6449674   -.0400326
_IpoiXdo~3_2 |        -.1   .1491387    -0.67   0.507    -.4024674    .2024674
_IpoiXdo~3_3 |       -.13   .1491387    -0.87   0.389    -.4324674    .1724674
_IpoiXdo~4_2 |        .15   .1491387     1.01   0.321    -.1524674    .4524674
_IpoiXdo~4_3 |     -.0825   .1491387    -0.55   0.584    -.3849674    .2199674
       _cons |      .4125   .0745694     5.53   0.000     .2612663    .5637337
------------------------------------------------------------------------------
```

You should do this and then examine the variables Stata has placed in the data set. This does not give us the tests from the ANOVA table. The test on interaction is easy:

```
. testparm _IpoiXdo*
 ( 1)  _IpoiXdos_2_2 = 0
 ( 2)  _IpoiXdos_2_3 = 0
 ( 3)  _IpoiXdos_3_2 = 0
 ( 4)  _IpoiXdos_3_3 = 0
 ( 5)  _IpoiXdos_4_2 = 0
 ( 6)  _IpoiXdos_4_3 = 0
       F(  6,    36) =    1.87
            Prob > F =    0.1123
```

but the tests on main effects (equality of marginal means) are a lot less obvious (don't worry, you won't have to do it this way for anova problems). Because this is such a mess, we probably would not test for main effects if interaction were present if we had to go through these steps. With no interaction the procedure is just like the test above.

```
 . test _Ipoison_2 + (_IpoiXdos_2_2+_IpoiXdos_2_3)/3 = _Ipoison_3 + (_IpoiXdos_3_2+_IpoiXdo
> s_3_3)/3 = _Ipoison_4 + (_IpoiXdos_4_2+_IpoiXdos_4_3)/3 =0
 ( 1)  _Ipoison_2 - _Ipoison_3 + .3333333 _IpoiXdos_2_2 + .3333333 _IpoiXdos_2_3 - .333333
> 3 _IpoiXdos_3_2 - .3333333 _IpoiXdos_3_3 = 0
 ( 2)  _Ipoison_2 - _Ipoison_4 + .3333333 _IpoiXdos_2_2 + .3333333 _IpoiXdos_2_3 - .333333
> 3 _IpoiXdos_4_2 - .3333333 _IpoiXdos_4_3 = 0
 ( 3)  _Ipoison_2 + .3333333 _IpoiXdos_2_2 + .3333333 _IpoiXdos_2_3 = 0
       F(  3,    36) =   13.81
            Prob > F =    0.0000

 . test _Idose_2+(_IpoiXdos_2_2+_IpoiXdos_3_2+_IpoiXdos_4_2)/4 =_Idose_3+(_IpoiXdos_2_3+_I
> poiXdos_3_3+_IpoiXdos_4_3)/4=0
 ( 1)  _Idose_2 - _Idose_3 + .25 _IpoiXdos_2_2 - .25 _IpoiXdos_2_3 + .25 _IpoiXdos_3_2 - .
> 25 _IpoiXdos_3_3 + .25 _IpoiXdos_4_2 - .25 _IpoiXdos_4_3 = 0
 ( 2)  _Idose_2 + .25 _IpoiXdos_2_2 + .25 _IpoiXdos_3_2 + .25 _IpoiXdos_4_2 = 0
       F(  2,    36) =   23.22
            Prob > F =    0.0000
```

Annoyingly, Stata has changed constraints on us (now the first level of a categorical variable gets zeroed out). We can set what level gets zeroed out (and thus becomes the reference level) with the `char` command

```
. char poison[omit] 4
. char dose[omit] 3
. xi: regress time i.poison i.dose i.poison*i.dose
```

Execute these commands and confirm the original parameters from `anova,regress` are reproduced. We rarely use `xi` and explicit regression on dummy variables in anova problems. We will need it with logistic regression, though.

8 Analysis of Covariance

Let us recall our previous one-way ANOVA problem, where we compared the mean birth weight (weight) for children in three groups defined by the mother's smoking habits. The three groups had mothers that did not smoke during pregnancy (group 1), mothers that smoked a pack or less of cigarettes per day during their pregnancy (group 2), and mothers that smoked more than one pack of cigarettes per day during their pregnancy (group 3). We concluded that children born to non-smoking mothers were, on average, heavier than children born to mothers in the two smoking groups (and there was no significant difference in birth weights between the two smoking groups).

A deficiency with the analysis is that the differences among groups may be due to other factors that could not be controlled; for example, the mother's intake of caffeine, the mother's pre-pregnancy weight (mweight), and so on. This, of course, is a standard problem with observational studies. If the primary interest is to assess the potential effect of mother's smoking on birth weight, then a proper analysis would need to account for the possible effect of these other features on birth weight. For simplicity, I will consider an analysis that accounts, or adjusts, for the effect of mother's pre-pregnancy weight (mweight) when assessing the effect of smoking. We will see how to adjust for more effects as well.

Let weight_{ij} be the birth weight for the j^{th} child born to a mother in group i ($i = 1, 2, 3$) with pre-pregnancy weight mweight_{ij}. The statistical technique for comparing weights across groups, adjusting for mother's mweight, is called the analysis of covariance (ANCOVA), and is based on the model:

$$\text{weight}_{ij} = \mu + \alpha_i + \beta\,\text{mweight}_{ij} + \epsilon_{ij},$$

where μ is a "grand mean", α_i is the i^{th} group effect, and β is a regression effect. If $\beta = 0$ this is the standard one-way ANOVA model for comparing weights across smoking groups. In words:

$$\text{weight} = \text{Grand Mean} + \text{Group Effect} + \text{mweight Effect} + \text{Residual}.$$

The ANCOVA model implies that the relationship between the mean weight and mother's mweight is linear in each group, but that the regression lines for the groups have different intercepts (and equal slopes). The intercept for group i is $\mu + \alpha_i$. Figure 1 illustrates one possible realization of the model (PPW is mweight).

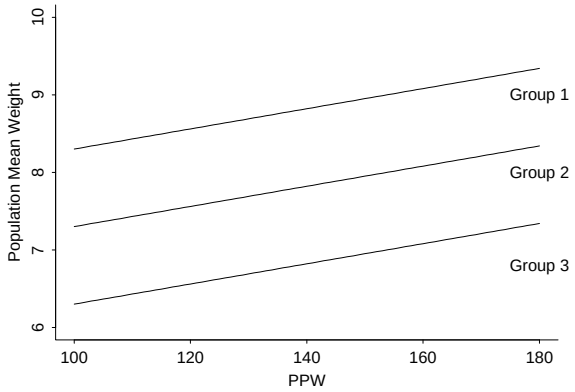


Figure 1: Possible population regression lines for ANCOVA model

Primary interest is in testing the hypothesis of no group effects, which is equivalent to testing that the intercepts of the population regression lines are equal: $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$. If H_0 is true then the relationship between weight and mweight does not depend on the smoking group to which the mother belongs, that is, there is no effect of mother's smoking on the child's weight, after adjusting for mweight (by including mweight in the model).

Fitting the ANCOVA Model in Stata

ANCOVA is a hybrid of ANOVA and Regression. In **Stata** both the `anova` and `regress` commands assume a continuous response (dependent or y-variable); with `regress` all predictors are continuous, with `anova` all predictors are by default categorical (and a separate indicator variable is created for each level of each predictor). ANCOVA is implemented most easily using the `anova` command (or by using `xi: regress`), but you need to specify what is continuous and what is categorical. The `regress` form is more awkward but is needed when we move to logistic regression. Continuous predictors are known as **covariates**.

Recall the syntax for one-way ANOVA was `anova weight ms_gp` where `weight` is child's birth weight and `ms_gp` is mother's smoking group (Note I coded group as 0,1,2 in the example. This actually serves to illustrate a point, because Stata will in subsequent analysis decide to recode groups as 1-3. I don't want you surprised by this). Page 1 of the **Stata** output has the analysis for this one-way problem. In order to adjust for mweight (Maternal pre-pregnancy weight) the usual method in **Stata** is the command `anova weight ms_gp mweight,cont(mweight)` where the option "`,cont(mweight)`" tells `anova` that `mweight` is continuous. Everything not listed as continuous is assumed to be categorical. You may also use the syntax `anova weight Smoke mweight,cat(Smoke)` to list the categorical variables, with remaining variables treated as continuous.

Page 2 of the **Stata** output begins the summaries of the ANCOVA model. At this time, I will not worry about whether the model fits the data. The F-test for the model gives a p-value for testing no significant effects in the model. The p-value of 0.0000 strongly suggests that either the smoking groups or mweight, or both, have an effect on the weight. The p-values for smoking group and mweight also are both 0.0000, indicating that the group and mweight effects are significant. In particular, there are significant differences in the intercepts of the population regression lines, or put another way, there are significant differences in the mean weights of the three groups defined by mother's smoking habits, **after adjusting for the effect of mweight.** Figure 2 shows the predicted values for the groups from this analysis.

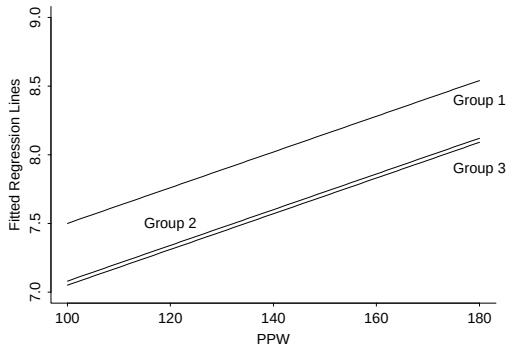


Figure 2: Fitted regression lines for ANCOVA model

A Little More Explanation of the Model

To better understand why ANCOVA is preferred to the one-way ANOVA on birth weights, suppose for argument's sake that weight is strongly positively related to mweight. If smoking behavior is strongly related to mother's mweight, then differences in the mean weights for the three groups could be due solely to differences in mother's mweight. For example, consider the hypothetical population in Figure 3 where I have plotted the relationship between the mean weight and mother's mweight in each group. Suppose that any data collected from these populations falls exactly on the lines in the plot.

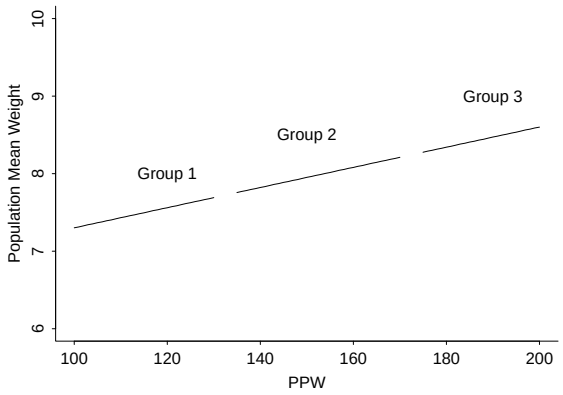


Figure 3: More possible population regression lines for ANCOVA model

A simple linear regression model relating weight to mweight, with no group effects, is appropriate, yet the distributions for weights, ignoring groups, would differ dramatically. The regression line suggests that you would have identical mean weights for mothers in the different smoking groups, if mothers with a fixed weight could be compared across groups.

A one-way ANOVA comparing smoking groups, ignoring mother's mweight would find significant differences in weight across groups. The ANCOVA would conclude, appropriately, that there are no differences across groups once mweight was taken into account. (A more fundamental question is whether these groups, as drawn, are even comparable!)

Interpreting Parameter Estimates

We have already tackled most of the confusing issues with parameter estimates in the two-way ANOVA problem. We have the same issues of constraints here (because we use more model parameters than we actually have "real" parameters). We have 3 intercepts and one slope; we have 1 μ, 3 α_i s, and one β in the model. One of the α_i s is redundant, and **Stata** will take care of that by setting the last one to 0. Again, that is just one possible solution, and different software packages can make different decisions.

We have three lines, and we want the equations for all three. We have forced parallel lines, so the slope is the same for each smoking group, .0118683. The intercept for the i^{th} smoking group has been modelled as $\mu + \alpha_i$, but α_3 has been constrained to 0, so

$$\text{Intercept group } 0 \quad = \quad \text{Constant for model} + \text{gp[0] effect}$$

$$\begin{aligned} &= 5.7517 + 0.4477 = 6.1994 \\ \text{Intercept group 1} &= \text{Constant for model} + \text{gp[1] effect} \\ &= 5.7517 + .0318 = 5.7835 \\ \text{Intercept group 2} &= \text{Constant for model} = 5.7517 \end{aligned}$$

The fitted relationships are

$$\begin{aligned} \text{Predicted weight} &= 6.1994 + .00187 \text{ mweight, for group 0} \\ &= 5.7835 + .00187 \text{ mweight, for group 1} \\ &= 5.7517 + .00187 \text{ mweight, for group 2} \end{aligned}$$

A plot of the fitted relationships is given in Figure 2.

Group Differences

Group differences in the ANCOVA model are differences in intercepts, i.e. vertical distances between the lines. Using a Bonferroni criterion, we see significant differences between Groups 0 and 2 (from `_b[ms_gp[1]]`), and Groups 0 and 1 (from the `lincom (_b[ms_gp[1]]-_b[Smoke[2]])` command), but no significant difference between Groups 1 and 2 (from `_b[Smoke[2]]`). (Note that Stata has renumbered groups as 1, 2, 3 even though we had made them 0, 1, 2 — I apologize for making this confusing, but this "feature" could catch you off guard someday.) This is very similar to what we found in the one-way problem both in direction and size of all effects. Adjusting for mweight has not given us all that much insight here, but needed to be done in order to believe the group differences were "real". The slope of the line has the same interpretation as in regression (what *is* the interpretation?), but is not of primary interest here.

Checking the ANCOVA Model

I have explained the basics of ANCOVA, without considering whether this model describes the CHDS data. The ANCOVA model constrains the slopes of the regression lines for the 3 groups to be identical, so we should check if this is sensible. We can fit three completely different lines (to see if we do any better than by forcing them to be parallel) by adding a smoking group by mweight interaction (crossed effect) to the ANCOVA model.

Pages 4-5 of the **Stata** output show results, as well as three separate linear regressions for the three groups. Let us spend some time in class making sure we see what the parameter constraints have done for us, and that the resulting lines are identical to separate linear regressions.

The p-value of .2491 for the test of no Smoke group * mweight interaction indicates there is no significant improvement by allowing different slopes, so the original ANCOVA model looks reasonable in that regard. The plot of all three fitted lines does not suggest much difference in slopes. The residual plot was not very suggestive of problems – there appears to be little difficulty with this model.

You should note that there is a huge disadvantage to needing a Group*covariate interaction term. With parallel lines we know what group differences mean – the distances between lines. With non-parallel lines the distance between lines depends upon what value of the covariate you are considering, and that distance thus *can be anything you want it to be.* Confirm this with a plot.

Extending the Analysis of Covariance

In the CHDS study, there are several possible effects in addition to mother's pre-pregnancy weight (mweight) that we may wish to consider when assessing whether mother's smoking impacts a child's

birth weight. For example, the mother's height and age, and the gestation length, may be important features to account for in the analysis.

The natural way to account for each effect is through a multiple regression model with a group effect:

$$\text{weight}_{ij} = \mu + \alpha_i + \beta_1 \text{mweight}_{ij} + \beta_2 \text{AGE}_{ij} + \beta_3 \text{HT}_{ij} + \beta_4 \text{GL}_{ij} + \epsilon_{ij}.$$

As before, weight_{ij} is the birth weight for the j^{th} child born to a mother in group i ($i = 1, 2, 3$) with pre-pregnancy weight mweight_{ij}, age AGE_{ij}, height HT_{ij}, and gestation length GL_{ij}.

The multiple regression model has 4 four predictors (mweight, AGE, HT, GL) and 1 factor (groups). The model assumes that the effect of each predictor is the same across groups, leading to a multiple regression model with identical regression effects for each predictor in each group. In words:

$$\text{weight} = \text{Grand Mean} + \text{Group Effect} + \text{mweight Effect} + \text{Age Effect}$$

$$+\text{HT Effect} + \text{GL Effect} + \text{Residual}.$$

A primary interest is testing the hypothesis of no group effects: $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$. If H_0 is true then the relationship between weight and mweight, AGE, HT, and GL does not depend on the smoking group to which the mother belongs, that is, there is no effect of mother's smoking on the child's weight, after adjusting for mweight, AGE, HT, and GL (by including them as predictors in the model). More generally, the model could include other factors (group variables) or predictors.

Stata Implementation

Six variables are needed to fit this model in Stata: child's weight (continuous), smoking group (1-3, categorical), mother's mweight (continuous), mother's age (continuous), mother's height (continuous), and gestation length (continuous). The command
`anova weight Smoke mweight mage mheight gest,cat(ms_gp)` fits the model. Results are on p. 6 of the output.

The F-tests show all effects but age are important in this model. Confirm that intercepts are -5.838, -6.159, and -6.244 for, respectively, groups 0, 1, and 2. What is of interest of course is the actual group differences. Notice that they are smaller than when ignoring covariates altogether (group differences are the `_b[ms_gp[i]]` terms as before).

Further Thoughts

The approach that I have taken here is consistent with the way epidemiologists assess the impact of a risk factor on a response, adjusting for the effects of confounders. In our analysis, the response is child's weight, mother's smoking habits play the role of a risk factor, and the other features play the roles of confounders (even if they are not strictly so).

A sensible further step in the analysis would be to eliminate, one at a time, the unimportant predictors of weight (i.e. backward elimination). This is easily automated. Once AGE is omitted, the remaining effects are significant at the 1% level. Furthermore, the differences for the smoking groups are nearly identical to those obtained with the previous model, so omitting AGE has little impact on our conclusions. Another reasonable question to examine is whether the smoking groups interact with any of the predictors in the model.

I will note that epidemiologists often adjust for all confounders (at least all that they have measured), regardless of their statistical significance.

Two Simulated Examples

Just to see how much difference covariates can make, I simulated two extreme examples. In the first, the group effect is not at all significant using a simple one-way ANOVA, yet introducing a covariate makes the group effect extremely significant. In this case the covariate is crucial for finding group differences.

In the second example, group differences seem clear from a one-way ANOVA, yet disappear completely when the covariate is introduced. Differences in covariate values completely explain apparent group differences.

We will discuss these examples in class.

Adjusted means

It is very common to report adjusted means in ANCOVA problems. Let us consider the second simulated data set. The table of (raw) means is obtained as follows:

```
. tabstat y,by(group) stat(mean semean)
 Summary for variables: y
     by categories of: group
   group |      mean  se(mean)
---------+--------------------
       1 |  4.472125  .1028853
       2 |   5.50423  .1216309
       3 |  6.472919  .0959465
---------+--------------------
   Total |  5.483091  .1203481
------------------------------
```

We can obtain a table of means adjusted for the covariate x as follows:

```
. adjust x, by(group) se
------------------------------------------------------------
     Dependent variable: y     Command: anova
  Covariate set to mean: x = 3.4761906
------------------------------------------------------------
---------------------------------
    group |         xb        stdp
----------+----------------------
        1 |    5.04042   (.227786)
        2 |    5.50423   (.101776)
        3 |    5.90462   (.227785)
---------------------------------
     Key:  xb    =  Linear Prediction
           stdp  =  Standard Error
```

What is the `adjust` command doing? Match its results with the following:

```
. tabstat x
    variable |      mean
-------------+----------
           x |   3.47619
-------------------------

. lincom(_b[_cons] + _b[group[1]]+_b[x]*3.47619)
 ( 1)  _cons + group[1] + 3.47619 x = 0
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   5.040422   .2277854    22.13   0.000     4.584625     5.49622
------------------------------------------------------------------------------

. lincom(_b[_cons] + _b[group[2]]+_b[x]*3.47619)
 ( 1)  _cons + group[2] + 3.47619 x = 0
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
         (1) |    5.50423   .1017763    54.08   0.000     5.300576    5.707883
------------------------------------------------------------------------------
. lincom(_b[_cons] + _b[x]*3.47619)
 ( 1)  _cons + 3.47619 x = 0
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   5.904621   .2277856    25.92   0.000     5.448823    6.360419
------------------------------------------------------------------------------
```

The problem is that when groups differ on covariate values, the differences in raw means may possibly be attributed just to that, not to real group differences. How can we devise a single summary of a group that is as simple as a mean yet does not suffer this limitation? We fix a single value of the covariate and take the point on the predicted least squares line at that value for each group. This is the estimated mean of the group for that value of the covariate. The default is to take the mean covariate value, but Stata will allow you to take any other value. In this case we are estimating the mean of each population at a value of $x = 3.47619$

Note that differences in adjusted means are just differences in intercepts, and those are the group differences we have been calculating. If we fit an interaction term it is no longer so simple, however. We still can calculate adjusted means, but differences are very specific to the x-value fit.

SAS does the same thing but uses the term LSMEANS (Least Squares Means) instead of adjusted means. The marginal means calculated at the end of the last chapter of notes are more easily obtained using SAS's approach, but can be calculated as adjusted means in Stata as well.

Returning to the preceding example, we see that the adjusted means are not nearly as different as are the raw means. The differences in covariate values explain the apparent differences in groups. The analysis in the separate output is the proper approach, but the table of adjusted means is easier for quick group comparisons on a familiar scale (assuming y has a familiar scale).

Now consider the first simulated data set:

```
. tabstat y,by(group) stat(mean semean)
Summary for variables: y
     by categories of: group
   group |      mean  se(mean)
---------+--------------------
       1 |  4.398373  .2309875
       2 |   4.62633  .2657725
       3 |  4.417013  .2081521
---------+--------------------
   Total |  4.480572  .1347717
----------------------------

. adjust x, by(group) se
-------------------------------------------------
     Dependent variable: y     Command: anova
  Covariate set to mean: x = 3.5
-------------------------------------------------
-----------------------------------
    group |         xb        stdp
----------+------------------------
        1 |    5.41749   (.135878)
        2 |    4.62633   (.114672)
        3 |     3.3979   (.135878)
-----------------------------------
     Key:  xb    =  Linear Prediction
           stdp  =  Standard Error
```

The very small differences in raw means become very large (and have much smaller standard errors when adjusted by x.

You should compute adjusted means in the CHDS data set as an exercise.

9 Review of Discrete Data Analysis

The material in this section was covered last semester. Since Stata differs from Minitab in how the methods are implemented, we will review those methods and see how to use Stata for them. The huge difference from what we have been doing is that the response or outcome variable is now categorical instead of continuous. Our goal is to extend all the t-test, regression, ANOVA, and ANCOVA methods we have studied to the case of categorical outcomes.

Comparing Two Proportions: Independent Samples

The New Mexico state legislature is interested in how the proportion of registered voters that support Indian gaming differs between New Mexico and Colorado. Assuming neither population proportion is known, the state's statistician might recommend that the state conduct a survey of registered voters sampled independently from the two states, followed by a comparison of the sample proportions in favor of Indian gaming.

Statistical methods for comparing two proportions using independent samples can be formulated as follows. Let p_1 and p_2 be the proportion of populations 1 and 2, respectively, with the attribute of interest. Let $\hat{p}_1$ and $\hat{p}_2$ be the corresponding sample proportions, based on independent random or representative samples of size n_1 and n_2 from the two populations.

Large Sample CI and Tests for $p_1 - p_2$

A large sample CI for $p_1 - p_2$ is $(\hat{p}_1 - \hat{p}_2) \pm z_{crit} SE_{CI}(\hat{p}_1 - \hat{p}_2)$, where z_{crit} is the standard normal critical value for the desired confidence level, and

$$SE_{CI}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

is the CI standard error.

A large sample p-value for a test of the null hypothesis $H_0 : p_1 - p_2 = 0$ against the two-sided alternative $H_A : p_1 - p_2 \neq 0$ is evaluated using tail areas of the standard normal distribution (identical to 1 sample evaluation) in conjunction with the test statistic

$$z_s = \frac{\hat{p}_1 - \hat{p}_2}{SE_{test}(\hat{p}_1 - \hat{p}_2)},$$

where

$$SE_{test}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}} = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

is the test standard error for $\hat{p}_1 - \hat{p}_2$. The **pooled proportion**

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

is the proportion of successes in the two samples combined. The test standard error has the same functional form as the CI standard error, with $\bar{p}$ replacing the individual sample proportions.

The pooled proportion is the best guess at the common population proportion when $H_0 : p_1 = p_2$ is true. The test standard error estimates the standard deviation of $\hat{p}_1 - \hat{p}_2$ assuming H_0 is true.

Example Two hundred and seventy nine French skiers were studied during two one-week periods in 1961. One group of 140 skiers receiving a placebo each day, and the other 139 receiving 1 gram of ascorbic acid (Vitamin C) per day. The study was double blind - neither the subjects

nor the researchers knew who received what treatment. Let p_1 be the probability that a member of the ascorbic acid group contracts a cold during the study period, and p_2 be the corresponding probability for the placebo group. Linus Pauling and I are interested in testing whether $p_1 = p_2$. The data are summarized below as a two-by-two table of counts (a contingency table)

Outcome	Ascorbic Acid	Placebo
# with cold	17	31
# with no cold	122	109
Totals	139	140

The sample sizes are $n_1 = 139$ and $n_2 = 140$. The sample proportion of skiers developing colds in the placebo and treatment groups are $\hat{p}_2 = 31/140 = .221$ and $\hat{p}_1 = 17/139 = .122$, respectively. The pooled proportion is the number of skiers that developed colds divided by the number of skiers in the study: $\bar{p} = 48/279 = .172$.

The test standard error is:

$$SE_{test}(\hat{p}_1 - \hat{p}_2) = \sqrt{.172 * (1 - .172)\left(\frac{1}{139} + \frac{1}{140}\right)} = .0452.$$

The test statistic is

$$z_s = \frac{.122 - .221}{.0452} = -2.19.$$

The p-value for a two-sided test is twice the area under the standard normal curve to the right of 2.19 (or twice the area to the left of -2.19), which is $2 * (.014) = .028$ At the 5% level, we reject the hypothesis that the probability of contracting a cold is the same whether you are given a placebo or Vitamin C.

A CI for $p_1 - p_2$ provides a measure of the size of the treatment effect. For a 95% CI

$$z_{crit} SE_{CI}(\hat{p}_1 - \hat{p}_2) = 1.96\sqrt{\frac{.221 * (1 - .221)}{140} + \frac{.122 * (1 - .122)}{139}} = 1.96 * (.04472) = .088.$$

The 95% CI for $p_1 - p_2$ is $(.122 - .221) \pm .088$, or $(-.187, -.011)$. We are 95% confident that p_2 exceeds p_1 by at least .011 but not by more than .187.

On the surface, we would conclude that a daily dose of Vitamin C decreases a French skier's chance of developing a cold by between .011 and .187 (with 95% confidence). This conclusion was somewhat controversial. Several reviews of the study felt that the experimenter's evaluations of cold symptoms were unreliable. Many other studies refute the benefit of Vitamin C as a treatment for the common cold.

To implement this test and obtain a CI using **Stata**'s `prtesti` command (immediate from of prtest command – uses data on the command line rather than in memory), we must provide the raw number of skiers receiving ascorbic acid (139) along with the proportion of these skiers that got a cold ($\hat{p}_1 = 0.122$), as well as the raw number of skiers receiving placebo (140) along with the proportion of these skiers that got a cold ($\hat{p}_2 = 0.221$). I actually like using the GUI `(Statistics -> Summaries, tables & tests -> Classical tests of hypotheses -> Two sample proportion calculator)` instead of the command line for this, in which case it all looks just like Minitab. Options and entries are a little more obvious from the GUI.

```
. prtesti 139 0.122 140 0.221
Two-sample test of proportion                     x: Number of obs =      139
                                                  y: Number of obs =      140
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |       .122     .02776                      .0675914    .1764086
           y |       .221   .0350672                      .1522696    .2897304
-------------+----------------------------------------------------------------
        diff |      -.099    .044725                     -.1866594   -.0113406
             |  under Ho:    .045153    -2.19   0.028
------------------------------------------------------------------------------

              Ho: proportion(x) - proportion(y) = diff = 0
      Ha: diff < 0            Ha: diff != 0             Ha: diff > 0
       z = -2.193              z = -2.193               z = -2.193
   P < z =  0.0142        P > |z| =  0.0283          P > z =  0.9858
```

It actually is a little more direct to use counts instead of proportions you calculate, by typing `prtesti 139 17 140 31, count`.

Example A case-control study was designed to examine risk factors for cervical dysplasia (Becker et al. 194). All the women in the study were patients at UNM clinics. The 175 cases were women, aged 18-40, who had cervical dysplasia. The 308 controls were women aged 18-40 who did not have cervical dysplasia. Each women was classified as positive or negative, depending on the presence of HPV (human papilloma virus).

The data are summarized below.

HPV Outcome	Cases	Controls
Positive	164	130
Negative	11	178
Sample size	175	308

Let p_1 be the probability that a case is HPV positive and let p_2 be the probability that a control is HPV positive. The sample sizes are $n_1 = 175$ and $n_2 = 308$. The sample proportions of positive cases and controls are $\hat{p}_1 = 164/175 = .937$ and $\hat{p}_2 = 130/308 = .422$.

For a 95% CI

$$z_{crit} SE_{CI}(\hat{p}_1 - \hat{p}_2) = 1.96\sqrt{\frac{.937 * (1 - .937)}{175} + \frac{.422 * (1 - .422)}{308}} = 1.96 * (.03336) = .0659.$$

A 95% CI for $p_1 - p_2$ is $(.937 - .422) \pm .066$, or $.515 \pm .066$, or $(.449, .581)$. I am 95% confident that p_1 exceeds p_2 by at least .45 but not by more than .58.

Not surprisingly, a two-sided test at the 5% level would reject $H_0 : p_1 = p_2$. In this problem one might wish to do a one-sided test, instead of a two-sided test. Can you find the p-value for the one-sided test in the **Stata** output below?

```
. prtesti 175 0.937 308 0.422
Two-sample test of proportion                     x: Number of obs =      175
                                                  y: Number of obs =      308
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |       .937   .0183663                      .9010028    .9729972
           y |       .422   .0281413                       .366844     .477156
-------------+----------------------------------------------------------------
        diff |       .515   .0336044                      .4491366    .5808634
             |  under Ho:   .0462016    11.15   0.000
------------------------------------------------------------------------------

              Ho: proportion(x) - proportion(y) = diff = 0
      Ha: diff < 0            Ha: diff != 0             Ha: diff > 0
       z = 11.147              z = 11.147               z = 11.147
   P < z =  1.0000        P > |z| =  0.0000          P > z =  0.0000
```

Appropriateness of the Large Sample Test and CI

The standard two sample CI and test used above are appropriate when each sample is large. A rule of thumb suggests a minimum of at least five successes (i.e. observations with the characteristic of interest) and failures (i.e. observations without the characteristic of interest) in each sample before using these methods. This condition is satisfied in our two examples.

Effect Measures in Two-by-Two Tables

Consider a study of a particular disease, where each individual is either exposed or not-exposed to a risk factor. Let p_1 be the proportion diseased among the individuals in the exposed population, and p_2 be the proportion diseased among the non-exposed population. This population information can be summarized as a two-by-two table of population proportions:

Outcome	Exposed population	Non-Exposed population
Diseased	p_1	p_2
Non-Diseased	$1-p_1$	$1-p_2$

A standard measure of the difference between the exposed and non-exposed populations is the **absolute difference:** $p_1 - p_2$. We have discussed statistical methods for assessing this difference.

In many epidemiological and biostatistical settings, other measures of the difference between populations are considered. For example, the relative risk

$$\text{RR} = \frac{p_1}{p_2}$$

is commonly reported when the individual risks p_1 and p_2 are small. The odds ratio

$$\text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

is another standard measure. Here $p_1/(1-p_1)$ is the odds of being diseased in the exposed group, whereas $p_2/(1-p_2)$ is the odds of being diseased in the non-exposed group.

Note that each of these measures can be easily estimated from data, using the sample proportions as estimates of the unknown population proportions. For example, in the vitamin C study:

Outcome	Ascorbic Acid	Placebo
# with cold	17	31
# with no cold	122	109
Totals	139	140

the proportion with colds in the placebo group is $\hat{p}_2 = 31/140 = .221$. The proportion with colds in the vitamin C group is $\hat{p}_1 = 17/139 = .122$.

The estimated absolute difference in risk is $\hat{p}_1 - \hat{p}_2 = .122 - .221 = -.099$. The estimated risk ratio and odds ratio are

$$\widehat{RR} = \frac{.122}{.221} = .55$$

and

$$\widehat{OR} = \frac{.122/(1-.122)}{.221/(1-.221)} = .49,$$

respectively.

In the literature it probably is most common to see OR (actually $\widehat{OR}$ or adjusted $\widehat{OR}$) reported, usually from a logistic regression analysis — that will be covered in the next section). We will be interested in testing $H_0 : OR = 1$ (or $H_0 : RR = 1$). We will estimate OR with $\widehat{OR}$ and will need the sampling distribution of $\widehat{OR}$ in order to construct tests and confidence intervals.

Testing for Homogeneity of Proportions

Example The following two-way table of counts summarizes the location of death and age at death from a study of 2989 cancer deaths (Public Health Reports, 1983):

(Obs Counts)	Location of death			
Age	Home	Acute Care	Chronic care	Row Total
15-54	94	418	23	535
55-64	116	524	34	674
65-74	156	581	109	846
75+	138	558	238	934
Col Total	504	2081	404	2989

The researchers want to compare the age distributions across locations. A one-way ANOVA would be ideal if the actual ages were given. Because the ages are grouped, the data should be treated as categorical. Given the differences in numbers that died at the three types of facilities, a comparison of proportions or percentages in the age groups is appropriate. A comparison of counts is not.

The table below summarizes the proportion in the four age groups at each location. For example, in the acute care facility 418/2081 = .201 and 558/2081 = .268. The **pooled proportions** are the Row Totals divided by the total sample size of 2989. The pooled summary gives the proportions in the four age categories, ignoring location of death.

The age distributions for home and for the acute care facilities are similar, but are very different from the age distribution at chronic care facilities.

To formally compare the observed proportions, one might view the data as representative sample of ages at death from the three locations. Assuming independent samples from the three locations (populations), a chi-squared statistic is used to test whether the population proportions of ages at death are identical (homogeneous) across locations. The **chi-squared test for homogeneity** of population proportions can be defined in terms of proportions, but is traditionally defined in terms of counts.

(Proportions)	Location of death			
Age	Home	Acute Care	Chronic care	Pooled
15-54	.187	.201	.057	.179
55-64	.230	.252	.084	.226
65-74	.310	.279	.270	.283
75+	.273	.268	.589	.312
Total	1.000	1.000	1.000	1.000

In general, assume that the data are independent samples from c populations (strata, groups, sub-populations), and that each individual is placed into one of r levels of a categorical variable. The raw data will be summarized as a $r \times c$ **contingency table** of counts, where the columns correspond to the samples, and the rows are the levels of the categorical variable. In the age distribution problem, $r = 4$ and $c = 3$. (SW uses k to identify the number of columns.)

To implement the test:

1. Compute the (estimated) **expected** count for each cell in the table as follows:

$$\mathrm{E} = \frac{\text{Row Total} * \text{Column Total}}{\text{Total Sample Size}}.$$

2. Compute the Pearson test statistic

$$\chi^2_S = \sum_{\text{all cells}} \frac{(O-E)^2}{E},$$

where O is the **observed** count.

3. For a size α test, reject the hypothesis of homogeneity if $\chi^2_S \geq \chi^2_{crit}$, where χ^2_{crit} is the upper α critical value from the chi-squared distribution with $df = (r-1)(c-1)$.

The p-value for the chi-squared test of homogeneity is equal to the area under the chi-squared curve to the right of χ^2_S; see Figure 1.

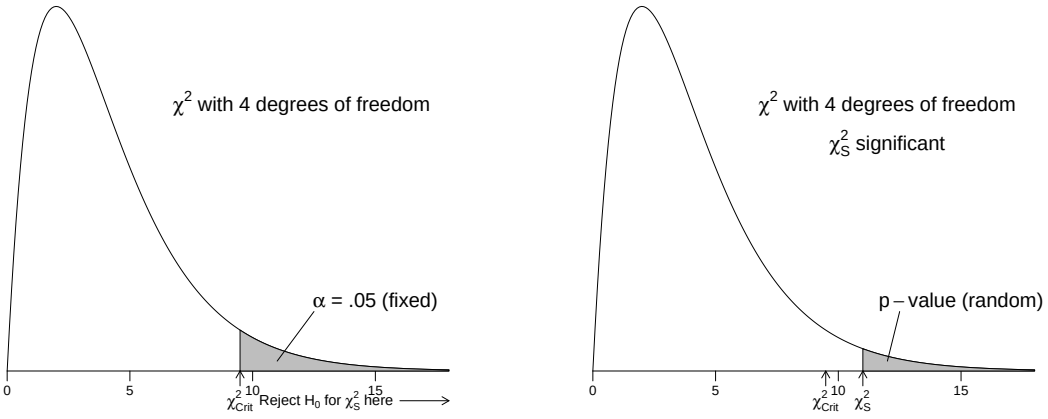


Figure 1: The p-value is the shaded area on the right

For a two-by-two table of counts, the chi-squared test of homogeneity of proportions is identical to the two-sample proportion test we discussed earlier.

Stata Analysis

One way to obtain the test statistic and p-value in **Stata** is to use the `tabi` command. The tables put out from that command are too poorly labelled to be very useful, though, so it's preferable to put the data into the worksheet so that it looks like this:

```
      Age   Location   Count
 1.     1          1      94
 2.     1          2     418
 3.     1          3      23
 4.     2          1     116
 5.     2          2     524
 6.     2          3      34
 7.     3          1     156
 8.     3          2     581
 9.     3          3     109
10.     4          1     138
11.     4          2     558
12.     4          3     238
```

The Hills and De Stavola book explains the following sequence,

```
. label define agemap 1  "15-54"  2  "55-64"  3   "65-74"   4  "75+"
. label define locmap 1 "Home" 2 "Acute Care" 3 "Chronic Care"
. label values Age agemap
. label values Location locmap
. list,clean
        Age       Location   Count
  1.   15-54           Home      94
  2.   15-54     Acute Care     418
  3.   15-54   Chronic Care      23
  4.   55-64           Home     116
  5.   55-64     Acute Care     524
  6.   55-64   Chronic Care      34
  7.   65-74           Home     156
  8.   65-74     Acute Care     581
  9.   65-74   Chronic Care     109
 10.     75+           Home     138
 11.     75+     Acute Care     558
 12.     75+   Chronic Care     238
```

If I typed `list,clean nolabel` I would get the original listing.

Why am I bothering with this? I actually could put those labels in as variable values, and not bother with labels. When I form tables, though, Stata wants to alphabetize according to variable values which will force Home as the last column. By keeping values numeric I can get Stata to order correctly and print the correct labels.

I find it easiest to go through the menu path `Summaries, tables, & tests -> Tables -> Two-way tables with measures of association` to generate the following commands. Note in particular the `[fweight = Count]` (frequency weight given by Count variable) syntax to tell Stata that each line represents many observations. Minitab and SAS have similar options.

```
. tabulate Age Location [fweight = Count], chi2 column expected lrchi2 row
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|   row percentage   |
| column percentage  |
+--------------------+

           |            Location
       Age |      Home  Acute Car  Chronic C |     Total
-----------+---------------------------------+----------
     15-54 |        94        418         23 |       535
           |      90.2      372.5       72.3 |     535.0
           |     17.57      78.13       4.30 |    100.00
           |     18.65      20.09       5.69 |     17.90
-----------+---------------------------------+----------
     55-64 |       116        524         34 |       674
           |     113.6      469.3       91.1 |     674.0
           |     17.21      77.74       5.04 |    100.00
           |     23.02      25.18       8.42 |     22.55
-----------+---------------------------------+----------
     65-74 |       156        581        109 |       846
           |     142.7      589.0      114.3 |     846.0
           |     18.44      68.68      12.88 |    100.00
           |     30.95      27.92      26.98 |     28.30
-----------+---------------------------------+----------
       75+ |       138        558        238 |       934
           |     157.5      650.3      126.2 |     934.0
           |     14.78      59.74      25.48 |    100.00
           |     27.38      26.81      58.91 |     31.25
-----------+---------------------------------+----------
     Total |       504      2,081        404 |     2,989
           |     504.0    2,081.0      404.0 |   2,989.0
           |     16.86      69.62      13.52 |    100.00
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(6) = 197.6241   Pr = 0.000
 likelihood-ratio chi2(6) = 200.9722   Pr = 0.000
```

The Pearson statistic is 197.6241 on 6 = (4-1)(3-1) *df*. The p-value is 0 to three places. The data strongly suggest that there are differences in the age distributions among locations.

Testing for Homogeneity in Cross-Sectional and Stratified Studies

Two-way tables of counts are often collected either by **stratified sampling** or by **cross-sectional** sampling.

In a stratified design, distinct groups, strata, or sub-populations are identified. Independent samples are selected from each group, and the sampled individuals are classified into categories. The HPV study is an illustration of a stratified design (and a case-control study). Stratified designs provide estimates for the strata (population) proportion in each of the categories. A test for **homogeneity of proportions** is used to compare the strata.

In a **cross-sectional design**, individuals are randomly selected from a population and classified by the levels of **two** categorical variables. With cross-sectional samples you can test homogeneity of proportions by comparing either the row proportions or by comparing the column proportions.

Example The following data (*The Journal of Advertising,* 1983, p. 34-42) are from a cross-sectional study that involved soliciting opinions on anti-smoking advertisements. Each subject was asked whether they smoked and their reaction (on a five-point ordinal scale) to the ad. The data are summarized as a two-way table of counts, given below:

	Str. Dislike	Dislike	Neutral	Like	Str. Like	Row Tot
Smoker	8	14	35	21	19	97
Non-smoker	31	42	78	61	69	281
Col Total	39	56	113	82	88	378

The row proportions are

(Row Prop)	Str. Dislike	Dislike	Neutral	Like	Str. Like	Row Tot
Smoker	.082	.144	.361	.216	.196	1.000
Non-smoker	.110	.149	.278	.217	.245	1.000

For example, the entry for the (Smoker, Str. Dislike) cell is: $8/97 = .082$.

Similarly, the column proportions are

(Col Prop)	Str. Dislike	Dislike	Neutral	Like	Str. Like
Smoker	.205	.250	.310	.256	.216
Non-smoker	.795	.750	.690	.744	.784
Total	1.000	1.000	1.000	1.000	1.000

Although it may be more natural to compare the smoker and non-smoker row proportions, the column proportions can be compared across ad responses. There is no advantage to comparing "rows" instead of "columns" in a formal test of homogeneity of proportions with cross-sectional data. The Pearson chi-squared test treats the rows and columns interchangeably, so you get the same result regardless of how you view the comparison. However, one of the two comparisons may be more natural to interpret.

Note that checking for homogeneity of proportions is meaningful in stratified studies only when the comparison is across strata! Further, if the strata correspond to columns of the table, then the column proportions or percentages are meaningful whereas the row proportions are not.

Question: How do these ideas apply to the age distribution problem?

Testing for Independence in a Two-Way Contingency Table

The row and column classifications for a population where each individual is cross-classified by two categorical variables are said to be independent if each **population** cell proportion in the two-way table is the product of the proportion in a given row and the proportion in a given column. One can show that independence is equivalent to homogeneity of proportions. In particular, the two-way table of population cell proportions satisfies independence if and only if the population column proportions are homogeneous. If the population column proportions are homogeneous then so are the population row proportions.

This suggests that a test for independence or **no association** between two variables based on a cross-sectional study can be implemented using the chi-squared test for homogeneity of proportions. This suggestion is correct. If independence is not plausible, I tend to interpret the dependence as a deviation from homogeneity, using the classification for which the interpretation is most natural.

Example: **Stata** output for testing independence between smoking status and ad reaction is given below. The Pearson chi-squared test is not significant (p-value = 0.559). The observed association between smoking status and the ad reaction is not significant. This suggests, for example, that the smoker's reactions to the ad were not statistically significantly different from the non-smoker's reactions, which is consistent with the smokers and non-smokers attitudes being fairly similar. The data were coded as opinion from 1 to 5 and smoke as 1 or 2, and then label define applied as before.

```
. tabulate Smoke Opinion [fweight=count],chi2 lrchi2 exp col row
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|   row percentage   |
| column percentage  |
+--------------------+
           |                          Opinion
     Smoke | Str. Disl    Dislike    Neutral       Like  Str. Like |     Total
-----------+-------------------------------------------------------+----------
    Smoker |         8         14         35         21         19 |        97
           |      10.0       14.4       29.0       21.0       22.6 |      97.0
           |      8.25      14.43      36.08      21.65      19.59 |    100.00
           |     20.51      25.00      30.97      25.61      21.59 |     25.66
-----------+-------------------------------------------------------+----------
Non-smoker |        31         42         78         61         69 |       281
           |      29.0       41.6       84.0       61.0       65.4 |     281.0
           |     11.03      14.95      27.76      21.71      24.56 |    100.00
           |     79.49      75.00      69.03      74.39      78.41 |     74.34
-----------+-------------------------------------------------------+----------
     Total |        39         56        113         82         88 |       378
           |      39.0       56.0      113.0       82.0       88.0 |     378.0
           |     10.32      14.81      29.89      21.69      23.28 |    100.00
           |    100.00     100.00     100.00     100.00     100.00 |    100.00
          Pearson chi2(4) =   2.9907   Pr = 0.559
 likelihood-ratio chi2(4) =   2.9797   Pr = 0.561
```

One-sample procedures

Last semester we spent some time on the situation where we obtained a SRS of n observations from a binomial population (binary outcome variable) with probability p of Success. We learned how to calculate CIs for p and tests of $H_0 : p = p_0$ for some fixed p_0. The large sample form of this is also done with the `prtesti` command or through the GUI, and the (preferable) exact binomial test is done through the `bitesti` command (or through the menus). The extension to 3 or more categories was the chi-squared goodness of fit test, done in Stata using the `csgof` command. That command is not automatically installed but you can locate and install it from the `findit csgof`

command. Since we do these one sample procedures relatively infrequently, I am going to leave it to you to learn them in Stata if you need them.

10 Logistic Regression - Two Introductory Examples

The chi-squared tests in the previous section are used very frequently, along with Fisher's exact test (asked for with the `,fisher` option in tabulate – note that it is often feasible to calculate only for small sample sizes). Those "classical" methods have been around a very long time and are often the best choice for analysis. In order to consider problems with more complicated predictors we need newer technology, so we now turn to logistic regression.

The data below are from a study conducted by Milicer and Szczotka on pre-teen and teenage girls in Warsaw. The subjects were classified into 25 age categories. The number of girls in each group (sample size) and the number that reached menarche (# RM) at the time of the study were recorded. The age for a group corresponds to the midpoint for the age interval.

Sample size	# RM	Age	Sample size	# RM	Age
376	0	9.21	106	67	13.33
200	0	10.21	105	81	13.58
93	0	10.58	117	88	13.83
120	2	10.83	98	79	14.08
90	2	11.08	97	90	14.33
88	5	11.33	120	113	14.58
105	10	11.58	102	95	14.83
111	17	11.83	122	117	15.08
100	16	12.08	111	107	15.33
93	29	12.33	94	92	15.58
100	39	12.58	114	112	15.83
108	51	12.83	1049	1049	17.58
99	47	13.08			

The researchers were interested in whether the proportion of girls that reached menarche (# RM/ sample size) varied with age. One could perform a test of homogeneity by arranging the data as a 2 by 25 contingency table with columns indexed by age and two rows: ROW1 = # RM and ROW2 = # that have not RM = sample size − # RM. A more powerful approach treats these as regression data, using the proportion of girls reaching menarche as the "response" and age as a predictor.

The data were imported into **Stata** using the `infile` command and labelled `menarche`, `total`, and `age`. A plot of the observed proportion of girls that have reached menarche (obtained in **Stata** with the two commands `generate phat = menarche / total` and `twoway (scatter phat age)`) shows that the proportion increases as age increases, but that the relationship is nonlinear.

The observed proportions, which are bounded between zero and one, have a lazy S-shape (a **sigmoidal function**) when plotted against age. The change in the observed proportions for a given change in age is much smaller when the proportion is near 0 or 1 than when the proportion is near 1/2. This phenomenon is common with regression data where the response is a proportion.

The trend is nonlinear so linear regression is inappropriate. A sensible alternative might be to transform the response or the predictor to achieve near linearity. A better approach is to use a non-linear model for the proportions. A common choice is the **logistic regression model**.

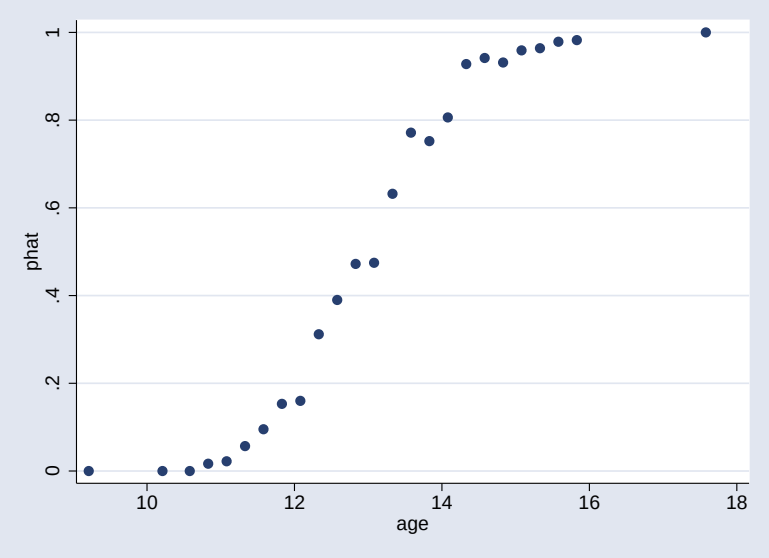


Figure 2: Estimated proportions $\hat{p}_i$ versus AGE_i, for $i = 1, \ldots, 25$.

The Simple Logistic Regression Model

The simple logistic regression model expresses the population proportion p of individuals with a given attribute (called a success) as a function of a single predictor variable X. The model assumes that p is related to X through

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta X \tag{1}$$

or, equivalently, as

$$p = \frac{exp(\alpha + \beta X)}{1 + exp(\alpha + \beta X)}.$$

The logistic regression model is a **binary response model**, where the response for each case falls into one of 2 exclusive and exhaustive categories, often called success (cases with the attribute of interest) and failure (cases without the attribute of interest). In many biostatistical applications, the success category is presence of a disease, or death from a disease.

I will often write p as $p(X)$ to emphasize that p is the proportion of all individuals with score X that have the attribute of interest. In the menarche data, $p = p(X)$ is the population proportion of girls at age X that have reached menarche.

The odds of success are $p/(1-p)$. For example, the odds of success are 1 (or 1 to 1) when $p = 1/2$. The odds of success are 2 (or 2 to 1) when $p = 2/3$. The logistic model assumes that the log-odds of success is linearly related to X. Graphs of the logistic model relating p to X are given in Figure 3. The sign of the slope refers to the sign of β.

There are a variety of other binary response models that are used in practice. The **probit** regression model or the **complementary log-log** regression model might be appropriate when the logistic model does not fit the data.

Data for Simple Logistic Regression

For the formulas below, I assume that the data are given in summarized or **aggregate** form:

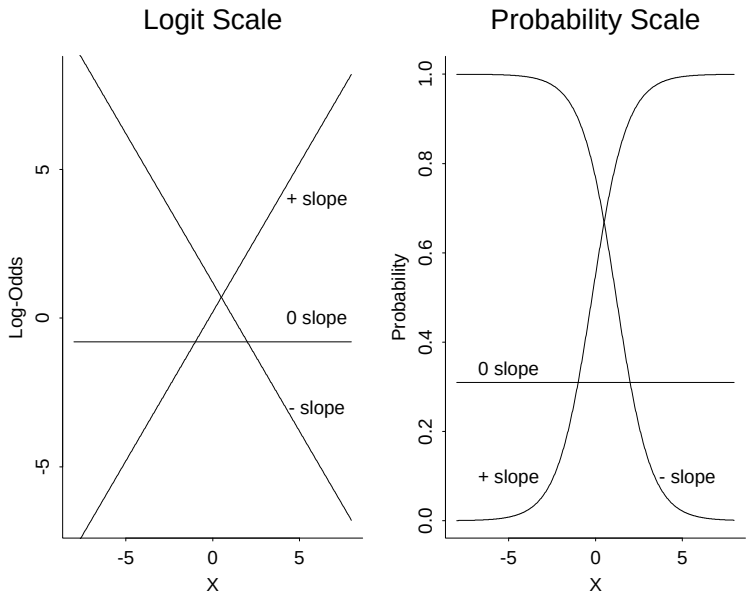


Figure 3: logit(p) and p as a function of X

X	n	D
X_1	n_1	d_1
X_2	n_2	d_2
.	.	.
.	.	.
X_m	n_m	d_m

where d_i is the number of individuals with the attribute of interest (number of diseased) among n_i randomly selected or representative individuals with predictor variable value X_i. The subscripts identify the group of cases in the data set. In many situations, the sample size is 1 in each group, and for this situation d_i is 0 or 1.

For **raw data** on individual cases, the sample size column n is usually omitted and D takes on 1 of two coded levels, depending on whether the case at X_i is a success or not. The values 0 and 1 are typically used to identify "failures" and "successes" respectively.

Estimating Regression Coefficients

The principle of maximum likelihood is commonly used to estimate the two unknown parameters in the logistic model:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta X.$$

The **maximum likelihood estimates** (MLE) of the regression coefficients are estimated iteratively by maximizing the so-called Binomial likelihood function for the responses, or equivalently, by minimizing the **deviance** function (also called the likelihood ratio LR chi-squared statistic)

$$\text{LR} = 2\sum_{i=1}^{m}\left\{d_i\log\left(\frac{d_i}{n_i p_i}\right) + (n_i - d_i)\log\left(\frac{n_i - d_i}{n_i - n_i p_i}\right)\right\}$$

over all possible values of α and β, where the p_is satisfy

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta X_i.$$

The ML method also gives standard errors and significance tests for the regression estimates.

The deviance is an analog of the residual sums of squares in linear regression. The choices for α and β that minimize the deviance are the parameter values that make the observed and fitted proportions as close together as possible in a "likelihood sense".

Suppose that $\hat{\alpha}$ and $\hat{\beta}$ are the MLEs of α and β. The deviance evaluated at the MLEs:

$$\text{LR} = 2\sum_{i=1}^{m}\left\{d_i \log\left(\frac{d_i}{n_i\tilde{p}_i}\right) + (n_i - d_i)\log\left(\frac{n_i - d_i}{n_i - n_i\tilde{p}_i}\right)\right\},$$

where the fitted probabilities $\tilde{p}_i$ satisfy

$$\log\left(\frac{\tilde{p}_i}{1-\tilde{p}_i}\right) = \hat{\alpha} + \hat{\beta}X_i,$$

is used to test the adequacy of the model. The deviance is small when the data fits the model, that is, when the observed and fitted proportions are close together. Large values of LR occur when one or more of the observed and fitted proportions are far apart, which suggests that the model is inappropriate.

If the logistic model holds, then LR has a chi-squared distribution with $m-r$ degrees of freedom, where m is the number of groups and r (here 2) is the number of estimated regression parameters. A p-value for the deviance is given by the area under the chi-squared curve to the right of LR. A small p-value indicates that the data does not fit the model.

Stata does not provide the deviance statistic, but rather the Pearson chi-squared test statistic, which is defined similarly to the deviance statistic and is interpreted in the same manner:

$$X^2 = \sum_{i=1}^{m}\frac{(d_i - n_i\tilde{p}_i)^2}{n_i\tilde{p}_i(1-\tilde{p}_i)}.$$

This statistic can be interpreted as the sum of standardized, squared differences between the *observed* number of successes d_i and *expected* number of successes $n_i\tilde{p}_i$ for each covariate X_i. When what we expect to see under the model agrees with what we see, the Pearson statistic is close to zero, indicating good model fit to the data. When the Pearson statistic is *large*, we have an indication of lack of fit. Often the *Pearson residuals* $r_i = (d_i - n_i\tilde{p}_i)/\sqrt{n_i\tilde{p}_i(1-\tilde{p}_i)}$ are used to determine exactly *where* lack of fit occurs. These residuals are obtained in **Stata** using the `predict` command after the `logistic` command. Examining these residuals is very similar to looking for large values of $\frac{(O-E)^2}{E}$ in a χ^2 analysis of a contingency table as discussed in the last lecture. We will not talk further of logistic regression diagnostics.

Age at Menarche Data: Stata Implementation

A logistic model for these data implies that the probability p of reaching menarche is related to age through

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \text{ AGE}.$$

If the model holds, then a slope of $\beta = 0$ implies that p does not depend on AGE, i.e. the proportion of girls that have reached menarche is identical across age groups. However, the power of the logistic regression model is that if the model holds, and if the proportions change with age, then you have a way to quantify the effect of age on the proportion reaching menarche. This is more appealing and useful than just testing homogeneity across age groups.

A logistic regression model with a single predictor can be fit using one of the many commands available in **Stata** depending on the data type and desired results: `logistic` (raw data, outputs

odds ratios), `logit` (raw data, outputs model parameter estimates), and `blogit` (grouped data). The `logistic` command has many more options than either `logit` or `blogit`, but requires you to reformat the data into individual records, one for each girl. For an example of how to do this, check out the online **Stata** help at **http://www.stata.com/support/faqs/stat/grouped.html**. The **Stata** command `blogit menarche total age` yields the following output:

```
Logit estimates                                   Number of obs   =       3918
                                                  LR chi2(1)      =    3667.18
                                                  Prob > chi2     =     0.0000
Log likelihood = -819.65237                       Pseudo R2       =     0.6911
------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.631968   .0589509    27.68   0.000     1.516427     1.74751
       _cons |  -21.22639   .7706558   -27.54   0.000    -22.73685   -19.71594
------------------------------------------------------------------------------
```

The output tables the MLEs of the parameters: $\hat{\alpha} = -21.23$ and $\hat{\beta} = 1.63$. Thus, the fitted or predicted probabilities satisfy:

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = -21.23 + 1.63\text{AGE}$$

or

$$\tilde{p}(AGE) = \frac{\exp(-21.23 + 1.63\text{AGE})}{1 + \exp(-21.23 + 1.63\text{AGE})}.$$

The p-value for testing $H_0 : \beta = 0$ (i.e. the slope for the regression model is zero) based upon the chi-squared test p-value (`P>|z|`) is 0.000, which leads to rejecting H_0 at any of the usual test levels. Thus, the proportion of girls that have reached menarche is not constant across age groups.

The likelihood ratio test statistic of no logistic regression relationship (`LR chi2(1) = 3667.18`) and p-value (`Prob > chi2 = 0.0000`) gives the logistic regression analogue of the overall F-statistic that *no predictors are important* to multiple regression. In general, the chi-squared statistic provided here is used to test the hypothesis that the regression coefficients are zero for each predictor in the model. There is a single predictor here, AGE, so this test and the test for the AGE effect are *both* testing $H_0 : \beta = 0$.

To obtain the Pearson goodness of fit statistic and p-value we must reformat the data and use the `logistic` command as described in the webpage above:

```
generate w0 = total - menarche
rename menarche w1
generate id = _n
reshape long w, i(id) j(y)
logistic y age [fw=w]
lfit
```

We obtain the following output:

```
Logistic regression                               Number of obs   =       3918
                                                  LR chi2(1)      =    3667.18
                                                  Prob > chi2     =     0.0000
Log likelihood = -819.65237                       Pseudo R2       =     0.6911
------------------------------------------------------------------------------
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   5.113931   .3014706    27.68   0.000     4.555917    5.740291
------------------------------------------------------------------------------

Logistic model for y, goodness-of-fit test

       number of observations =      3918
 number of covariate patterns =        25
             Pearson chi2(23) =        21.87
                  Prob > chi2 =         0.5281
```

Using properties of exponential functions, the odds of reaching menarche is $\exp(1.632) = 5.11$ times larger for every year older a girl is. To see this, let $p(\text{Age}+1)$ and $p(\text{Age})$ be probabilities of reaching menarche for ages one year apart. The odds ratio OR satisfies

$$\begin{aligned}
\log(OR) &= \log\left(\frac{p(\text{Age}+1)/(1-\text{p}(\text{Age}+1))}{p(\text{Age})/(1-\text{p}(\text{Age}))}\right) \\
&= \log\left(p(\text{Age}+1)/(1-\text{p}(\text{Age}+1))\right) - \log\left(p(\text{Age})/(1-\text{p}(\text{Age}))\right) \\
&= (\alpha+\beta(\text{Age}+1)) - (\alpha+\beta\ \text{Age}) \\
&= \beta
\end{aligned}$$

so $OR = e^{\beta}$. If we considered ages 5 years apart, the same derivation would give us $OR = e^{5\beta} = (e^{\beta})^5$. You often see a continuous variable with a significant though apparently small OR, but when you examine the OR for a reasonable range of values (by raising to the power of the range in this way), then the OR is substantial.

You should pick out the estimated regression coefficient $\hat{\beta} = 1.632$ and the estimated odds ratio $\exp(\hat{\beta}) = \exp(1.632) = 5.11$ from the output obtained using the `blogit` and `logistic` commands respectively. We would say that, for example, that the odds of 15 year old girls having reached menarche are between 4.5 and 5.7 times larger than for 14 year old girls.

The Pearson chi-square statistic is 21.87 on 23 df, with a p-value of 0.5281. The large p-value suggests no gross deficiencies with the logistic model.

Logistic Regression with Two Effects: Leukemia Data

Feigl and Zelen reported the survival time in weeks and the white cell blood count (WBC) at time of diagnosis for 33 patients who eventually died of acute leukemia. Each person was classified as AG+ or AG- (coded as IAG = 1 and 0, respectively), indicating the presence or absence of a certain morphological characteristic in the white cells. The researchers are interested in modelling the probability p of surviving at least one year as a function of WBC and IAG. They believe that WBC should be transformed to a log scale, given the skewness in the WBC values. Where `Live`=0, 1 indicates whether the patient died or lived respectively, the data are

IAG	WBC	Live	IAG	WBC	Live	IAG	WBC	Live
1	75	1	1	230	1	1	430	1
1	260	1	1	600	0	1	1050	1
1	1000	1	1	1700	0	1	540	0
1	700	1	1	940	1	1	3200	0
1	3500	0	1	5200	0	1	10000	1
1	10000	0	1	10000	0	0	440	1
0	300	1	0	400	0	0	150	0
0	900	0	0	530	0	0	1000	0
0	1900	0	0	2700	0	0	2800	0
0	3100	0	0	2600	0	0	2100	0
0	7900	0	0	10000	0	0	10000	0

As an initial step in the analysis, consider the following model:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{LWBC} + \beta_2 \text{IAG},$$

where LWBC = log WBC. This is a logistic regression model with 2 effects, fit using the `logistic` command. The parameters α, β_1 and β_2 are estimated by maximum likelihood.

The model is best understood by separating the AG+ and AG- cases. For AG- individuals, IAG=0 so the model reduces to

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{LWBC} + \beta_2 * 0 = \alpha + \beta_1 \text{LWBC}.$$

For AG+ individuals, IAG=1 and the model implies

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{LWBC} + \beta_2 * 1 = (\alpha + \beta_2) + \beta_1 \text{LWBC}.$$

The model without IAG (i.e. $\beta_2 = 0$) is a simple logistic model where the log-odds of surviving one year is linearly related to LWBC, and is independent of AG. The reduced model with $\beta_2 = 0$ implies that there is no effect of the AG level on the survival probability *once LWBC has been taken into account.*

Including the **binary predictor** IAG in the model implies that there is a linear relationship between the log-odds of surviving one year and LWBC, with a constant slope for the two AG levels. This model includes an effect for the AG morphological factor, but more general models are possible. Thinking of IAG as a **factor**, the proposed model is a logistic regression analog of ANCOVA.

The parameters are easily interpreted: α and $\alpha + \beta_2$ are intercepts for the population logistic regression lines for AG- and AG+, respectively. The lines have a common slope, β_1. The β_2 coefficient for the IAG indicator is the difference between intercepts for the AG+ and AG- regression lines. A picture of the assumed relationship is given below for $\beta_1 < 0$. The population regression lines are parallel on the logit (i.e. log odds) scale only, but the order between IAG groups is preserved on the probability scale.

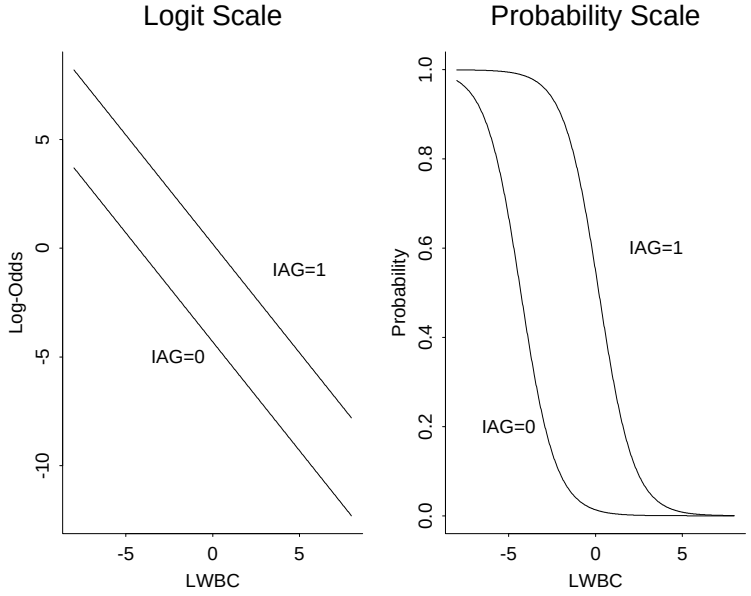


Figure 4: Predicted relationships on the logit and probability scales

The data are in the **raw data** form for individual cases. There are three columns: the binary or **indicator variable** `iag` (with value 1 for AG+, 0 for AG-), `wbc` (continuous), `live` (with value 1 if the patient lived at least 1 year and 0 if not). Note that a frequency column is not needed with

raw data (and hence using the `logistic` command) and that the success category corresponds to surviving at least 1 year.

Before looking at output for the equal slopes model, note that the data set has 30 distinct IAG and WBC combinations, or 30 "groups" or samples that could be constructed from the 33 individual cases. Only two samples have more than 1 observation. The majority of the observed proportions surviving at least one year (number surviving $\geq$ 1 year/ group sample size) are 0 (i.e. 0/1) or 1 (i.e. 1/1). This sparseness of the data makes it difficult to graphically assess the suitability of the logistic model (Why?). Although significance tests on the regression coefficients do not require large group sizes, the chi-squared approximations to the deviance and Pearson goodness-of-fit statistics are suspect in sparse data settings. With small group sizes as we have here, most researchers would not interpret the p-values for the deviance or Pearson tests literally. Instead, they would use the p-values to informally check the fit of the model. Diagnostics would be used to highlight problems with the model.

We obtain the following modified output:

```
. infile iag wbc live using c:/biostat/notes/leuk.txt
. generate lwbc = log(wbc)
. logistic live iag lwbc
. logit
. lfit
Logistic regression                                     Number of obs   =         33
                                                        LR chi2(2)      =      15.18
                                                        Prob > chi2     =     0.0005
Log likelihood = -13.416354                             Pseudo R2       =     0.3613
------------------------------------------------------------------------------
       live | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        iag |   12.42316    13.5497     2.31   0.021     1.465017    105.3468
       lwbc |   .3299682   .1520981    -2.41   0.016     .1336942    .8143885
------------------------------------------------------------------------------
Logit estimates                                         Number of obs   =         33
                                                        LR chi2(2)      =      15.18
                                                        Prob > chi2     =     0.0005
Log likelihood = -13.416354                             Pseudo R2       =     0.3613
------------------------------------------------------------------------------
       live |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        iag |   2.519562   1.090681     2.31   0.021     .3818672    4.657257
       lwbc |  -1.108759   .4609479    -2.41   0.016      -2.0122   -.2053178
      _cons |   5.543349   3.022416     1.83   0.067     -.380477    11.46718
------------------------------------------------------------------------------
Logistic model for live, goodness-of-fit test
       number of observations =        33
 number of covariate patterns =        30
            Pearson chi2(27) =        19.81
                 Prob > chi2 =         0.8387
```

The large p-value (0.8387) for the lack-of-fit chi-square (i.e. the Pearson statistic) indicates that there are no gross deficiencies with the model. Given that the model fits reasonably well, a test of $H_0 : \beta_2 = 0$ might be a primary interest here. This checks whether the regression lines are identical for the two AG levels, which is a test for whether AG affects the survival probability, after taking LWBC into account. The test that $H_0 : \beta_2 = 0$ is equivalent to testing that the odds ratio $\exp(\beta_2)$ is equal to 1: $H_0 : e^{\beta_2} = 1$. The p-value for this test is 0.021. The test is rejected at any of the usual significance levels, suggesting that the AG level affects the survival probability (assuming a very specific model). In fact we estimate that the odds of surviving past a year in the AG+ population is 12.4 times the odds of surviving past a year in the AG- population, with a 95% CI of (1.4, 105.4); see below for this computation carried out explicitly.

The estimated survival probabilities satisfy

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = 5.54 - 1.11\text{LWBC} + 2.52\text{IAG}.$$

For AG- individuals with IAG=0, this reduces to

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = 5.54 - 1.11\text{LWBC},$$

or equivalently,

$$\tilde{p} = \frac{\exp(5.54 - 1.11\text{LWBC})}{1 + \exp(5.54 - 1.11\text{LWBC})}.$$

For AG+ individuals with IAG=1,

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = 5.54 - 1.11\text{LWBC} + 2.52 * (1) = 8.06 - 1.11\text{LWBC},$$

or

$$\tilde{p} = \frac{\exp(8.06 - 1.11\text{LWBC})}{1 + \exp(8.06 - 1.11\text{LWBC})}.$$

Using the **logit scale**, the difference between AG+ and AG- individuals in the estimated log-odds of surviving at least one year, at a fixed but arbitrary LWBC, is the estimated IAG regression coefficient:

$$(8.06 - 1.11\text{LWBC}) - (5.54 - 1.11\text{LWBC}) = 2.52.$$

Using properties of exponential functions, the odds that an AG+ patient lives at least one year is $\exp(2.52) = 12.42$ times larger than the odds that an AG- patient lives at least one year, regardless of LWBC.

Although the equal slopes model appears to fit well, a more general model might fit better. A natural generalization here would be to add an **interaction**, or product term, IAG $*$ LWBC to the model. The logistic model with an IAG effect and the IAG $*$ LWBC interaction is equivalent to fitting separate logistic regression lines to the two AG groups. This interaction model provides an easy way to test whether the slopes are equal across AG levels. I will note that the interaction term is not needed here.

11 Logistic Regression - Interpreting Parameters

Let us expand on the material in the last section, trying to make sure we understand the logistic regression model and can interpret **Stata** output. Consider first the case of a single binary predictor, where

$$x = \begin{cases} 1 & \text{if exposed to factor} \\ 0 & \text{if not} \end{cases} \text{, and } y = \begin{cases} 1 & \text{if develops disease} \\ 0 & \text{does not} \end{cases} .$$

Results can be summarized in a simple 2 X 2 contingency table as

	Exposure	
Disease	1	0
1 (+)	a	b
0 (–)	c	d

where $\widehat{OR} = \frac{ad}{bc}$ (why?) and we interpret $\widehat{OR} > 1$ as indicating a risk factor, and $\widehat{OR} < 1$ as indicating a protective factor.

Recall the logistic model: $p(x)$ is the probability of disease for a given value of x, and

$$\text{logit}(p(x)) = \log\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \beta x.$$

Then for x = 0 (unexposed), $\text{logit}(p(x)) = \text{logit}(p(0)) = \alpha + \beta(0) = \alpha$
x = 1 (exposed), $\text{logit}(p(x)) = \text{logit}(p(1)) = \alpha + \beta(1) = \alpha + \beta$

Also,
odds of disease among unexposed: $p(0)/(1-p(0))$
exposed: $p(1)/(1-p(1))$

Now

$$OR = \frac{\text{odds of disease among exposed}}{\text{odds of disease among unexposed}} = \frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}$$

and

$$\begin{aligned} \beta &= \text{logit}(p(1)) - \text{logit}(p(0)) \\ &= \log\left(\tfrac{p(1)}{(1-p(1))}\right) - \log\left(\tfrac{p(0)}{(1-p(0))}\right) \\ &= \log\left(\tfrac{p(1)/(1-p(1))}{p(0)/(1-p(0))}\right) \\ &= \log(OR) \end{aligned}$$

The regression coefficient in the population model is the $\log(OR)$, hence the OR is obtained by exponentiating β,

$$e^{\beta} = e^{\log(OR)} = OR$$

Remark: If we fit this simple logistic model to a 2 X 2 table, the estimated unadjusted OR (above) and the regression coefficient for x have the same relationship.

Example: Leukemia Survival Data (Section 10 p. 108). We can find the counts in the following table from the `tabulate live iag` command:

Surv ≥ 1 yr?	Ag+ (x=1)	Ag- (x=0)
Yes	9	2
No	8	14

and (unadjusted) $\widehat{OR} = \frac{9(14)}{2(8)} = 7.875$.

Before proceeding with the **Stata** output, let me comment about coding of the outcome variable. Some packages are less rigid, but **Stata** enforces the (reasonable) convention that 0 indicates a negative outcome and all other values indicate a positive outcome. If you try to code something like 2 for survive a year or more and 1 for not survive a year or more, **Stata** coaches you with the error message

```
outcome does not vary; remember:
                                0 = negative outcome,
        all other nonmissing values = positive outcome
```

This data set uses 0 and 1 codes for the `live` variable; 0 and -100 would work, but not 1 and 2. Let's look at both regression estimates and direct estimates of unadjusted odds ratios from **Stata**.

```
. logit live iag
Logit estimates                                   Number of obs   =         33
                                                  LR chi2(1)      =       6.45
                                                  Prob > chi2     =     0.0111
Log likelihood = -17.782396                       Pseudo R2       =     0.1534
------------------------------------------------------------------------------
        live |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iag |   2.063693   .8986321     2.30   0.022     .3024066     3.82498
       _cons |   -1.94591   .7559289    -2.57   0.010    -3.427504   -.4643167
------------------------------------------------------------------------------
. logistic live iag
Logistic regression                               Number of obs   =         33
                                                  LR chi2(1)      =       6.45
                                                  Prob > chi2     =     0.0111
Log likelihood = -17.782396                       Pseudo R2       =     0.1534
------------------------------------------------------------------------------
        live | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iag |      7.875   7.076728     2.30   0.022     1.353111    45.83187
------------------------------------------------------------------------------
```

Stata has fit $\text{logit}(\hat{p}(x)) = \log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) = \hat{\alpha} + \hat{\beta}x = -1.946 + 2.064\ IAG$, with $\widehat{OR} = e^{2.064} = 7.875$. This is identical to the "hand calculation" above. A 95% Confidence Interval for β (IAG coefficient) is $.3024066 \leq \beta \leq 3.82498$. This logit scale is where the real work and theory is done. To get a Confidence Interval for the odds ratio, just exponentiate everything

$$\begin{array}{rcccl} e^{.3024066} & \leq & e^{\beta} & \leq & e^{3.82498} \\ 1.353111 & \leq & OR & \leq & 45.83187 \end{array}$$

What do you conclude?

A More Complex Model

$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2$, where x_1 is binary (as before) and x_2 is a continuous predictor. The regression coefficients are *adjusted log-odds ratios*.

To interpret β_1, fix the value of x_2:
For $x_1 = 0$

$$\begin{array}{rcl} \text{log odds of disease} & = & \alpha + \beta_1(0) + \beta_2 x_2 = \alpha + \beta_2 x_2 \\ \text{odds of disease} & = & e^{\alpha+\beta_2 x_2} \end{array}$$

For $x_1 = 1$

$$\begin{array}{rcl} \text{log odds of disease} & = & \alpha + \beta_1(1) + \beta_2 x_2 = \alpha + \beta_1 + \beta_2 x_2 \\ \text{odds of disease} & = & e^{\alpha+\beta_1+\beta_2 x_2} \end{array}$$

Thus the odds ratio (going from $x_1 = 0$ to $x_1 = 1$ is

$$OR = \frac{\text{odds when } x_1 = 1}{\text{odds when } x_1 = 0} = \frac{e^{\alpha+\beta_1+\beta_2 x_2}}{e^{\alpha+\beta_2 x_2}} = e^{\beta_1}$$

(remember $e^{a+b} = e^a e^b$, so $\frac{e^{a+b}}{e^a} = e^b$), i.e. $\beta_1 = \log(OR)$. Hence e^{β_1} is the relative increase in the odds of disease, going from $x_1 = 0$ to $x_1 = 1$ holding x_2 fixed (or *adjusting for* x_2).

To interpret β_2, fix the value of $\mathbf{x_1}$:
For $x_2 = k$ (any given value k)

$$\begin{aligned} \text{log odds of disease} &= \alpha + \beta_1 x_1 + \beta_2 k \\ \text{odds of disease} &= e^{\alpha+\beta_1 x_1+\beta_2 k} \end{aligned}$$

For $x_2 = k + 1$

$$\begin{aligned} \text{log odds of disease} &= \alpha + \beta_1 x_1 + \beta_2(k+1) \\ &= \alpha + \beta_1 x_1 + \beta_2 k + \beta_2 \\ \text{odds of disease} &= e^{\alpha+\beta_1 x_1+\beta_2 k+\beta_2} \end{aligned}$$

Thus the odds ratio (going from $x_2 = k$ to $x_2 = k + 1$ is

$$OR = \frac{\text{odds when } x_2 = k+1}{\text{odds when } x_2 = k} = \frac{e^{\alpha+\beta_1 x_1+\beta_2 k+\beta_2}}{e^{\alpha+\beta_1 x_1+\beta_2 k}} = e^{\beta_2}$$

i.e. $\beta_2 = \log(OR)$. Hence e^{β_2} is the relative increase in the odds of disease, going from $x_2 = k$ to $x_2 = k + 1$ holding x_1 fixed (or *adjusting for* x_1). Put another way, for every increase of 1 in x_2 the odds of disease increases by a factor of e^{β_2}. More generally, if you increase x_2 from k to $k + \Delta$ then

$$OR = \frac{\text{odds when } x_2 = k+\Delta}{\text{odds when } x_2 = k} = e^{\beta_2 \Delta} = \left(e^{\beta_2}\right)^{\Delta}$$

The Leukemia Data

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{ IAG} + \beta_2 \text{ LWBC}$$

where IAG is a binary variable and LWBC is a continuous predictor. **Stata** output seen earlier

```
------------------------------------------------------------------------------
        live |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iag |   2.519562   1.090681     2.31   0.021     .3818672    4.657257
        lwbc |  -1.108759   .4609479    -2.41   0.016      -2.0122   -.2053178
       _cons |   5.543349   3.022416     1.83   0.067     -.380477    11.46718
------------------------------------------------------------------------------
```

shows a fitted model of

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 5.54 + 2.52 \text{ IAG} - 1.11 \text{ LWBC}$$

The estimated (adjusted) OR for IAG is $e^{2.52} = 12.42$, which of course we saw earlier in the **Stata** output

```
------------------------------------------------------------------------------
        live | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iag |   12.42316    13.5497     2.31   0.021     1.465017    105.3468
        lwbc |   .3299682   .1520981    -2.41   0.016     .1336942    .8143885
------------------------------------------------------------------------------
```

The estimated odds that an Ag+ individual (IAG=1) survives at least one year is 12.42 greater than the corresponding odds for an Ag- individual (IAG=0), regardless of the LWBC (although the LWBC must be the same for both individuals).

The estimated OR for LWBC is $e^{-1.11} = .33 \quad (\approx \frac{1}{3})$. For each increase in 1 unit of LWBC, the estimated odds of surviving at least a year decreases by roughly a factor of 3, regardless of ones

IAG. Stated differently, if two individuals have the same Ag factor (either + or -) but differ on their values of LWBC by one unit, then the individual with the higher value of LWBC has about 1/3 the estimated odds of survival for a year as the individual with the lower LWBC value.

Confidence intervals for coefficients and ORs are related as before. For IAG the 95% CI for β_1 yields the 95% CI for the adjusted IAG OR as follows:

$$\begin{array}{rcccl} .382 & \leq & \beta_1 & \leq & 4.657 \\ e^{.382} & \leq & e^{\beta_1} & \leq & e^{4.657} \\ 1.465 & \leq & OR & \leq & 105.35 \end{array}$$

We estimate that the odds of an Ag+ individual (IAG=1) surviving at least a year to be 12.42 times the odds of an Ag- individual surviving at least one year. We are 95% confident the odds ratio is between 1.465 and 105.35. How does this compare with the unadjusted odds ratio?

Similarly for LWBC, the 95% CI for β_2 yields the 95% CI for the adjusted LWBC OR as follows:

$$\begin{array}{rcccl} -2.012 & \leq & \beta_2 & \leq & -.205 \\ e^{-2.012} & \leq & e^{\beta_2} & \leq & e^{-.205} \\ .134 & \leq & OR & \leq & .814 \end{array}$$

We estimate the odds of surviving at least a year is reduced by a factor of 3 (i.e. 1/3) for each increase of 1 LWBC unit. We are 95% confindent the reduction in odds is between .134 and .814.

Note that while this is the usual way of defining the OR for a continuous predictor variable, software may try to trick you. JMP IN for instance would report

$$\widehat{OR} = e^{-1.11(\max(LWBC)-\min(LWBC))} = .33^{\max(LWBC)-\min(LWBC)},$$

the change from the smallest to the largest LWBC. That is a lot smaller number. You just have to be careful and check what is being done by knowing these relationships.

General Model

We can have a lot more than complicated models than we have been analyzing, but the principles remain the same. Suppose we have k predictor variables where k can be considerably more than 2 and the variables are a mix of binary and continuous. then we write

$$\log\left(\frac{p}{1-p}\right) = \text{log odds of disease} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

which is a logistic multiple regression model. Now fix values of $x_2, x_3, \ldots, x_k$, and we get

$$\begin{array}{rl} \text{odds of disease for } x_1 = c: & e^{\alpha+\beta_1 c+\beta_2 x_2+\ldots+\beta_k x_k} \\ x_1 = c+1: & e^{\alpha+\beta_1(c+1)+\beta_2 x_2+\ldots+\beta_k x_k} \end{array}$$

The odds ratio, increasing x_1 by 1 and holding $x_2, x_3, \ldots, x_k$ fixed at any values is

$$OR = \frac{e^{\alpha+\beta_1(c+1)+\beta_2 x_2+\ldots+\beta_k x_k}}{e^{\alpha+\beta_1 c+\beta_2 x_2+\ldots+\beta_k x_k}} = e^{\beta_1}$$

That is, e^{β_1} is the increase in odds of disease obtained by increasing x_1 by 1 unit, holding $x_2, x_3, \ldots, x_k$ fixed (i.e. adjusting for levels of $x_2, x_3, \ldots, x_k$). For this to make sense

- x_1 needs to be binary or continuous
- None of the remaining effects $x_2, x_3, \ldots, x_k$ can be an interaction (product) effect with x_1. I will say more about this later! The essential problem is that if one or more of $x_2, x_3, \ldots, x_k$ depends upon x_1 then you cannot mathematically increase x_1 and simultaneously hold $x_2, x_3, \ldots, x_k$ fixed.

Example: The UNM Trauma Data

The data to be analyzed here were collected on 3132 patients admitted to The University of New Mexico Trauma Center between the years 1991 and 1994. For each patient, the attending physician recorded their age, their revised trauma score (RTS), their injury severity score (ISS), whether their injuries were blunt (i.e. the result of a car crash: BP=0) or penetrating (i.e. gunshot wounds: BP=1), and whether they eventually survived their injuries (DEATH = 1 if died, DEATH = 0 if survived). Approximately 9% of patients admitted to the UNM Trauma Center eventually die from their injuries.

The ISS is an overall index of a patient's injuries, based on the approximately 1300 injuries cataloged in the Abbreviated Injury Scale. The ISS can take on values from 0 for a patient with no injuries to 75 for a patient with 3 or more life threatening injuries. The ISS is the standard injury index used by trauma centers throughout the U.S. The RTS is an index of physiologic injury, and is constructed as a weighted average of an incoming patient's systolic blood pressure, respiratory rate, and Glasgow Coma Scale. The RTS can take on values from 0 for a patient with no vital signs to 7.84 for a patient with normal vital signs.

Champion et al. (1981) proposed a logistic regression model to estimate the probability of a patient's survival as a function of RTS, the injury severity score ISS, and the patient's age, which is used as a surrogate for physiologic reserve. Subsequent survival models included the binary effect BP as a means to differentiate between blunt and penetrating injuries. We will develop a logistic model for predicting *death* from ISS, AGE, BP, and RTS.

Figure 1 shows side-by-side boxplots of the distributions of ISS, AGE, and RTS for the survivors and non-survivors, and a bar chart showing proportion penetrating injuries for survivors and non-survivors. Survivors tend to have lower ISS scores, tend to be slightly younger, and tend to have higher RTS scores, than non-survivors. The importance of the effects individually towards predicting survival is directly related to the separation between the survivors and non-survivors scores. There are no dramatic differences in injury type (BP) between survivors and non-survivors.

Figure 1 was generated with the following Stata code. Earlier in the semester I was avoiding using the `relabel` option; it is much better to do things this way, but note the 1 and 2 refer to alphabetic order of values, not to the actual values. Bar graphs in Stata are a little tricky – this one worked, but had there been several values of BP or had they been coded other than 0 and 1 this would not have worked. In the latter case one needs to create separate indicator variables of categories (as an option to tabulate): See
`http://www.stata.com/support/faqs/graphics/piechart.html` for a discussion.

```
graph box iss, over(death, relabel(1 "Survived" 2 "Died" ) descending) ///
      ytitle(ISS) title(ISS by Death) name(iss)
graph box rts, over(death, relabel(1 "Survived" 2 "Died" ) descending) ///
      ytitle(RTS) title(RTS by Death) name(rts)
graph box age, over(death, relabel(1 "Survived" 2 "Died" ) descending) ///
      ytitle(Age) title(Age by Death) name(age)
graph bar bp,over(death,relabel(1 "Survived" 2 "Died") descending) ///
      ytitle("Proportion Penetrating") title("Penetrating by Death") name(bp)
graph combine iss rts age bp
```

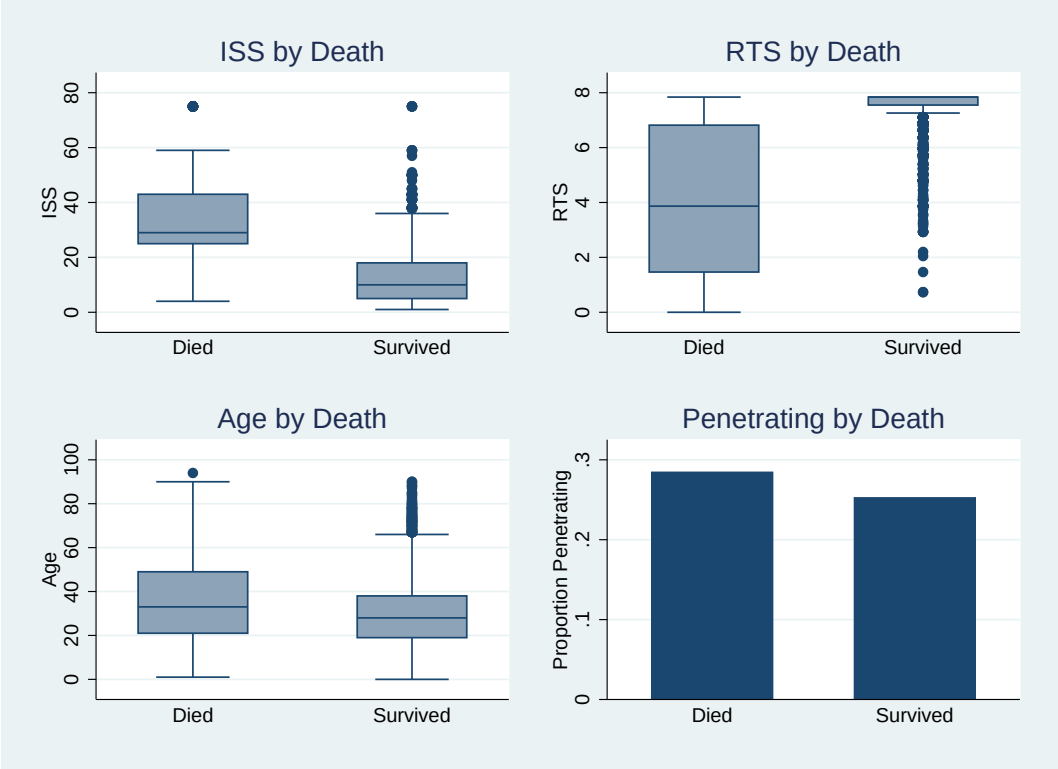


Figure 1: Relationship of predictor variables to death

Stata Analysis of Trauma Data

```
. logistic death iss bp rts age,coef
Logistic regression                                     Number of obs   =       3132
                                                        LR chi2(4)      =     933.34
                                                        Prob > chi2     =     0.0000
Log likelihood = -446.01414                             Pseudo R2       =     0.5113
------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   .0651794   .0071603     9.10   0.000     .0511455    .0792134
          bp |   1.001637    .227546     4.40   0.000     .5556555    1.447619
         rts |  -.8126968   .0537066   -15.13   0.000    -.9179597   -.7074339
         age |    .048616   .0052318     9.29   0.000     .0383619     .05887
       _cons |  -.5956074   .4344001    -1.37   0.170    -1.447016    .2558011
------------------------------------------------------------------------------
. logistic death iss bp rts age
Logistic regression                                     Number of obs   =       3132
                                                        LR chi2(4)      =     933.34
                                                        Prob > chi2     =     0.0000
Log likelihood = -446.01414                             Pseudo R2       =     0.5113
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.067351   .0076426     9.10   0.000     1.052476    1.082435
          bp |   2.722737   .6195478     4.40   0.000     1.743083    4.252978
         rts |     .44366   .0238275   -15.13   0.000      .399333    .4929074
         age |   1.049817   .0054924     9.29   0.000     1.039107    1.060637
------------------------------------------------------------------------------
. estat gof
Logistic model for death, goodness-of-fit test
       number of observations =      3132
 number of covariate patterns =      2096
```

```
          Pearson chi2(2091) =      2039.73
                Prob > chi2 =         0.7849
. estat gof,group(10)
Logistic model for death, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
      number of observations =      3132
            number of groups =        10
     Hosmer-Lemeshow chi2(8) =        10.90
                Prob > chi2 =         0.2072
```

There are four effects in our model: ISS, BP (a binary variable), RTS, and AGE. Looking at the goodness of fit tests, there is no evidence of gross deficiencies with the model. The small p-value ($< .0001$) for the LR chi-squared statistic implies that one or more of the 4 effects in the model is important for predicting the probability of death. The tests for parameters suggest that each of the effects in the model is significant at the .001 level (p-values $< .001$).

The fitted logistic model is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -.596 + .065\text{ISS} + 1.002\text{BP} - .813\text{RTS} + .049\text{AGE},$$

where $\hat{p}$ is the estimated probability of death.

The table below is in a form similar to Fisher et al's AJPH article (with this lecture). The estimated odds ratio was obtained by exponentiating the regression estimate. The CI endpoints for the ORs were obtained by exponentiating the CI endpoints for the corresponding regression parameter. JMP-IN (and some authors) would report different ORs for the continuous variables, for instance 124.37 for ISS (instead of the 1.067 we are reporting). (Why?). Everybody will agree on the coefficient, but you need to be very careful what OR is being reported and how you interpret it.

The p-value for each regression effect is smaller than .05, so the 95% CI for each OR excludes 1 (i.e. each regression coefficient is significantly different from zero so each OR is significantly different from 1). Thus, for example, the odds of dying from a penetrating injury (BP=1) is 2.72 times greater than the odds of dying from a blunt trauma (BP=0). We are 95% confident that the population odds ratio is between 1.74 and 4.25.

Do the signs of the estimated regression coefficients make sense? That is, which coefficients would you expect to be positive (leading to an OR greater than 1).

Effect	Estimate	Std Error	P-value	Odds Ratio	95% CI
ISS	.065	.007	$< .001$	1.067	(1.052 , 1.082)
BP	1.002	.228	$< .001$	2.723	(1.743 , 4.253)
RTS	-.813	.054	$< .001$	0.444	(0.399 , 0.493)
AGE	.049	.005	$< .001$	1.050	(1.039 , 1.061)

Logistic Models with Interactions

Consider the hypothetical problem with two binary predictors x_1 and x_2

	$x_2 = 0$		$x_2 = 1$	
	x_1		x_1	
Disease	1	0	1	0
+	1	9	9	1
–	45	45	45	45

The OR for $x_1 = 1$ versus $x_1 = 0$ when $x_2 = 0$: $\widehat{OR} = \frac{1(45)}{9(45)} = \frac{1}{9}$
The OR for $x_1 = 1$ versus $x_1 = 0$ when $x_2 = 1$: $\widehat{OR} = \frac{9(45)}{1(45)} = 9$

A simple logistic model for these data is $\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2$. For this model, OR for $x_1 = 1$ versus $x_1 = 0$ for fixed x_2 is e^{β_1}. That is, the adjusted OR for x_1 is *independent of the value of* x_2. This model would appear to be inappropriate for the data set above where the OR of x_1 is very different for $x_2 = 0$ than it is for $x_2 = 1$.

A simple way to allow for the odds ratio to depend on the level of x_2 is through the interaction model

$$\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \; x_1 * x_2$$

where the *interaction* term $x_1 * x_2$ is the product (in this case) of x_1 and x_2. In some statistical packages the interaction variable must be created in the spreadsheet (that always works), and in others it can (much more conveniently) be added to the model directly. **Stata** is in the former category, although the `xi` structure allows interaction terms to be generated automatically. That becomes much more important with multi-level (3 or more) factors.

To interpret the model, let us consider the 4 possible combinations of the binary variables:

Group	x_1	x_2	$x_1 * x_2$
A	0	0	0
B	0	1	0
C	1	0	0
D	1	1	1

Group	Log Odds of Disease	Odds of Disease
A	$\alpha + \beta_1(0) + \beta_2(0) + \beta_3(0) = \alpha$	e^{α}
B	$\alpha + \beta_1(0) + \beta_2(1) + \beta_3(0) = \alpha + \beta_2$	$e^{\alpha+\beta_2}$
C	$\alpha + \beta_1(1) + \beta_2(0) + \beta_3(0) = \alpha + \beta_1$	$e^{\alpha+\beta_1}$
D	$\alpha + \beta_1(1) + \beta_2(1) + \beta_3(1) = \alpha + \beta_1 + \beta_2 + \beta_3$	$e^{\alpha+\beta_1+\beta_2+\beta_3}$

Group A is the baseline or reference group. The parameters α, β_1, and β_2 are easily interpreted. The odds of disease for the baseline group ($x_1 = x_2 = 0$) is e^{α} – the same interpretation applies when interaction is absent. To interpret β_1 note OR for Group C vs. Group A is $\frac{e^{\alpha+\beta_1}}{e^{\alpha}} = e^{\beta_1}$. This is OR for $x_1 = 1$ vs. $x_1 = 0$ when $x_2 = 0$. Similarly OR for Group B vs. Group A is $\frac{e^{\alpha+\beta_2}}{e^{\alpha}} = e^{\beta_2}$. This is OR for $x_2 = 1$ vs. $x_2 = 0$ when $x_1 = 0$.

In an interaction model, the OR for $x_1 = 1$ vs. $x_1 = 0$ depends on the level of x_2. Similarly the OR for $x_2 = 1$ vs. $x_2 = 0$ depends on the level of x_1. For example,

$$OR \text{ for group D vs. B } = \frac{e^{\alpha+\beta_1+\beta_2+\beta_3}}{e^{\alpha+\beta_2}} = e^{\beta_1+\beta_3}$$

This is OR for $x_1 = 1$ vs. $x_1 = 0$ when $x_2 = 1$. Recalling that e^{β_1} is OR for $x_1 = 1$ vs. $x_1 = 0$ when $x_2 = 0$, we have

$$\begin{array}{ccccc} OR(x_1 = 1 \text{ vs. } x_1 = 0 \text{ when } x_2 = 1) & = & OR(x_1 = 1 \text{ vs. } x_1 = 0 \text{ when } x_2 = 0) & * & e^{\beta_3} \\ e^{\beta_1+\beta_3} & = & e^{\beta_1} & * & e^{\beta_3} \end{array}$$

Thus e^{β_3} is the factor that relates the OR for $x_1 = 1$ vs. $x_1 = 0$ when $x_2 = 0$ to the OR when $x_2 = 1$. If $\beta_3 = 0$ the two OR are identical, i.e. x_1 and x_2 do not interact. Similarly,

$$\begin{array}{ccccc} OR(x_2 = 1 \text{ vs. } x_2 = 0 \text{ when } x_1 = 1) & = & OR(x_2 = 1 \text{ vs. } x_2 = 0 \text{ when } x_1 = 0) & * & e^{\beta_3} \\ e^{\beta_2+\beta_3} & = & e^{\beta_2} & * & e^{\beta_3} \end{array}$$

so e^{β_3} is also the factor that relates the OR for $x_2 = 1$ vs. $x_2 = 0$ at the two levels of x_1. An important and no doubt fairly obvious point to take away from this is that the regression coefficients are harder to interpret in models with interactions!

Stata Analysis: Let's fit this interaction example (data from page 118) using **Stata**. We could actually do this particular example easily without using `xi`, but we won't be so lucky in the future.

```
. list,clean
       x2   x1   Disease   Count
  1.    0    1         1       1
  2.    0    1         0      45
  3.    0    0         1       9
  4.    0    0         0      45
  5.    1    1         1       9
  6.    1    1         0      45
  7.    1    0         1       1
  8.    1    0         0      45

. xi: logistic Disease i.x1 i.x2  i.x1*i.x2 [fw=Count],coef
i.x1              _Ix1_0-1            (naturally coded; _Ix1_0 omitted)
i.x2              _Ix2_0-1            (naturally coded; _Ix2_0 omitted)
i.x1*i.x2         _Ix1Xx2_#_#         (coded as above)
note: _Ix1_1 dropped due to collinearity
note: _Ix2_1 dropped due to collinearity
Logistic regression                               Number of obs   =        200
                                                  LR chi2(3)      =      13.44
                                                  Prob > chi2     =     0.0038
Log likelihood = -58.295995                       Pseudo R2       =     0.1034
------------------------------------------------------------------------------
     Disease |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Ix1_1 |  -2.197225   1.074892    -2.04   0.041    -4.303975    -.090474
      _Ix2_1 |  -2.197225   1.074892    -2.04   0.041    -4.303975    -.090474
 _Ix1Xx2_1_1 |   4.394449   1.520128     2.89   0.004     1.415054    7.373844
       _cons |  -1.609438   .3651484    -4.41   0.000    -2.325116   -.8937603
------------------------------------------------------------------------------

. xi: logistic Disease i.x1 i.x2  i.x1*i.x2 [fw=Count]
i.x1              _Ix1_0-1            (naturally coded; _Ix1_0 omitted)
i.x2              _Ix2_0-1            (naturally coded; _Ix2_0 omitted)
i.x1*i.x2         _Ix1Xx2_#_#         (coded as above)
note: _Ix1_1 dropped due to collinearity
note: _Ix2_1 dropped due to collinearity
Logistic regression                               Number of obs   =        200
                                                  LR chi2(3)      =      13.44
                                                  Prob > chi2     =     0.0038
Log likelihood = -58.295995                       Pseudo R2       =     0.1034
------------------------------------------------------------------------------
     Disease | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Ix1_1 |   .1111111   .1194325    -2.04   0.041     .0135147     .913498
      _Ix2_1 |   .1111111   .1194325    -2.04   0.041     .0135147     .913498
 _Ix1Xx2_1_1 |         81   123.1303     2.89   0.004     4.116709    1593.749
------------------------------------------------------------------------------
```

The fitted model is

$$\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3\, x_1 * x_2 = -1.61 - 2.20x_1 - 2.20x_2 + 4.39x_1 * x_2$$

Note that $e^{\hat{\beta}_1} = e^{-2.20} = \frac{1}{9} =$ estimated OR for $x_1 = 1$ vs. $x_1 = 0$ when $x_2 = 0$. Also,

$$e^{\hat{\beta}_1 + \hat{\beta}_3} = e^{-2.20+4.39} = e^{2.19} = 9 = \text{ estimated } OR \text{ for } x_1 = 1 \text{ vs. } x_1 = 0 \text{ when } x_2 = 1$$

Note that

$$e^{\hat{\beta}_3} = e^{4.39} = 81 = \text{ mult. factor that relates } OR \text{ for } x_1 = 1 \text{ vs. } x_1 = 0 \text{ at the 2 levels of } x_2$$

Make sure you see how **Stata** agrees with these calculations.

A More Complex Interaction Model

The treatment regime to be adopted for patients who have been diagnosed as having prostate cancer is crucially dependent on whether the cancer has spread to the surrounding lymph nodes. A laparatomy (a surgical incision into the abdominal cavity) may be performed to ascertain the extent of this nodal involvement. There are a number of variables that are indicative of nodal involvement which can be measured without surgery. The aim of the study for which the data were collected was to determine if a combination of 5 variables could be used to predict whether cancer has spread to the lymph nodes. The 5 variables are: age of patient at diagnosis (years), level of serum acid phosphatase (in King-Armstrong units), result of X-ray examination (0=negative, 1=positive), size of tumor by rectal examination (0=small, 1=large), and a summary of pathological grade of tumor from biopsy (0=less serious, 1=serious). The response variable is involvement of lymph node (0=no, 1=yes). Fifty-three patients were enrolled in the study.

A published analysis suggested the following model for the probability p of nodal involvement

$$\begin{aligned}\log\left(\tfrac{p}{1-p}\right) = \alpha + \beta_1\ \text{Xray} + \beta_2\ \text{size} + \beta_3\ \text{grade} + \beta_4\ \log(\text{acid}) \\ + \beta_5\ \text{size*grade} + \beta_6\ \log(\text{acid})\text{*grade}\end{aligned}$$

The model contains **3** binary variables (Xray, size, and grade), **1** continuous variable (log(acid)), and **2** interactions, or product effects (size*grade) and log(acid)*grade). the size*grade interaction involves two binary variables, as considered in the previous example, whereas the log(acid)*grade interaction term involves a binary and a continuous variable. for each case in the data set

$$\log(\text{acid})\text{*grade} = \begin{cases} 0 & \text{if grade} = 0 \\ \log(\text{acid}) & \text{if grade} = 1 \end{cases}$$

Note that the model excludes age.

Interpreting the Regression Coefficients

For any regression variable that is *not included* in an interaction, the regression coefficient is an adjusted log OR, and is independent of levels of the other factors in the model. For example, for fixed size, grade, and acid levels

$$(\text{OR for Xray} = 1 \text{ vs. Xray} = 0) = \frac{e^{\alpha+\beta_1(1)+\beta_2\ \text{size}\ +\ \cdots}}{e^{\alpha+\beta_1(0)+\beta_2\ \text{size}\ +\ \cdots}} = e^{\beta_1}$$

The size*grade interaction means that the adjusted OR for size = 1 vs. size = 0 depends on grade. The log(acid)*grade interaction means that the adjusted OR for log(acid) depends on grade. To see this, let LA = log(acid). Then odds of nodal involvement = $e^{\alpha+\beta_1\text{Xray}+\beta_2\text{size}+\beta_3\text{grade}+\beta_4\text{LA}+\beta_5\text{size}*\text{grade}+\beta_6\text{LA}*g}$ so for fixed Xray, size, and grade

$$\begin{aligned}\frac{\text{odds of nodal involvment at LA + 1}}{\text{odds of nodal involvment at LA}} &= \frac{exp(\alpha+\beta_1\text{Xray}+\beta_2\text{size}+\beta_3\text{grade}+\beta_4(\text{LA}+1)+\beta_5\text{size}*\text{grade}+\beta_6(\text{LA}+1)*\text{grade})}{exp(\alpha+\beta_1\text{Xray}+\beta_2\text{size}+\beta_3\text{grade}+\beta_4\text{LA}+\beta_5\text{size}*\text{grade}+\beta_6\text{LA}*\text{grade})} \\ &= e^{\beta_4+\beta_6\text{grade}} \\ &= \begin{cases} e^{\beta_4} & \text{grade} = 0 \\ e^{\beta_4+\beta_6} & \text{grade} = 1 \end{cases}\end{aligned}$$

This adjusted OR depends on grade (because LA and grade interact), but not on size or Xray (because LA does not interact with either). We can interpret β_6, the LA*grade coefficient, as a measure of how the adjusted OR for LA changes with grade.

Given that the model contains a size*grade and a log(acid)*grade interaction, the adjusted OR for grade depends on the size and log(acid) levels. I'll note, but you can easily show,

$$\frac{\text{odds for nodal involvement for grade} = 1}{\text{odds for nodal involvement for grade} = 0} = e^{\beta_3 + \beta_5 \text{size} + \beta_6 \log(\text{acid})}$$

where β_5 is the grade*size coefficient and β_6 is the log(acid)*grade coefficient.

In summary, interactions among variables make interpretations of effects of individual variables on OR harder (OK, *lots* harder!) The ideal world has no interactions — but we don't live in such a world.

Stata Analysis

Raw data are available on the web page. Output from fitting the model in **Stata** follows:

```
. gen logacid=log(acid)
. xi: logistic nodal i.xray i.size i.grade logacid i.size*i.grade i.grade*logacid
i.xray            _Ixray_0-1          (naturally coded; _Ixray_0 omitted)
i.size            _Isize_0-1          (naturally coded; _Isize_0 omitted)
i.grade           _Igrade_0-1         (naturally coded; _Igrade_0 omitted)
i.size*i.grade    _IsizXgra_#_#       (coded as above)
i.grade*logacid   _IgraXlogac_#       (coded as above)
note: _Isize_1 dropped due to collinearity
note: _Igrade_1 dropped due to collinearity
note: _Igrade_1 dropped due to collinearity
note: logacid dropped due to collinearity
Logistic regression                               Number of obs   =         53
                                                  LR chi2(6)      =      33.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -18.143573                       Pseudo R2       =     0.4835
------------------------------------------------------------------------------
       nodal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Ixray_1 |   10.38589   11.26382     2.16   0.031     1.239622    87.01579
    _Isize_1 |   23.05661   26.99148     2.68   0.007     2.324485    228.6989
   _Igrade_1 |   21187.36   98875.09     2.13   0.033     2.258257    1.99e+08
     logacid |   5.520827   7.841721     1.20   0.229     .3411671    89.33903
_IsizXgra_~1 |   .0035255   .0085831    -2.32   0.020     .0000298    .4164339
_IgraXloga~1 |   33724.72   223942.7     1.57   0.116     .0751111    1.51e+10
------------------------------------------------------------------------------
. xi: logistic nodal i.xray i.size i.grade logacid i.size*i.grade i.grade*logacid,coef
i.xray            _Ixray_0-1          (naturally coded; _Ixray_0 omitted)
i.size            _Isize_0-1          (naturally coded; _Isize_0 omitted)
i.grade           _Igrade_0-1         (naturally coded; _Igrade_0 omitted)
i.size*i.grade    _IsizXgra_#_#       (coded as above)
i.grade*logacid   _IgraXlogac_#       (coded as above)
note: _Isize_1 dropped due to collinearity
note: _Igrade_1 dropped due to collinearity
note: _Igrade_1 dropped due to collinearity
note: logacid dropped due to collinearity
Logistic regression                               Number of obs   =         53
                                                  LR chi2(6)      =      33.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -18.143573                       Pseudo R2       =     0.4835
------------------------------------------------------------------------------
       nodal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Ixray_1 |   2.340448   1.084531     2.16   0.031     .2148063     4.46609
    _Isize_1 |   3.137952   1.170661     2.68   0.007     .8434986    5.432406
   _Igrade_1 |    9.96116   4.666701     2.13   0.033     .8145935    19.10773
     logacid |   1.708528   1.420389     1.20   0.229    -1.075383    4.492438
_IsizXgra_~1 |  -5.647741   2.434592    -2.32   0.020    -10.41945   -.8760275
_IgraXloga~1 |   10.42599   6.640313     1.57   0.116    -2.588787    23.44076
       _cons |  -2.552712   1.039703    -2.46   0.014    -4.590494   -.5149311
------------------------------------------------------------------------------
```

```
***************** Make sure you understand what variables are being fit!
-------------------------------------
variable name     variable label
-------------------------------------
_Ixray_1           xray==1
_Isize_1           size==1
_Igrade_1          grade==1
_IsizXgra_1_1      size==1 & grade==1
_IgraXlogac_1     (grade==1)*logacid
-------------------------------------
```

Note that I did not actually need to use xi here since the variables were already binary and coded as 0 and 1, but this is the safe way to do things. The fitted model is

$$\begin{aligned}\log\left(\frac{\hat{p}}{1-\hat{p}}\right) &= -2.55 + 2.34\ \text{Xray} + 3.14\ \text{size} + 9.96\ \text{grade} + 1.71\ \log(\text{acid}) \\ &\quad -5.65\ \text{size} * \text{grade} + 10.43\ \log(\text{acid}) * \text{grade}\end{aligned}$$

If a primary question was the impact of a positive Xray, we can conclude that for fixed levels of size, grade, and log(acid)

$$\widehat{OR} \text{ for Xray} = 1 \text{ vs. Xray} = 0 \text{ is } e^{2.34} = 10.39$$

i.e. the odds of nodal involvement are 10.39 times higher for patients with positive X-rays than for patients with a negative X-rays (adjusting for size, grade, and log(acid)). The lack of interaction makes this a clean interpretation.

If a primary question was the impact of log(acid) (LA) level, then for fixed size tumor and X-ray result, recalling 1.709 is the LA coefficient and 10.43 is the grade coefficient,

$$\begin{aligned}\widehat{OR} \text{ for LA} + 1 \text{ vs. LA is } & e^{1.709+10.43*\text{grade}} \\ &= \begin{cases} e^{1.709} & = 5.52 \quad \text{if grade} = 0 \\ e^{1.709+10.43} & = 186,838 \quad \text{if grade} = 1 \end{cases}\end{aligned}$$

For less serious tumors (grade = 0) the odds of nodal involvement increase by 5.52 for each increase in 1 LA unit. For more serious tumors (grade=1) the odds increase by 186,838.

Remark: The log(acid)*grade interaction is not significant at the 10% level (p-value = .116). An implication is that the estimated adjusted OR for log(acid) when grade = 1 (i.e. 186,838) is not statistically different from the adjusted OR for log(acid) when grade = 0 (i.e. 5.52) — why? Because in a model without the log(acid)*grade interaction, those estimated ORs would be equal.

A sensible strategy would be to refit the model without this interaction. We will discuss such strategies later.

12 Odds Ratios for Multi-level Factors; Examples

The Framingham Study

The Framingham study was a prospective (follow-up, cohort) study of the occurrence of coronary heart disease (CHD) in Framingham, Mass. The study involved 2187 men and 2669 women aged between 30 and 62. More details on the study are given as a supplement to the lecture. Variables and values of the variables are as follows:

Variable Name	Codes
Gender	0 = Female, 1 = male
Age Group	0 is 30-49, 1 is 50-62
SCL (Serum Cholesterol)	1 is < 190, 2 is 190-219, 3 is 220-249, 4 is 250+
CHD (Coronary Heart Disease)	1 is Yes, 0 is No
Freq	Count

I will consider a simple analysis of the association between serum cholesterol level (SCL) at the start of the study and whether a subject had, or developed CHD, during the 12 year follow-up period. A table with **Stata** analysis of counts relating CHD to SCL is given below.

```
. tabulate chd scl [fw=frequency],chi2 lrchi2 exp col
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
| column percentage  |
+--------------------+

           |                        SCL
       CHD |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         0 |     1,022      1,203      1,119      1,125 |     4,469
           |     978.3    1,169.7    1,127.4    1,193.6 |   4,469.0
           |     96.14      94.65      91.35      86.74 |     92.03
-----------+--------------------------------------------+----------
         1 |        41         68        106        172 |       387
           |      84.7      101.3       97.6      103.4 |     387.0
           |      3.86       5.35       8.65      13.26 |      7.97
-----------+--------------------------------------------+----------
     Total |     1,063      1,271      1,225      1,297 |     4,856
           |   1,063.0    1,271.0    1,225.0    1,297.0 |   4,856.0
           |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(3) =  86.7040   Pr = 0.000
 likelihood-ratio chi2(3) =  85.8644   Pr = 0.000
```

The Pearson χ^2 statistic, which can be viewed as testing that the probability of developing CHD is independent of SCL, is highly significant (p-value < .001). Clearly observed counts of CHD are below expected counts for this hypothesis with low SCL, and above with high SCL, so it looks like CHD increases as SCL increases.

Let us do a closer look at the data for CHD vs. SCL using odds ratios. There are a lot of possible ways to do this. Since SCL categories are ordered, many analysts would compare SCL level 2 to 1, then 3 to 2, then 4 to 3. It is a little more conventional (and slightly more direct to implement in **Stata**) to consider all OR relative to a fixed baseline SCL category, say SCL < 190 (Cat. 1).

	SCL		
CHD	2	1	
Y	68	41	
N	1203	1022	$\widehat{OR}(2\text{vs.}1) = \frac{68 \cdot 1022}{41 \cdot 1203} = 1.409$

	3	1	
Y	106	41	
N	1119	1022	$\widehat{OR}(3\text{vs.}1) = \frac{106 \cdot 1022}{41 \cdot 1119} = 2.361$

	4	1	
Y	172	41	
N	1125	1022	$\widehat{OR}(4\text{vs.}1) = \frac{172 \cdot 1022}{41 \cdot 1125} = 3.811$

Any OR may be computed from this set of OR's. For example,

	SCL		
CHD	4	2	
Y	172	68	
N	1125	1203	$\widehat{OR}(4\text{vs.}2) = \frac{172 \cdot 1203}{1125 \cdot 68} = 2.705 = \frac{3.811}{1.409} = \frac{\widehat{OR}(4\text{vs.}1)}{\widehat{OR}(2\text{vs.}1)}$

Think of this relationship as $\frac{4}{2} = \frac{4/1}{2/1}$. An important point to recognize is that the effect of SCL on CHD can be captured through 3 effects (ORs), which is #SC levels - 1.

To get these ORs directly from **Stata**, we need to use `xi`. Actually, there are other, better, options you can download and install, like `xi3` and `desmat`. Since `xi` is built-in and commonly used, we will stick with it but it does not allow higher order interaction terms in models, unlike `xi3` and `desmat`.

The code and output follow:

```
. xi:logistic chd i.scl [fweight=frequency]
 i.scl            _Iscl_1-4        (naturally coded;_Iscl_1 omitted)
Logistic regression                               Number of obs   =       4856
                                                  LR chi2(3)      =      85.86
                                                  Prob > chi2     =     0.0000
Log likelihood = -1307.1541                       Pseudo R2       =     0.0318
------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iscl_2 |   1.408998   .2849726     1.70   0.090     .9478795    2.094438
     _Iscl_3 |   2.361255    .446123     4.55   0.000     1.630502    3.419514
     _Iscl_4 |   3.811035   .6825005     7.47   0.000     2.682905    5.413532
------------------------------------------------------------------------------
. xi:logistic chd i.scl [fweight=frequency],coef
------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iscl_2 |   .3428787    .202252     1.70   0.090    -.0535279    .7392852
     _Iscl_3 |   .8591931   .1889347     4.55   0.000     .4888878    1.229498
     _Iscl_4 |   1.337901   .1790853     7.47   0.000        .9869    1.688902
       _cons |  -3.215945   .1592756   -20.19   0.000    -3.528119    -2.90377
------------------------------------------------------------------------------
```

Remember what `xi` is doing. It creates indicator variables with the first omitted, so we are fitting a model for p = probability of CHD of

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_j \text{ \_Iscl\_j, where I\_scl\_j} = \begin{cases} 1 & \text{SCL} = j, j \geq 2 \\ 0 & \text{SCL} \neq j, j \geq 2 \\ 0 & j = 1 \text{ (i.e.naturally coded; \_Iscl\_1 omitted)} \end{cases}$$

and proceeding as in the last lecture, for $j > 1$,

$$\begin{aligned} \beta_j = & \ (\alpha + \beta_j) - \alpha \\ = & \ \log(\text{Odds for SCL} = j) - \log(\text{Odds for SCL} = 1) \\ = & \ \log(\text{OR(for SCL} = \text{j vs. SCL} = 1)) \end{aligned}$$

which yields the result that

$$e^{\beta_j} = \text{OR(for SCL} = \text{j vs. SCL} = 1) \qquad j > 1$$

with confidence intervals for ORs produced by exponentiating limits of confidence intervals for coefficients. The **Stata** output above gives us exactly the values of $\widehat{OR}$(2vs.1), $\widehat{OR}$(3vs.1), and $\widehat{OR}$(4vs.1) we calculated previously, along with confidence limits. We also saw that $\widehat{OR}(4\text{vs.}2) = \frac{\widehat{OR}(4\text{vs.}1)}{\widehat{OR}(2\text{vs.}1)} = \frac{3.811}{1.409} = 2.705$ but this does not produce a confidence interval for OR(4vs.2). In order to get full information about this OR, note that

$$\frac{\widehat{OR}(4\text{vs.}1)}{\widehat{OR}(2\text{vs.}1)} = \frac{e^{\beta_4}}{e^{\beta_2}} = e^{\beta_4 - \beta_2}$$

This looks like `lincom` should work, and it is exactly the solution.

```
. lincom _b[_Iscl_4] - _b[_Iscl_2]
 ( 1) - _Iscl_2 + _Iscl_4 = 0
------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.704784   .4033665     6.67   0.000      2.01926    3.623039
------------------------------------------------------------------------------
```

`lincom` reports OR after `logistic`. If you actually want the difference in coefficients, you need to use the `logit` form of the command, and then `lincom` reports

```
. lincom _b[_Iscl_4] - _b[_Iscl_2]
 ( 1) - _Iscl_2 + _Iscl_4 = 0
------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .9950222   .1491307     6.67   0.000     .7027313    1.287313
------------------------------------------------------------------------------
```

This section has shown you how to generate unadjusted ORs in **Stata**. In practice we would add confounding variables, such as Age and Sex, to the model and then evaluate adjusted ORs for the SCL levels. You will get to do this in lab.

Model Building

There are a variety of systematic approaches to logistic regression models. Automated methods such as the backward elimination approach described below are well suited for producing good predictive models. Systematic approaches such as those advocated in Kleinbaum's book on Logistic Regression focus more attention on understanding the complex interdependencies among the predictors, and their impact on odds ratios.

Backward Elimination

1. Identify predictors or factors for which an association with outcome is biologically plausible (based on literature, science, knowledge, etc.).
2. Identify biologically plausible interactions.
3. Fit the logistic model with all candidate effects identified in the first 2 steps.
4. Following the *hierarchy principle*, identify the least significant effect in the model, and sequentially eliminate the least significant effect until the step where the least significant effect is "too important" to omit.

- The *hierarchy principle* implies that a main effect in a model can only be considered for exclusion if the model does not contain an interaction involving the main effect.
- The impact of an effect is measured by a p-value for testing that the regression coefficient for the effect is zero.

$$\text{p} - \text{value} \begin{cases} \leq \alpha & \text{effect stays in model} \\ > \alpha & \text{effect is removed} \end{cases}$$

In backwards elimination it is not uncommon to set $\alpha = .10$ or .15 rather than $\alpha = .05$.

Example: UNM Trauma Data

Response	:	Death (1=yes, 0 = no)
Predictors	:	ISS
		Age
		RTS
		BP (0=Blunt, 1=Penetratin)

The surgeon who collected the data, Dr. Turner Osler, believes that all these effects are associated with the probability of death and that the three interactions involving BP (BP*ISS, BP*Age, BP*RTS) are plausible.

Steps

0. Fit full model

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{ISS} + \beta_2 \text{BP} + \beta_3 \text{RTS} + \beta_4 \text{Age} + \beta_5(\text{BP} * \text{ISS}) + \beta_6(\text{BP} * \text{RTS}) + \beta_7(\text{BP} * \text{Age})$$

where p=probability of death from injuries. **Stata** does not allow specification of interaction terms directly with `logit` or `logistic`, so we need to use `xi`.

```
. xi:logistic death iss i.bp rts age i.bp*iss i.bp*rts i.bp*age
i.bp              _Ibp_0-1          (naturally coded; _Ibp_0 omitted)
i.bp*iss          _IbpXiss_#        (coded as above)
i.bp*rts          _IbpXrts_#        (coded as above)
i.bp*age          _IbpXage_#        (coded as above)
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(7)      =     937.59
                                                  Prob > chi2     =     0.0000
Log likelihood = -443.88603                       Pseudo R2       =     0.5136
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.070319   .0090198     8.06   0.000     1.052785    1.088144
         age |   1.047169   .0058718     8.22   0.000     1.035724    1.058741
  _IbpXiss_1 |   .9927199   .0161392    -0.45   0.653     .9615863    1.024861
```

```
         rts |   .4714732   .0285804   -12.40   0.000     .4186563    .5309533
  _IbpXrts_1 |   .7540543   .1137513    -1.87   0.061     .5610433    1.013465
      _Ibp_1 |   12.53281    14.3303     2.21   0.027       1.3328    117.8506
  _IbpXage_1 |   1.013542   .0148866     0.92   0.360     .9847811    1.043143
------------------------------------------------------------------------------
. estat gof,group(10)
Logistic model for death, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =      3132
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        16.84
                 Prob > chi2 =         0.0318
```

At this point I am not happy about the goodness-of-fit test. The objection raised earlier in class that age probably does not have a strictly linear effect may be coming back to bite us here. I hate to proceed with a full model that does not seem to fit well. I experimented with a couple of approaches to be more flexible with age. One was to create age groupings, the other was to fit an additional term that allowed curvature in age in the logit scale. The latter approach is more parsimonious and I liked the results more, although there was not a lot of difference (older patients are fit with greatly reduced odds of survival either way). The distribution of age already is skewed right in this data set, so instead of introducing a term for square of age I introduced a term for square root of age – the difference in logit fits being slight for older patients but substantial for very young ones. Now I fit the model above with this new age term introduced along with the interaction:

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= \alpha+\beta_1\text{ISS}+\beta_2\text{BP}+\beta_3\text{RTS}+\beta_4\text{Age}+\beta_5\sqrt{\text{Age}}+\beta_6(\text{BP}*\text{ISS})\\ &\quad+\beta_7(\text{BP}*\text{RTS})+\beta_8(\text{BP}*\text{Age})+\beta_9(\text{BP}*\sqrt{\text{Age}})\end{aligned}$$

```
. xi:logistic death iss i.bp rts age agesqrt i.bp*iss i.bp*rts i.bp*age i.bp*agesqrt
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
i.bp*iss          _IbpXiss_#          (coded as above)
i.bp*rts          _IbpXrts_#          (coded as above)
i.bp*age          _IbpXage_#          (coded as above)
i.bp*agesqrt      _IbpXagesq_#        (coded as above)
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(9)      =     944.90
                                                  Prob > chi2     =     0.0000
Log likelihood = -440.23492                       Pseudo R2       =     0.5176
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.073518   .0092108     8.27   0.000     1.055616    1.091723
         age |   1.110938   .0256408     4.56   0.000     1.061802    1.162347
     agesqrt |   .4900429   .1303889    -2.68   0.007     .2909038    .8255033
  _IbpXiss_1 |   .9906573   .0161779    -0.57   0.565     .9594513    1.022878
         rts |   .4779762   .0289156   -12.20   0.000     .4245337    .5381464
  _IbpXrts_1 |   .7347408   .1127582    -2.01   0.045     .5438801    .9925791
  _IbpXage_1 |    .872296   .0932179    -1.28   0.201     .7074574    1.075542
      _Ibp_1 |   .0760752   .3033795    -0.65   0.518     .0000307    188.6858
_IbpXagesq_1 |   6.322103   8.321638     1.40   0.161     .4791204     83.4216
------------------------------------------------------------------------------
. estat gof,group(10)
Logistic model for death, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =      3132
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        11.61
                 Prob > chi2 =         0.1695
```

We will look shortly at what is being fit in terms of age, but note how much larger is the p-value for the goodness-of-fit test. We should be ready to proceed with reducing the model.

I will consider a backward elimination with $\alpha = .10$. Following the hierarchy principle, the only candidates for exclusion at step 1 are the interactions. Each of the 5 main effects is

involved in 1 or more interactions, so we cannot eliminate any main effects initially. The least significant interaction is BP*ISS with a p-value of .565, so this effect is removed (.565 > .10).

1. Omit BP*ISS and fit the model

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{ISS} + \beta_2 \text{BP} + \beta_3 \text{RTS} + \beta_4 \text{Age} + \beta_5 \sqrt{\text{Age}} + \beta_7(\text{BP} * \text{RTS}) + \beta_8(\text{BP} * \text{Age}) + \beta_9(\text{BP} * \sqrt{\text{Age}})$$

```
. xi:logistic death iss i.bp rts age agesqrt i.bp*rts i.bp*age i.bp*agesqrt
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
i.bp*rts          _IbpXrts_#          (coded as above)
i.bp*age          _IbpXage_#          (coded as above)
i.bp*agesqrt      _IbpXagesq_#        (coded as above)
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(8)      =     944.57
                                                  Prob > chi2     =     0.0000
Log likelihood = -440.39915                       Pseudo R2       =     0.5175
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.070782    .007818     9.37   0.000     1.055568    1.086216
         age |   1.109481   .0254389     4.53   0.000     1.060726    1.160478
     agesqrt |   .4963908   .1314641    -2.64   0.008     .2953871    .8341726
         rts |   .4751856   .0283819   -12.46   0.000     .4226906       .5342
  _IbpXrts_1 |   .7451743     .11272    -1.94   0.052     .5539868    1.002343
  _IbpXage_1 |   .8700633   .0941304    -1.29   0.198     .7038191    1.075575
      _Ibp_1 |   .0480244   .1897495    -0.77   0.442     .0000208    110.8277
_IbpXagesq_1 |   6.604057   8.789526     1.42   0.156     .4863207    89.68069
------------------------------------------------------------------------------
```

At this step the candidates for exclusion are ISS, BP*Age, BP*$\sqrt{\text{Age}}$, and BP*RTS, of which BP*Age is least significant with a p-value of .198. This interaction is then omitted. Why is ISS a candidate for exclusion at this point?

2. Omit BP*Age and fit

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{ISS} + \beta_2 \text{BP} + \beta_3 \text{RTS} + \beta_4 \text{Age} + \beta_5 \sqrt{\text{Age}} + \beta_7(\text{BP} * \text{RTS}) + \beta_9(\text{BP} * \sqrt{\text{Age}})$$

```
. xi:logistic death iss i.bp rts age agesqrt i.bp*rts  i.bp*agesqrt
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
i.bp*rts          _IbpXrts_#          (coded as above)
i.bp*agesqrt      _IbpXagesq_#        (coded as above)
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(7)      =     942.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -441.34771                       Pseudo R2       =     0.5164
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.070154   .0078046     9.30   0.000     1.054966    1.085561
         rts |    .475377   .0283584   -12.47   0.000     .4229219    .5343383
         age |   1.100623   .0245862     4.29   0.000     1.053474    1.149881
      _Ibp_1 |   5.772836    6.82775     1.48   0.138     .5683829    58.63236
  _IbpXrts_1 |   .7505693   .1117116    -1.93   0.054     .5606625    1.004801
     agesqrt |   .5432324   .1411161    -2.35   0.019     .3264886    .9038644
_IbpXagesq_1 |   1.228323    .213722     1.18   0.237     .8733895    1.727497
------------------------------------------------------------------------------
```

At this step the candidates for exclusion are ISS, BP*$\sqrt{\text{Age}}$, and BP*RTS, of which BP*$\sqrt{\text{Age}}$ is least significant with a p-value of .237. This interaction is then omitted.

3. Omit BP*$\sqrt{\text{Age}}$ and fit

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{ISS} + \beta_2 \text{BP} + \beta_3 \text{RTS} + \beta_4 \text{Age} + \beta_5 \sqrt{\text{Age}} + \beta_7 (\text{BP} * \text{RTS})$$

```
. xi:logistic death iss i.bp rts age agesqrt i.bp*rts
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
i.bp*rts          _IbpXrts_#          (coded as above)
note: _Ibp_1 dropped due to collinearity
note: rts dropped due to collinearity
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(6)      =     941.24
                                                  Prob > chi2     =     0.0000
Log likelihood =  -442.0633                       Pseudo R2       =     0.5156
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.070038    .007773     9.32   0.000     1.054912    1.085382
         rts |   .4705327   .0280666   -12.64   0.000     .4186169    .5288869
         age |   1.102008   .0246182     4.35   0.000     1.054798     1.15133
     agesqrt |    .550232   .1433209    -2.29   0.022     .3302406     .916772
      _Ibp_1 |   13.35317   12.52076     2.76   0.006     2.125423    83.89257
  _IbpXrts_1 |   .7906284   .1090461    -1.70   0.089     .6033535    1.036032
------------------------------------------------------------------------------
```

The candidates for exclusion at this point are ISS, Age, $\sqrt{\text{Age}}$, and BP*RTS. The least significant effect is BP*RTS with a p-value of .089, which is less than our criterion of .10.

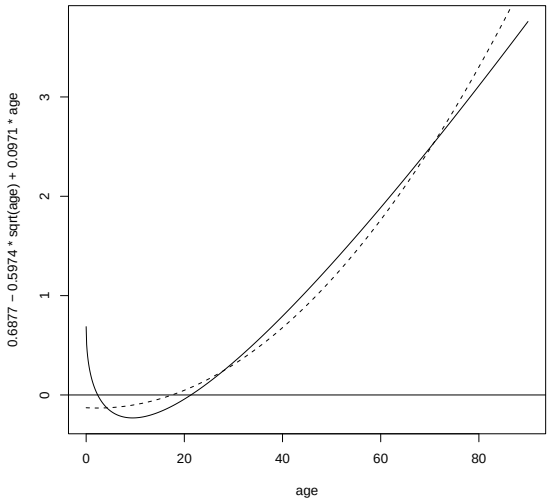
4. If we stick to the algorithm, we would stop and conclude that the important predictors are ISS, BP, RTS, AGE, and $\sqrt{\text{Age}}$ with an interaction between BP and RTS. All these steps can be automated with `sw`, as in the following output. I used logit here in order to see coefficients. Since I cannot combine `xi` and `sw`, I need to use `xi` alone to create indicator variables and then use those in `sw` for variable selection. `lockterm1` forces the first term in parentheses to stay in the model.

```
. xi i.bp i.bp*iss i.bp*rts i.bp*age i.bp*agesqrt
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
i.bp*iss          _IbpXiss_#          (coded as above)
i.bp*rts          _IbpXrts_#          (coded as above)
i.bp*age          _IbpXage_#          (coded as above)
i.bp*agesqrt      _IbpXagesq_#        (coded as above)

. sw  logit death (iss  _Ibp_1 rts  age agesqrt)   _IbpXiss_1  _IbpXrts_1 _IbpXage_1
>  _IbpXagesq_1, pr(.1) lockterm1
                    begin with full model
p = 0.5654 >= 0.1000  removing _IbpXiss_1
p = 0.1983 >= 0.1000  removing _IbpXage_1
p = 0.2372 >= 0.1000  removing _IbpXagesq_1
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(6)      =     941.24
                                                  Prob > chi2     =     0.0000
Log likelihood =  -442.0633                       Pseudo R2       =     0.5156
------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   .0676945   .0072642     9.32   0.000      .053457    .0819321
      _Ibp_1 |   2.591754   .9376617     2.76   0.006     .7539709    4.429537
         rts |  -.7538899   .0596486   -12.64   0.000    -.8707991   -.6369807
         age |   .0971337   .0223394     4.35   0.000     .0533494    .1409181
     agesqrt |  -.5974152   .2604735    -2.29   0.022    -1.107934   -.0868965
  _IbpXrts_1 |  -.2349272   .1379234    -1.70   0.089     -.505252    .0353977
       _cons |   .6877421   .8100227     0.85   0.396    -.8998732    2.275357
------------------------------------------------------------------------------

. estat gof,group(10)
Logistic model for death, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =      3132
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        14.46
                 Prob > chi2 =         0.0704
```

The fitted model is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = .688 + .068 \text{ ISS} + 2.59 \text{ BP} - .754 \text{ RTS} + .097 \text{ Age} - .597\sqrt{\text{Age}} - .235 \text{ BP} * \text{RTS}$$

The regression effect for ISS is easily interpreted as a risk factor for death (why?). The effect of age needs to be examined graphically since it is not simply linear. In the plot below the solid line is for the fitted model above, and the dotted line is what happens if we use AGE and AGE^2 instead. Can you see why I preferred using $\sqrt{\text{AGE}}$ to AGE^2? The fitted model shows increased risk of death for very young children, lowest risk for children and young adults, and substantially increased risk for older adults.



The effects of BP and RTS are more difficult to interpret because they interact. For example, for any fixed ISS and Age,

$$\begin{aligned}
\widehat{OR} &= \frac{\widehat{\text{odds}} \text{ of death for BP=1 (Penetrating)}}{\widehat{\text{odds}} \text{ of death for BP=0 (Blunt)}} \\
&= \frac{e^{.688+.068\text{ISS}+2.59(1)-.754\text{RTS}+.097\text{Age}-.597\sqrt{\text{Age}}-.235(1)\text{RTS}}}{e^{.688+.068\text{ISS}+2.59(0)-.754\text{RTS}+.097\text{Age}-.597\sqrt{\text{Age}}-.235(0)\text{RTS}}} \\
&= e^{2.59-.235\text{RTS}}
\end{aligned}$$

which decreases for increasing RTS. Looking at the ends of the RTS spectrum,

	RTS	$\widehat{OR}$
(no vitals)	0	13.35
(normal)	7.84	2.12

So, depending on ones RTS, the estimated odds of dying from a penetrating injury vary from 2 to 13 times the odds of dying from a blunt trauma, adjusting for ISS and Age. Before jumping on this large difference very hard, though, let's look at confidence intervals, which do overlap quite a bit here.

```
. lincom _b[_Ibp_1],or
 ( 1)  _Ibp_1 = 0
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   13.35317   12.52076     2.76   0.006     2.125423    83.89257
```

```
------------------------------------------------------------------------------
. lincom _b[_Ibp_1]+7.84*_b[_IbpXrts_1],or
 ( 1)  _Ibp_1 + 7.84 _IbpXrts_1 = 0
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.116841    .628437     2.53   0.012     1.183009    3.787814
------------------------------------------------------------------------------
```

Remarks

1. Some epidemiologists force *confounders* to be included in a logistic regression model regardless of their statistical significance.

2. The BP*RTS interaction was barely significant at the $\alpha = .10$ level. It might be interesting to see whether ones conclusions change when this effect is omitted.

```
. xi:logistic death iss i.bp rts age agesqrt
i.bp              _Ibp_0-1            (naturally coded; _Ibp_0 omitted)
Logistic regression                               Number of obs   =       3132
                                                  LR chi2(5)      =     937.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -443.73652                       Pseudo R2       =     0.5138
------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         iss |   1.069022   .0077489     9.21   0.000     1.053942    1.084318
      _Ibp_1 |   2.911404   .6711942     4.64   0.000     1.852963    4.574443
         rts |   .4474509   .0239457   -15.03   0.000     .4028958    .4969333
         age |   1.102117   .0249277     4.30   0.000     1.054327    1.152073
     agesqrt |   .5548027   .1463495    -2.23   0.026     .3308286    .9304094
------------------------------------------------------------------------------
. estat gof,group(10)
Logistic model for death, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =      3132
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        14.77
                 Prob > chi2 =         0.0637
```

We see that remaining effects are highly significant and there is no evidence of gross deficiencies.

3. The $\widehat{OR}$s for ISS and Age are similar for the two models. If a primary interest was estimating OR for ISS or Age, then it would not matter much which model we used. If BP is the interesting effect, the simpler model yields an $\widehat{OR}$ of 2.91, which is between the minimum and maximum $\widehat{OR}$ for the previous model.

4. The model without BP*RTS is simpler to interpret because it contains no interactions. However, most scientists are wary of omitting potentially important interactions, because of the potentially misleading conclusions that might be reached in models that ignore them. I would be inclined here to use the slightly more complex model with the BP*RTS interaction.

Case-Control Data

In epidemiological studies, the logistic model $\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is used to relate p, say the probability of disease or death, to the collection $x_1, x_2, \ldots, x_k$ of risk factors and confounders. With prospective or cross-sectional studies, we have noted that risk (i.e. probability of disease or death), relative risks, and ORs can be estimated using the logistic model – however most of our focus has been on ORs.

In practice, data are often sampled retrospectively using a case-control design. Although it is well known that risks and relative risks cannot be estimated using case-control data, ORs are

estimable and agree with ORs defined from a prospective study. In terms of logistic regression, the intercept cannot be estimated without bias using data from a case-control study, but regression coefficients for predictors and confounders, which correspond to adjusted ORs, are estimated appropriately. Thus, we can use standard methods to estimate regression effects and build regression models using case-control data.

Diverticular Disease Example
There is a description of this data set on the web page as a supplement to this lecture. The data set also is provided there. The data set has 64 rows with this content:

Variable Name	Contents
Age	Midpoint of age range (8 levels)
Sex	Values are f and m
Cancer	Colonic Cancer (1 is yes - case, 0 is no - control)
Lab	Case - Control label (not used)
Disease	Diverticular disease (values dd (yes) and ndd (no))
Count	No. individuals with this combination of variables

There are a lot of possible strategies for building a model to predict Cancer. I proceeded this way:

1. The primary interest is the potential association with diverticular disease (DD) and colonic cancer (CC). DD is considered an exposure variable.

2. Age and sex are viewed as confounders (potentially). Confounders are variables that are risk factors for the disease and associated with, but not a consequence of, presence or absence of the exposure variable.

   Because age and sex are likely to confound the relationship between the occurrence of DD and CC, most epidemiologists would argue that the effect of DD has to be assessed after adjusting for the effects of age and sex. As a result, many epidemiologists would include age and sex effects in a model, regardless of their statistical significance. Others might adopt a slightly different view and consider the effect of removing insignificant sex and age effect on adjusted ORs for DD. If removing insignificant effects has little impact on the estimate and precision of the adjusted OR for DD it does not matter much whether they are included or excluded. If the adjusted OR for DD changes dramatically upon removal, the insignificant effect would typically remain in the model.

3. We have the option of treating Age, using midpoint of the age range, as a categorical variable or on a continuous scale. If we consider Age as categorical, the odds of CC will be allowed to vary freely across age categories – that is, the odds is not required to vary smoothly with Age. If we choose this approach, interpretation of Age effects and interactions with Age will be cumbersome. However, almost every logistic model with Age, Sex, and DD effects fits well (using goodness of fit measures) when Age is categorical but fits poorly when Age is continuous. This implies that the log odds of CC does not change linearly with Age, but follows a more complex pattern. Consequently, I considered adding a quadratic term in Age, and this improved the fit dramatically.

4. I then posed a full model with the following effects: Sex, DD, Age, Age^2, Sex*Age, Sex*DD, Age*DD. I then proceeded with a Backward Elimination. I decided to force DD, Sex, and Age to be included in the model, regardless of their significance, but all other effects were candidates for exclusion. Note: Count must be defined as a frequency variable.

5. **Stata** is not going to let us use character variables directly in `logistic`, but that's no problem here since we need to create appropriate indicator variables and interactions anyway. `xi` is accommodating, though, so first we generate the indicators and then perform the sw procedure with the constraints listed above.

```
 . xi i.sex i.disease i.sex*age i.sex*i.disease i.disease*age
 i.sex             _Isex_1-2           (_Isex_1 for sex==f omitted)
 i.disease         _Idisease_1-2       (_Idisease_1 for disease==dd omitted)
 i.sex*age         _IsexXage_#         (coded as above)
 i.sex*i.disease   _IsexXdis_#_#       (coded as above)
 i.disease*age     _IdisXage_#         (coded as above)
. sw logit  cancer (age _Isex_2 _Idisease_2) agesq _IsexXage_2  _IsexXdis_2_2
        _IdisXage_2 [fweight=count],pr(.1) lockterm1
                      begin with full model
p = 0.4841 >= 0.1000  removing _IsexXdis_2_2
Logit estimates                                   Number of obs   =        193
                                                  LR chi2(6)      =      46.54
                                                  Prob > chi2     =     0.0000
Log likelihood = -104.72637                       Pseudo R2       =     0.1818
------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.6267873   .2185572    -2.87   0.004    -1.055151   -.1984232
     _Isex_2 |   4.009058   2.022184     1.98   0.047     .0456503    7.972465
 _Idisease_2 |  -4.604635   3.183622    -1.45   0.148    -10.84442    1.635149
       agesq |   .0053892   .0016092     3.35   0.001     .0022352    .0085432
 _IsexXage_2 |  -.0737373   .0318366    -2.32   0.021    -.1361359   -.0113386
 _IdisXage_2 |   .0806418   .0479469     1.68   0.093    -.0133323    .1746159
       _cons |   16.93369   7.510259     2.25   0.024     2.213853    31.65353
------------------------------------------------------------------------------
```

The `lockterm1` option forces `(age _Isex_2 _Idisease_2)` to stay in the model no matter what. Only the DD*Sex interaction term was removed in the backward elimination, so we have a model left with Age, Sex, DD, Age^2, Age*Sex, and Age*DD effects.

6. The goodness of fit test for the final model shows no problems.

```
Logistic model for cancer, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =       193
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        10.80
                Prob > chi2 =         0.2131
```

7. The parameter estimates table is given only for the final model when using `sw`.

8. A primary interest is the effect of disease on CC. `xi` produced `_Idisease_2` and `_IdisXage_2` where `_Idisease_2` is 1 for ndd, 0 for dd; and `_IdisXage_2` is 0 for dd and Age for ndd. We want to measure odds of cancer for ndd and dd. Using the same reasoning as previously (write the model, cancel common terms – the ones that are the same for dd and ndd),

$$\widehat{OR}\text{ (NDD vs. DD)} = e^{-4.604635+.0806418\text{Age}}$$

We could use this formula directly, but it is considerably easier to use `lincom` as before. I just computed the estimated OR for each of the ages in the data set, with the following results.

```
. lincom _b[ _Idisease_2]+44.5*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 44.5 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .3620128   .3995044    -0.92   0.357     .0416264    3.148324
------------------------------------------------------------------------------
. lincom _b[ _Idisease_2]+52*_b[  _IdisXage_2],or
```

```
 ( 1)  _Idisease_2 + 52 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .6628131   .5182929    -0.53   0.599     .1431482    3.068995
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+57*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 57 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    .991979   .5875997    -0.01   0.989     .3106651     3.16747
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+62*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 62 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.484615   .6725879     0.87   0.383     .6109234    3.607788
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+67*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 67 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.221904    .928302     1.91   0.056     .9797086    5.039108
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+72*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 72 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   3.325345   1.691708     2.36   0.018     1.226884    9.013008
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+77*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 77 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   4.976776   3.368095     2.37   0.018     1.320952    18.75035
------------------------------------------------------------------------------

. lincom _b[ _Idisease_2]+84.5*_b[  _IdisXage_2],or
 ( 1)  _Idisease_2 + 84.5 _IdisXage_2 = 0
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   9.112032   8.985077     2.24   0.025     1.319086    62.94442
------------------------------------------------------------------------------
```

The confidence intervals indicate that OR really doesn't differ significantly between DD and NDD for patients under 70, but for older patients DD appears actually to be protective. We should check to see if this is a real pattern in the data, or a fluke of the model we have fit. How would you do such an analysis? We also need to make sure it makes some sense to someone who knows the medicine.

I hope you see the value in including terms like Disease in the model, even though it is not actually significant in this case. We needed to assess the potential for this variable to affect CC through adjusted ORs, and we did find an interesting relationship (because age is so important).

As an aside, I will note that if you remove the effect for Sex (and its interaction with age), this has little effect on adjusted OR for DD. If age is completely ignored in the analysis the adjusted OR for DD is reduced dramatically, implying that age is clearly an important confounding variable in the relationship between DD and CC.

You can calculate any estimated adjusted OR using the above method. Remember, however, that this is a case-control study, so risks or odds should not be evaluated!

In-Lab Exercise

Return to the Framingham study data. Run the following code (make sure you understand what I am doing here):

```
graph bar chd [fw=freq],over(scl, ///
               relabel(1 "<190" 2 "190-219" 3 "220-249" 4 "250+")) ///
               over(agegroup, relabel(1 "30-49" 2 "50-62")) ///
               over(gender, relabel(1 "Female" 2 "Male")) ///
               ytitle("Proportion CHD") ///
               title("CHD vs. Gender, Age, and SCL")
bysort gender agegroup:tabulate chd scl [fw=frequency],chi2 exp col
```

Examine the output of the bar graphs and chi-squared tests.

1. What main effects appear to be present?
2. What interactions appear to be present?
3. Find a suitable model using logistic regression.
4. Summarize important odds ratios from your logistic regression model.
5. Give an overall summary of the analysis.

13 Introduction to Survival Analysis

In many biomedical studies, the outcome variable is a survival time, or more generally a time to an event. We will describe some of the standard tools for analyzing survival data.

Most studies of survival last a few years, and at completion many subjects may still be alive. For those individuals, the actual survival time is not known – all we know is how long they survived from their entry in the study. Similarly, certain individuals may drop out from the study or be lost to follow-up. Each of these cases is said to be *censored*, and the recorded time for such individuals is their time until the censoring event.

Example: HPA staining for breast cancer survival

We consider data from a retrospective study of 45 women who had surgery for breast cancer. Tumor cells, surgically removed from each woman, were classified according to the results of staining on a marker taken from the Roman snail, the *Helix pomatia agglutinin* (HPA). The marker binds to cancer cells associated with metastasis to nearby lymph nodes. Upon microscopic examination, the cancer cells stained with HPA are classified as positive, corresponding to a tumor with the potential for metastasis, or negative. It is of interest to determine the relationship of HPA staining and the survival of women with breast cancer.

The survival times in months time_i and staining results ($\text{group}_i = 0$ for negative and $\text{group}_i = 1$ for positive) for the 45 women are presented in the following table. Also included is a *censoring indicator* cens_i. Contrary to the normal definition of an indicator variable, the censoring indicator is zero if the observation is right-censored, and one if the observation is uncensored. So it's really a *non-censoring* indicator! A woman's survival time was right censored if the woman was alive at the end of the study or if the woman died of causes unrelated to breast cancer.

	time	group	cens		time	group	cens
1.	23	0	1	24.	40	1	1
2.	47	0	1	25.	41	1	1
3.	69	0	1	26.	48	1	1
4.	70	0	0	27.	50	1	1
5.	71	0	0	28.	59	1	1
6.	100	0	0	29.	61	1	1
7.	101	0	0	30.	68	1	1
8.	148	0	1	31.	71	1	1
9.	181	0	1	32.	76	1	0
10.	198	0	0	33.	105	1	0
11.	208	0	0	34.	107	1	0
12.	212	0	0	35.	109	1	0
13.	224	0	0	36.	113	1	1
14.	5	1	1	37.	116	1	0
15.	8	1	1	38.	118	1	1
16.	10	1	1	39.	143	1	1
17.	13	1	1	40.	154	1	0
18.	18	1	1	41.	162	1	0
19.	24	1	1	42.	188	1	0
20.	26	1	1	43.	212	1	0
21.	26	1	1	44.	217	1	0
22.	31	1	1	45.	225	1	0
23.	35	1	1				

This is the format the data should be in to work with it in **Stata**, but succinctly, the *sorted* survival times for the negative stained women are

$$23, 47, 69, 70^*, 71^*, 100^*, 101^*, 148, 181, 198^*, 208^*, 212^*, 224^*,$$

where $*$ denotes a right-censored observation. The survival times for the positive stained group are

$$5, 8, 10, 13, 18, 24, 26, 26, 31, 35, 40, 41, 48, 50, 59, 61, 68, 71, 76^*, 105^*,$$

$$107^*, 109^*, 113, 116^*, 118, 143, 154^*, 162^*, 188^*, 212^*, 217^*, 225^*.$$

In the breast cancer study, 8 individuals in the negative stained group, and 11 in the positive stained group are censored. Although it is common for studies to have *right-censored* cases, such as we have here, left-censoring and interval-censoring are found in other clinical studies.

Survival Curves

A first step in survival analysis is often to estimate the *survival curve*, or *survival time distribution*. Suppose we are considering a single (homogeneous) population. Let T be the survival time (from some reference point) for a randomly selected individual from the population. Where t is any arbitrary positive value, the survival time distribution is defined to be

$$\begin{aligned} S(t) &= Pr(T \geq t) \\ &= \text{probability randomly selected individual survives at least until time } t \\ &= \text{proportion of population that survives at least until time } t. \end{aligned}$$

The function might look like Figure 1.

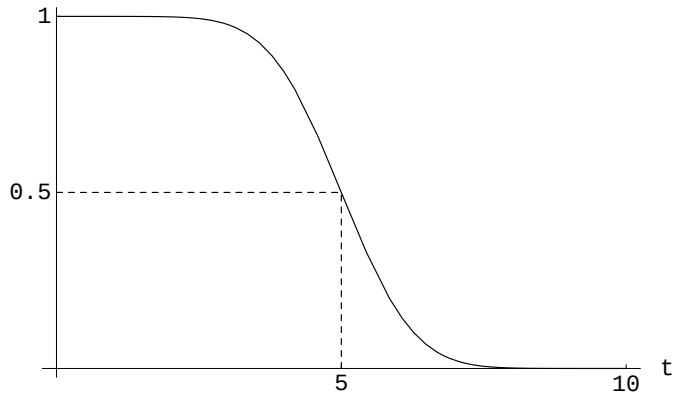


Figure 1: $S(t)$ versus t;median survival time for population is 5.

Estimating the Survival Curve

Case I: No censoring

If we have a random sample from the population, we use the *empirical survival function*:

$$\hat{S}(t) = \text{sample proportion that survive at least until time t}$$

to estimate $S(t)$. This is easy to compute and plot as a function of t.

Suppose we have a sample of 5 survival times (in days): 5, 8, 20, 30, and 33. $\hat{S}(t)$ has "jumps" of size 1/5 (i.e. 1 divided by the sample size) at each survival time; see Figure 2.

Case II: Right censoring

Recall the data on the survival of women with breast cancer whose cells were negatively stained with HPA:

$$23, 47, 69, 70^*, 71^*, 100^*, 101^*, 148, 181, 198^*, 208^*, 212^*, 224^*,$$

where the $*$ superscript identifies a right-censored observation.

The following algorithm describes the Kaplan-Meier (KM) method for estimating the survival curve (Kaplan-Meier product-limit estimate).

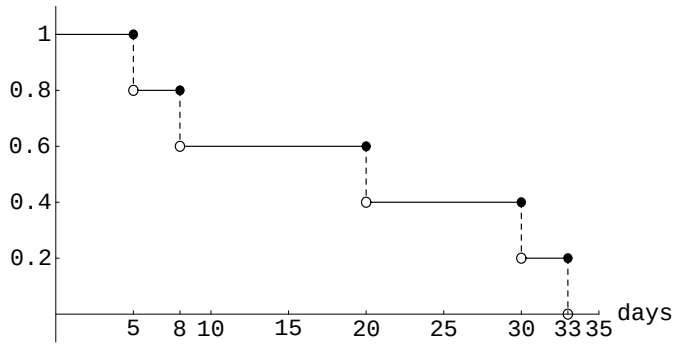


Figure 2: Empirical survival function $\hat{S}(t)$ for the data 5, 8, 20, 30, and 33.

1. Identify times for non-censored cases $0 = t_0 < t_1 < t_2 < \cdots < t_r$. That is, t_1 is the smallest non-censored survival time, t_2 is the second smallest, et cetera. For the example $r = 5$ and $t_0 = 0$, $t_1 = 23$, $t_2 = 47$, $t_3 = 69$, $t_4 = 148$, and $t_5 = 181$.

2. For the j^{th} interval, where $t_{j-1} \leq t < t_j$, evaluate

$$\begin{aligned} n_j &= \text{number at risk (of dying) at beginning of interval,} \\ d_j &= \text{number of deaths in interval,} \\ \frac{n_j - d_j}{n_j} &= \text{estimated probability of surviving past } t_{j-1}\text{, given survival to } t_{j-1} \\ &= \hat{P}(T \geq t_{j-1} | T \geq t_{j-2}). \end{aligned}$$

3. For $t_{j-1} \leq t < t_j$,

$$\begin{aligned} \hat{S}(t) &= \hat{P}(T \geq t) \\ &= \hat{P}(T \geq t_{j-1} | T \geq t_{j-2}) \times \\ &\quad \hat{P}(T \geq t_{j-2} | T \geq t_{j-3}) \times \cdots \times \\ &\quad \hat{P}(T \geq t_1 | T \geq t_0) \\ &= \frac{n_j - d_j}{n_j} \times \frac{n_{j-1} - d_{j-1}}{n_{j-1}} \times \cdots \times \frac{n_1 - d_1}{n_1}. \end{aligned}$$

Remark: Censored observations are taken into account by being treated as cases at risk at the beginning of the interval in which they fail.

To illustrate the calculation for our data, consider the table:

j	Interval	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$\hat{S}(t)$
1	$0 \leq t < 23$	13	0	$\frac{13-0}{13} = 1$	1.0
2	$23 \leq t < 47$	13	1	$\frac{13-1}{13} = \frac{12}{13} \doteq 0.923$	$1.0 \times 0.923 = 0.923$
3	$47 \leq t < 69$	12	1	$\frac{12-1}{12} = \frac{11}{12} \doteq 0.917$	$0.923 \times 0.917 = 0.846$
4	$69 \leq t < 148$	11	1	$\frac{10}{11} \doteq 0.909$	$0.846 \times 0.909 = 0.769$
5	$148 \leq t < 181$	6	1	$\frac{5}{6} \doteq 0.833$	$0.769 \times 0.833 = 0.641$
6	$181 \leq t$	5	1	$\frac{4}{5} = 0.8$	$0.641 \times 0.8 = 0.513$

To obtain the KM estimate in **Stata** we must declare the data we are working with to be survival data. **Stata** then uses the survival time variable and the censoring variable together in analyses. For the breast cancer data we first read in the variables using something like `infile time group cens using c:/breast.txt`. We declare the data to be survival data using `stset time, failure(cens)`. Stata creates several internal variables when we do this. Note that the option `,failure(cens)` makes the variable `cens` into an indicator of known death ("failure"). Finally we obtain the KM survival curve estimates across the two groups with the command `sts graph, by(group)`. In Figure 3 we have a picture of $\hat{S}(t)$ from the negatively stained group as well as the estimate from the positively stained group. Note that the negatively stained group tends to live longer, as we would expect.

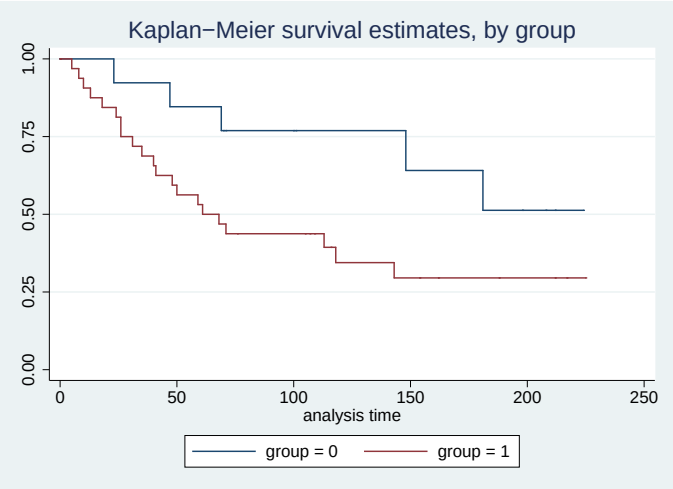


Figure 3: KM survival curves for positively and negatively stained groups.

The estimated quartiles for survival across the two groups are obtained by `stsum, by(group)`. Annotated output follows; for example, we see that the median survival in the positive stained group is estimated to be 61 months.

```
. stsum,by(group)
         failure _d:  cens
   analysis time _t:  time
         |                incidence       no. of    |------ Survival time -----|
group    | time at risk     rate        subjects        25%       50%       75%
---------+--------------------------------------------------------------------
       0 |         1652   .0030266            13        148         .         .
       1 |         2679   .0078387            32         26        61         .
---------+--------------------------------------------------------------------
   total |         4331   .0060032            45         40       113         .
```

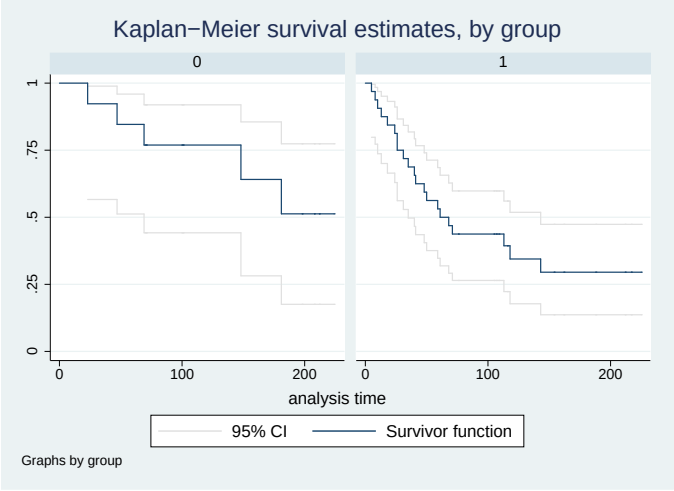
Some remarks:

- The estimated survival curve "drops to zero" only if the last case is not censored.
- The KM curve allows us to estimate percentiles of the survival distribution, with a primary interest being the median survival time (50^{th} percentile). In the example above (group 0), the 90^{th} percentile is approximately 47 months (i.e. we estimate that 90% of the population will survive at least 47 months). The median cannot be estimated here – all we can say is that we estimate the median to be at least 181 months.
- The KM estimate is the usual empirical estimate if no cases are censored.
- Statistical methods are available to
  - Estimate the mean survival time. This sounds good, but survival time distributions tend to be highly skewed right, so usually we are much more interested in the median.

```
. stci,by(group) rmean
         failure _d:  cens
   analysis time _t:  time
             |    no. of  restricted
group        |  subjects        mean      Std. Err.    [95% Conf. Interval]
-------------+------------------------------------------------------------
           0 |        13    167.7436(*)   20.80779      126.961    208.526
           1 |        32    104.0477(*)    15.6278      73.4177    134.678
-------------+------------------------------------------------------------
       total |        45    121.9075(*)    13.3793      95.6846     148.13
(*) largest observed analysis time is censored, mean is underestimated.

. stci,by(group) emean
         failure _d:  cens
   analysis time _t:  time
             |    no. of    extended
group        |  subjects        mean
-------------+----------------------
           0 |        13    339.7513
           1 |        32    158.5235
-------------+----------------------
       total |        45    196.3361
```

  - Get a C.I. for the survival curve. You need to ask for pointwise Greenwood confidence bands.

```
. sts graph,by(group) gwood
         failure _d:  cens
   analysis time _t:  time
```



– Compare survival curves across groups – you can think of this as the censored data analogue of (non-parametric) ANOVA. There are a lot of available tests, though most common probably is the log-rank test (Stata's default). A few of them follow.

```
. sts test group
         failure _d:  cens
   analysis time _t:  time

Log-rank test for equality of survivor functions
      |   Events         Events
group |  observed       expected
------+-------------------------
0     |         5           9.57
1     |        21          16.43
------+-------------------------
Total |        26          26.00
            chi2(1) =       3.51
            Pr>chi2 =     0.0608
. sts test group,wilc
         failure _d:  cens
   analysis time _t:  time

Wilcoxon (Breslow) test for equality of survivor functions
      |   Events         Events        Sum of
group |  observed       expected        ranks
------+----------------------------------------
0     |         5           9.57         -159
1     |        21          16.43          159
------+----------------------------------------
Total |        26          26.00            0
            chi2(1) =       4.18
            Pr>chi2 =     0.0409
. sts test group,cox
         failure _d:  cens
   analysis time _t:  time

Cox regression-based test for equality of survival curves
      |   Events         Events       Relative
group |  observed       expected        hazard
------+----------------------------------------
0     |         5           9.57        0.5633
1     |        21          16.43        1.3966
------+----------------------------------------
Total |        26          26.00        1.0000
         LR chi2(1) =       3.87
            Pr>chi2 =     0.0491
. sts test group,peto
```

```
        failure _d:  cens
  analysis time _t:  time

Peto-Peto test for equality of survivor functions
      |   Events         Events        Sum of
group |  observed       expected        ranks
------+-------------------------------------
0     |         5           9.57    -3.509069
1     |        21          16.43     3.509069
------+-------------------------------------
Total |        26          26.00            0
           chi2(1) =       4.12
           Pr>chi2 =     0.0425

. sts test group,tware
        failure _d:  cens
  analysis time _t:  time

Tarone-Ware test for equality of survivor functions
      |   Events         Events        Sum of
group |  observed       expected        ranks
------+-------------------------------------
0     |         5           9.57   -26.921999
1     |        21          16.43    26.921999
------+-------------------------------------
Total |        26          26.00            0
           chi2(1) =       4.05
           Pr>chi2 =     0.0441
```

These tests do not necessarily all agree with each other – some emphasize different parts of the distribution than others, and different principles for approximating distributions are employed. In this case there is little difference, except the minor differences are right at the 0.05 significance level.

The Cox Proportional Hazards Model

The risk of failing at time t is defined to be the probability of an individual dying in the "next instant" (e.g. in a time frame of length Δ) given this individual has survived at least until time t:

$$P(t \leq T < t + \Delta | t \leq T).$$

We define the *hazard function* $h(t)$ such that for small enough Δ,

$$P(t \leq T < t + \Delta | t \leq T) = h(t)\Delta.$$

The hazard function is proportional to the instantaneous "risk of failing" at any time t, given that an individual has lived at least to time t.

Now consider two individuals, 1 and 2, each with their own hazard functions $h_1(t)$ and $h_2(t)$. If we assume that one individual's instantaneous rate of failing is a constant multiple of the other's, i.e. $h_2(t) = ah_1(t)$ for some constant a, then these two individuals have *proportional hazard functions*. Figure 4 shows an example of this phenomenon where the hazard ratio is $1/2$.

Proportional hazards may or may not be a reasonable assumption to make. For example, consider two people, roughly the same age and demographic except that at the age of 20, person 2 takes up smoking while person 1 does not. You will hopefully agree with me that initially, the smoker and the non-smoker will most likely have *identical* hazards. As the years roll by, and smoking takes its toll, we would think that the smoker's instantaneous rate of failing, which is proportional to the probability of dying in the next minute, say, will increase relative to the hazard for the non-smoker. In this example proportional hazards probably is an unreasonable assumption.

The proportional hazards *model* generalizes the above concept for n individuals, each with their own covariate value x_i or set of p covariate values $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. In the case where the

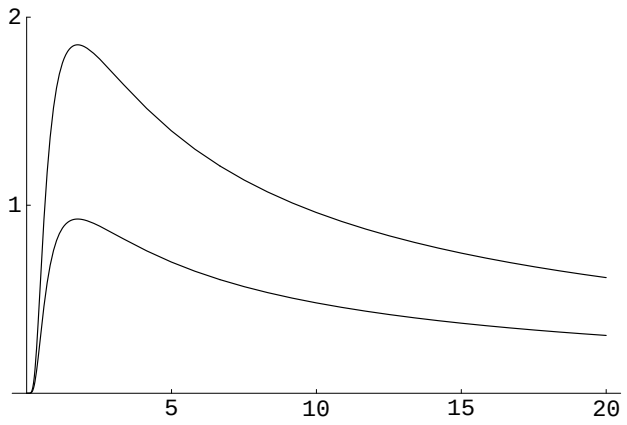


Figure 4: An example of proportional hazard functions; here the constant of proportionality is 0.5.

n individuals only have one covariate, the model stipulates for individuals i and j, with hazard functions $h_i(t)$ and $h_j(t)$ respectively, that

$$h_i(t)e^{-\beta x_i} = h_j(t)e^{-\beta x_j}.$$

Note that this implies

$$\frac{h_i(t)}{h_j(t)} = \frac{e^{\beta x_i}}{e^{\beta x_j}} = e^{\beta(x_i - x_j)}.$$

Here, $e^{\beta(x_i-x_j)}$ is the relative risk of instantaneous failure at *any time* t for individuals i and j. That is the power of the proportional hazards assumption: the relative risk of dying for two individuals is a simple function of the model parameters and holds for all t, independent of the value of t. If individual i has covariate value $x+1$ and individual j has covariate value x, i.e. their covariate values only differ by 1 unit on the covariate measurement scale, then

$$\frac{h_i(t)}{h_j(t)} = \frac{e^{\beta(x+1)}}{e^{\beta x}} = e^{\beta}.$$

Thus, e^{β} is the relative risk of failing in the next instant when we increase the covariate by one unit. Note that if x_i is a simple zero/one variable denoting which group individual i falls into, then e^{β} is the relative risk of failing in the next instant for the group denoted by $x_i = 1$ versus $x_i = 0$.

The Cox PH model is fit as follows (everything had to be `stset` prior to this command). We do not have to specify the dependent variable as in other regression routines because it is defined (survival time) with `stset`.

```
. stcox group
         failure _d:  cens
   analysis time _t:  time
Iteration 0:   log likelihood = -86.983777
Iteration 1:   log likelihood = -85.087844
Iteration 2:   log likelihood = -85.048003
Iteration 3:   log likelihood = -85.047944
Refining estimates:
Iteration 0:   log likelihood = -85.047944
Cox regression -- Breslow method for ties
No. of subjects =           45                     Number of obs   =        45
No. of failures =           26
Time at risk    =         4331
```

```
                                                  LR chi2(1)      =      3.87
Log likelihood  =   -85.047944                    Prob > chi2     =    0.0491
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |   2.479398   1.241987     1.81   0.070     .9288808    6.618086
------------------------------------------------------------------------------
*************
************* Note:  stcox group,nohr  reports coefficients, not hazard ratios
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |   .9080157   .5009228     1.81   0.070    -.0737749    1.889806
------------------------------------------------------------------------------
```

We have an estimate of $\hat{\beta} = 0.908$ and the estimated relative risk is $e^{\hat{\beta}} = e^{0.908} \doteq 2.5$. That is, those with positive staining are estimated to have a risk of dying in the next instant about 2.5 times as great as those with negative staining. Note that the p-value for $H_0 : \beta = 0$ is small but not significant at the 5% level. There is definitely *some* indication that staining affects survival, with positive staining decreasing survival. A 95% C.I. for the risk may be obtained by exponentiating the endpoints for the C.I. for β. Here, we estimate the relative risk of expiring (for positive compared to negative staining) to be within $(e^{-0.073}, e^{1.89}) = (0.93, 6.62)$ with 95% confidence.

Remark: The hazard function for individual i can be defined to be a scale multiple $e^{x_i\beta}$ of a *baseline* hazard function denoted $h_0(t)$. The model may be recast as $h_i(t) = h(t|x_i) = e^{x_i\beta}h_0(t)$. This baseline hazard function $h_0(t)$ and β together thus completely determine the model. The baseline hazard $h_0(t)$ may be estimated from the data as well as survival curves, median and mean survival, et cetera, for any covariate value x. These sorts of inferences are quite easy to get out of Stata but a bit beyond what is comfortable to cover in this class.

Checking the Proportional Hazards Assumption

The assumption matters because estimates and inferences based on them may be completely invalid if hazards are not proportional, and we have no easy way to assess that from the regression output. There are standard diagnostics that often work pretty well but are not at all easy to derive without some mathematics.

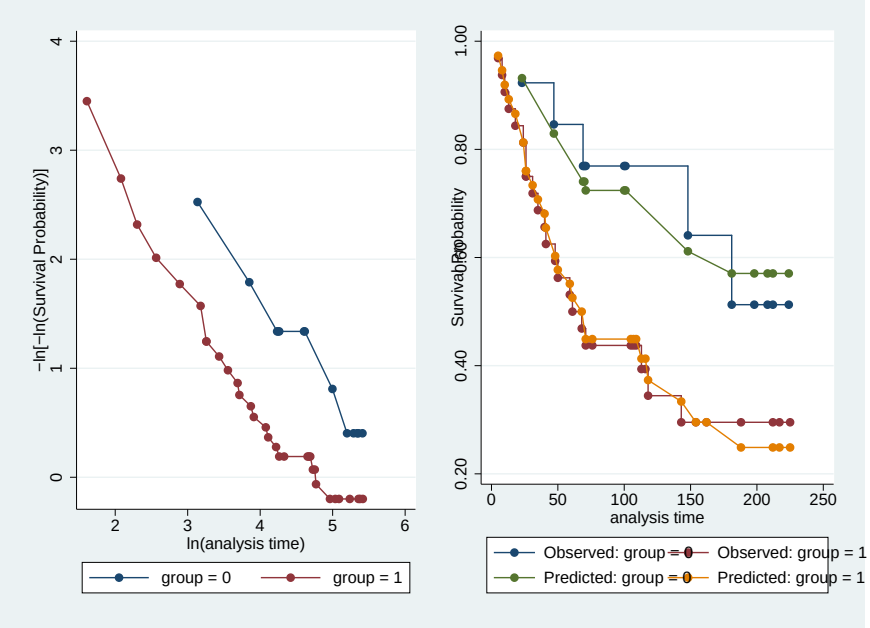
What you would like to do is estimate the hazard functions from the data, and plot them on the same scale to assess proportionality. That is very similar to estimating a population frequency distribution, and is a fairly hard problem that takes quite a bit of data (Stata will do it, but that doesn't mean it works well). What works better and is much easier to estimate (effectively from Kaplan-Meier) is the integrated or cumulative hazard $H(t) = \int_0^t h(s)ds$. If hazard functions are proportional then so are cumulative hazard functions. A very fortunate mathematical relationship is that the survival curve is related to the cumulative hazard as $H(t) = -\log S(t)$ so $\log H(t) = \log[-\log S(t)]$. If $h_0(t)$ is the baseline hazard function, then the proportional hazards assumptions says the cumulative hazard function for individual i is $e^{x_i\beta}H_0(t)$, and $\log[-logS_i(t)] = x_i\beta + \log[-logS_0(t)]$. What this says is that if x is a group indicator as above then plots of $\log[-\log S(t)]$ for groups should be parallel (and separated by the β's for groups). Stata does (almost) this in its `stphplot` command.

Another approach is to compare the completely nonparametric Kaplan-Meier curve to the curve predicted under proportional hazards. That is the `stcoxkm` command. Stata's description of these commands from the help system is as follows:

```
Description

    stphplot plots -ln{-ln(survival)} curves for each category of a nominal or
    ordinal covariate versus ln(analysis time). These are often referred to as
    "log-log" plots.  Optionally, these estimates can be adjusted for
    covariates.  The proportional-hazards assumption is not violated when the
    curves are parallel.

    stcoxkm plots Kaplan-Meier observed survival curves and compares them to the
    Cox predicted curves for the same variable.  The closer the observed values
    are to the predicted, the less likely it is that the proportional-hazards
    assumption has been violated.  Do not run stcox prior to this command;
    stcoxkm will execute stcox itself to estimate the model and obtain predicted
    values.
```

Doing this with the breast cancer data using these commands yields the following graph:

```
. stphplot,by(group) name(loglog)
. stcoxkm,by(group) name(phkm)
. graph combine loglog phkm
```



This all looks great. We do not have any evidence that the proportional hazards assumption is not reasonable.

A final example

We examine a data set consisting of the time spent running on a treadmill for 14 people aged 15 and older. Each subject's gender and age were recorded. It is of interest to the experimenter how age and gender affect ones endurance.

When fitting the PH model with gender and age as main effects,

$$h(t|\text{age, gender}) = e^{age\times\beta_1+gender\times\beta_2}h_0(t),$$

we are going to let `xi` determine the gender indicator (it will set the first group alphabetically, females, to 0 and males to 1). The baseline group (i.e. those with covariates age = 0 and gender indicator = 0, and thus a hazard function of $e^{0\beta_1+0\beta_2} = e^0 h_0(t) = h_0(t)$) consists of females of age zero, which is not interpretable in this context. Observations were censored due to a subject having to leave the treadmill for reasons other than being tired. The data follow:

```
      gender   age   minutes   cens   weight
 1.     male    34        16      1      215
 2.     male    15        35      0      135
 3.   female    22        55      0      145
 4.   female    18        95      1       97
 5.     male    18        55      0      225
 6.   female    32        55      1      185
 7.   female    37        25      1      155
 8.   female    67        15      1      142
 9.   female    55        22      1      132
10.     male    55        13      1      183
11.     male    62        13      1      168
12.   female    33        57      0      132
13.   female    17        52      0      112
14.     male    24        54      1      175
```

We need to start by declaring a survival data set:

```
. stset minutes,failure(cens)
     failure event:  cens != 0 & cens < .
obs. time interval:  (0, minutes]
 exit on or before:  failure
--------------------------------------------------------------------------------
       14  total obs.
        0  exclusions
--------------------------------------------------------------------------------
       14  obs. remaining, representing
        9  failures in single record/single failure data
      562  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =         0
                                 last observed exit t =        95
```

The fit of the model with only gender $h(t|\text{gender}) = e^{\text{gender indicator}\times\beta_1}h_0(t)$:

```
. xi: stcox i.gender
i.gender          _Igender_1-2        (_Igender_1 for gender==female omitted)
         failure _d:  cens
   analysis time _t:  minutes
Iteration 0:   log likelihood = -18.061924
Iteration 1:   log likelihood = -17.622711
Iteration 2:   log likelihood = -17.620368
Refining estimates:
Iteration 0:   log likelihood = -17.620368
Cox regression -- Breslow method for ties
No. of subjects =           14                     Number of obs   =        14
No. of failures =            9
Time at risk    =          562
                                                   LR chi2(1)      =      0.88
Log likelihood  =   -17.620368                     Prob > chi2     =    0.3474
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Igender_2 |   1.971276   1.411726     0.95   0.343     .4843518    8.022949
------------------------------------------------------------------------------
```

The test for a gender effect yields a p-value of 0.343. We would accept at any reasonable significance level that there is not a gender effect. The estimated hazard ratio is

$$h(t|\text{gender} = \text{male})/h(t|\text{gender} = \text{female}) = 1.97$$

for all t. That is, the probability of a randomly picked man failing (stepping off the treadmill) in the next second is estimated be twice the probability of a randomly picked female. The confidence interval for the hazard ratio is from 0.48 to 8.02, which includes 1 since the effect was not significant.

Let's look at the model fit with only age $h(t|\text{age}) = e^{\text{age}\times\beta_1}h_0(t)$:

```
. stcox age
         failure _d:  cens
```

```
   analysis time _t:  minutes
Iteration 0:   log likelihood = -18.061924
Iteration 1:   log likelihood = -11.184791
Iteration 2:   log likelihood = -11.184559
Refining estimates:
Iteration 0:   log likelihood = -11.184559
Cox regression -- Breslow method for ties
No. of subjects =           14                     Number of obs   =        14
No. of failures =            9
Time at risk    =          562
                                                   LR chi2(1)      =     13.75
Log likelihood  =   -11.184559                     Prob > chi2     =    0.0002
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.118133    .043125     2.90   0.004     1.036725    1.205934
------------------------------------------------------------------------------
```

A year from now, a randomly selected individual will be 1.118 times as likely to step off the treadmill after 15 minutes (or any amount of time) than now. In ten years it will be $1.118^{10} = 3.05$ times as likely. When we fit the model with both of these predictors $h(t|\text{age, gender}) = e^{\text{age}\times\beta_1+\text{gender indicator}\times\beta_2}h_0(t)$ we see that estimated regression effects, and therefore model interpretation, change somewhat:

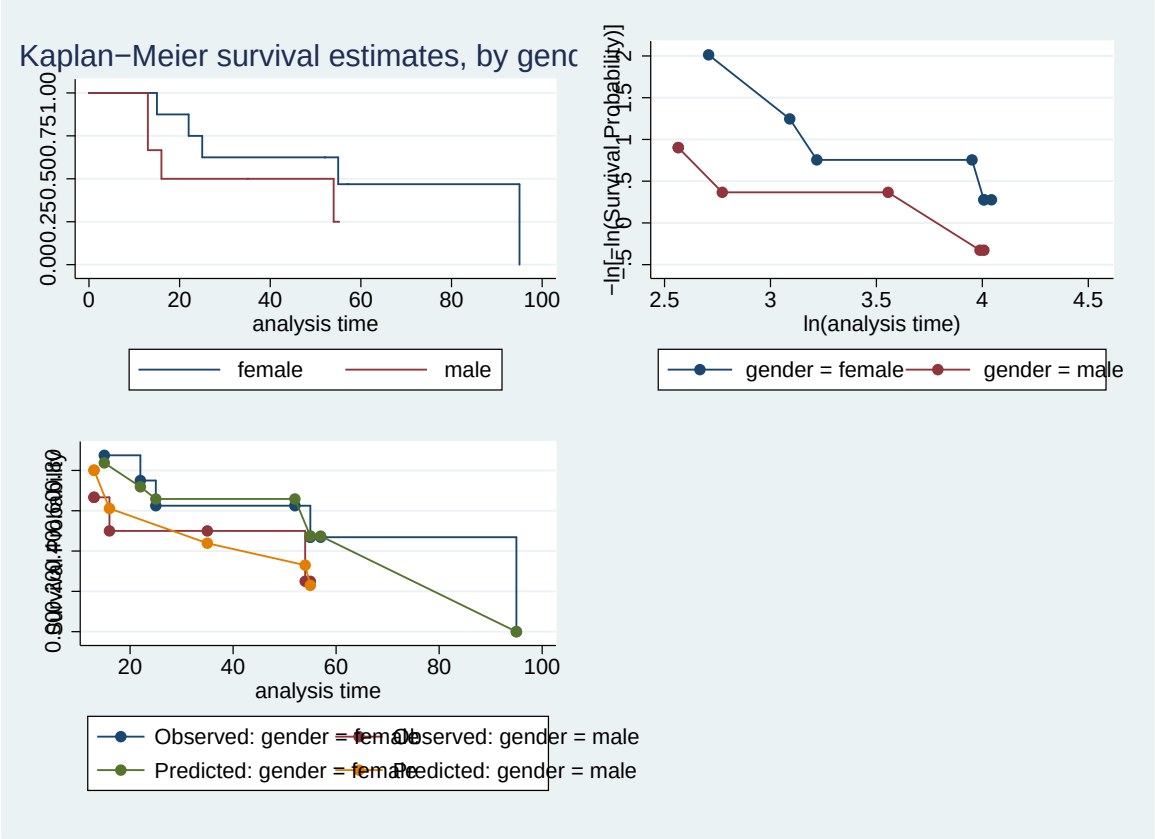
```
. xi: stcox i.gender age
i.gender          _Igender_1-2        (_Igender_1 for gender==female omitted)
         failure _d:  cens
   analysis time _t:  minutes
Iteration 0:   log likelihood = -18.061924
Iteration 1:   log likelihood = -8.3270231
Iteration 2:   log likelihood = -7.2765366
Iteration 3:   log likelihood = -7.1238759
Iteration 4:   log likelihood = -7.1166297
Iteration 5:   log likelihood = -7.1166049
Refining estimates:
Iteration 0:   log likelihood = -7.1166049
Cox regression -- Breslow method for ties
No. of subjects =           14                     Number of obs   =        14
No. of failures =            9
Time at risk    =          562
                                                   LR chi2(2)      =     21.89
Log likelihood  =   -7.1166049                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Igender_2 |   34.87809   55.05714     2.25   0.024     1.580812     769.529
         age |   1.244367   .1064681     2.56   0.011     1.052251    1.471558
------------------------------------------------------------------------------
```

At a given age, a random male running alongside a random female is about 35 times as likely to step off the treadmill at any time. A woman 20 years older than another woman is about $1.244^{20} \doteq 79$ times as likely to step off compared to the younger woman. Note that in the presence of age, gender is now significant, although by itself gender is not a significant factor. In this case age is said to be a *suppressor* variable.

In the model fit that included an interaction between age and gender, the interaction term was not significant. Weight is not significant in the presence of gender and age (or by itself).

We should check on the proportional hazards assumption.

```
. sts graph,by(gender) name(km)
. stphplot,by(gender) name(loglog)
. stcoxkm,by(gender) name(phkm)
. graph combine km loglog phkm
```



There is not a lot of data here, so we have to expect deviations from ideal patterns. The log-log plot only needs to be parallel, not linear, and this is plausibly parallel. The PH model is not disagreeing with the Kaplan-Meier fit. The survival curves by gender are consistent with the tests we ran. There is no clear reason to question the PH assumptions.

14 Poisson Regression

In class we will cover Chapter 12 (Analysis of Rates with Poisson Regression) from Steve Selvin's text *Practical Biostatistical Methods* (1995, Wadsworth). There are many more applications of Poisson regression than covered there, but this chapter has a treatment quite relevant to you.

The appendix in Selvin discusses the Poisson distribution mostly as an approximation to the binomial distribution when n is large and p is small. A broader more detailed perspective can be found at Wikipedia (`http://en.wikipedia.org/wiki/Poisson_distribution`). One entry there states that

> The word **law** is sometimes used as a synonym of probability distribution, and *convergence in law* means convergence in distribution. Accordingly, the Poisson distribution is sometimes called the **law of small numbers** because it is the probability distribution of the number of occurrences of an event that happens rarely but has very many opportunities to happen. *The Law of Small Numbers* is a book by Ladislaus Bortkiewicz about the Poisson distribution, published in 1898. Some historians of mathematics have argued that the Poisson distribution should have been called the Bortkiewicz distribution.

Wikipedia provides several examples of the Poisson as well:

> The Poisson distribution arises in connection with Poisson processes. It applies to various phenomena of discrete nature (that is, those that may happen 0, 1, 2, 3, ... times during a given period of time or in a given area) whenever the probability of the phenomenon happening is constant in time or space. Examples of events that can be modelled as Poisson distributions include:
>
> - The number of cars that pass through a certain point on a road during a given period of time.
> - The number of spelling mistakes a secretary makes while typing a single page.
> - The number of phone calls at a call center per minute.
> - The number of times a web server is accessed per minute. For instance, the number of edits per hour recorded on Wikipedia's Recent Changes page follows an approximately Poisson distribution.
> - The number of roadkill found per unit length of road.
> - The number of mutations in a given stretch of DNA after a certain amount of radiation.
> - The number of unstable nuclei that decayed within a given period of time in a piece of radioactive substance. The radioactivity of the substance will weaken with time, so the total time interval used in the model should be significantly less than the mean lifetime of the substance.
> - The number of pine trees per unit area of mixed forest.
> - The number of stars in a given volume of space.
> - The number of soldiers killed by horse-kicks each year in each corps in the Prussian cavalry. This example was made famous by a book of Ladislaus Josephovich Bortkiewicz (1868-1931).
> - The distribution of visual receptor cells in the retina of the human eye.

- The number of V2 rocket attacks per area in England, according to the fictionalized account in Thomas Pynchon's Gravity's Rainbow.
- The number of light bulbs that burn out in a certain amount of time.

For our purposes two more interesting examples probably are number of deaths in a subpopulation in a certain amount of time and number of cases of a disease in a subpopulation in a fixed period of time. What is key in these and the other examples is that the Poisson distribution describes random counts. Selvin deals with rates (proportions) which probably are more common but require some special consideration.

Mathematical Background

Let the random variable Y have a Poisson distribution with parameter λ. This means that

$$P(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!};\ k = 0, 1, 2, 3, \ldots$$

and both the mean $E(Y)$ and variance $Var(Y)$ of Y are λ. As with the binomial distribution where we found the logit function and logistic regression to be more useful than considering the binomial distribution directly, we fit regression models to $\log Y$ (natural log), so we get models of the form

$$\log E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

where there are good theoretical reasons for logit with binomial and log for the Poisson. These are special cases of a large class of such models called *generalized linear models* (handled with the `glm` command in Stata in general, more easily handled with the `poisson` command for this case).

Poisson regression fits linear models to log(counts of number of events) – remember, fitting linear models means looking for group differences, interactions, adjusting for covariates and confounders,etc.. The events we are looking at probably will be deaths or diagnosis of disease. In the applications you are likely to encounter most often, we actually want to fit linear models to rates, where rates are usually of the form

$$r = \frac{\text{count of events}}{\text{population size}}, \quad \text{or} \quad r = \frac{\text{count of events}}{\text{total exposure}}.$$

Either way, we probably want to model r more than we want to model counts directly. The way we do this is to fit a linear model to log of r

$$\begin{aligned}
\log r &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p && (1)\\
\log\left(\frac{\text{count of events}}{\text{population size}}\right) &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p && (2)\\
\log(\text{count of events}) - \log(\text{population size}) &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p && (3)\\
\log(\text{count of events}) &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p + \log(\text{population size}) && (4)
\end{aligned}$$

Equation (4) shows that we model $\log r$ by writing a linear model for log (count of events), which is regular Poisson regression, except that we have a special new variable with a known coefficient of 1, i.e. log (population size). Such a variable is called an *offset*. The general form would be the same if we had a rate using exposure instead of population size.

In order to interpret coefficients, consider the simple case of two groups (say F and M) where $x_1 = \begin{cases} 0 & \text{Group} = \text{F} \\ 1 & \text{Group} = \text{M} \end{cases}$, the usual group indicator variable, and we fit a simple model to compare

groups, $\log r = \hat{\beta}_0 + \hat{\beta}_1 x_1$. Then F is the reference group and $\hat{\beta}_1$ is the difference between groups M and F in the log scale, just as we usually have in linear models, i.e. $\log r_F = \hat{\beta}_0$ and $\log r_M = \hat{\beta}_0 + \hat{\beta}_1$, so $\log r_M - \log r_F = \hat{\beta}_1$ and $\frac{r_M}{r_F} = e^{\log r_M - \log r_F} = e^{\hat{\beta}_1}$. Similar to the way we obtained estimated Odds Ratios in logistic regression, we obtain estimated *incidence-rate ratios* by exponentiating estimated regression coefficients. Factors with multiple levels and continuous predictor variables are handled similarly to the way we have handled then in least squares regression and in logistic regression.

Stata Implementation

A portion of the help viewer in Stata for the Poisson command shows

```
Syntax
        poisson depvar [indepvars] [if] [in] [weight] [, options]

    options                       description
    -------------------------------------------------------------------------
    Model
      noconstant                  suppress constant term
      exposure(varname_e)         include ln(varname_e) in model with
                                    coefficient constrained to 1
      offset(varname_o)           include varname_o in model with coefficient
    Reporting
      level(#)                    set confidence level; default is level(95)
      irr                         report incidence-rate ratios

    irr reports estimated coefficients transformed to incidence-rate ratios,
        that is, exp(b) rather than b.  Standard errors and confidence
        intervals are similarly transformed.  This option affects how results
        are displayed, not how they are estimated or stored.  irr may be
        specified at estimation or when replaying previously estimated
        results.

    offset(varname) specifies that varname be included in the model with the
        coefficient constrained to be 1.

    exposure(varname) specifies a variable that reflects the amount of
        exposure over which the depvar events were observed for each
        observation; ln(varname) with coefficient constrained to be 1 is
        entered into the log-link function.
```

From a practical perspective, what difference does it make if you declare a variable an offset or an exposure? Read carefully – you need to already have taken the log for an offset, but Stata will go ahead and take the log of an exposure variable. You can use either form, just be careful you have the variable in the correct form.

Example

Consider the very simple data set

```
      group   deaths   popsize
 1.       F       10     10000
 2.       M       15      8000
```

and compare mortality rates for the two groups. Define `lpopsize` from the command `gene lpopsize = log(popsize)`. We want to fit using both the `offset` and `exposure` forms of the command to see they agree.

```
. xi:poisson deaths i.group, exposure(popsize)
i.group           _Igroup_1-2         (_Igroup_1 for group==F omitted)
Iteration 0:   log likelihood =   -4.35708
Iteration 1:   log likelihood =   -4.35708
Poisson regression                                Number of obs   =          2
                                                  LR chi2(1)      =       2.43
                                                  Prob > chi2     =     0.1188
Log likelihood =   -4.35708                       Pseudo R2       =     0.2183
------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Igroup_2 |   .6286087   .4082483     1.54   0.124    -.1715433    1.428761
       _cons |  -6.907755   .3162278   -21.84   0.000     -7.52755    -6.28796
     popsize | (exposure)
------------------------------------------------------------------------------

. xi:poisson deaths i.group, offset(lpopsize)
i.group           _Igroup_1-2         (_Igroup_1 for group==F omitted)
Iteration 0:   log likelihood =   -4.35708
Iteration 1:   log likelihood =   -4.35708
Poisson regression                                Number of obs   =          2
                                                  LR chi2(1)      =       2.43
                                                  Prob > chi2     =     0.1188
Log likelihood =   -4.35708                       Pseudo R2       =     0.2183
------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Igroup_2 |   .6286087   .4082483     1.54   0.124    -.1715433    1.428761
       _cons |  -6.907755   .3162278   -21.84   0.000     -7.52755    -6.28796
    lpopsize |   (offset)
------------------------------------------------------------------------------

. poisson,irr
Poisson regression                                Number of obs   =          2
                                                  LR chi2(1)      =       2.43
                                                  Prob > chi2     =     0.1188
Log likelihood =   -4.35708                       Pseudo R2       =     0.2183
------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Igroup_2 |      1.875   .7654656     1.54   0.124     .8423638    4.173523
    lpopsize |   (offset)
------------------------------------------------------------------------------
```

But look what happens if you mix offset and exposure:

```
. xi:poisson deaths i.group, offset(popsize)
Poisson regression                                Number of obs   =          2
                                                  Wald chi2(1)    =  2.401e+07
Log likelihood =   -4.35708                       Prob > chi2     =     0.0000
------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Igroup_2 |   2000.405   .4082483  4899.97   0.000     1999.605    2001.206
       _cons |  -9997.697   .3162278 -3.2e+04   0.000    -9998.317   -9997.078
     popsize |   (offset)
------------------------------------------------------------------------------
```

We will do a number of the examples in Selvin's chapter during class.

14 Power and Sample Size

Consider the simple problem from last semester where we sample from a normal population with a known σ and want to test a hypothesis of the form $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$. The rule for the test is to reject H_0 if $|Z| > z_{critical}$ where $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$. $z_{critical}$ is chosen to satisfy $P(|Z| > z_{critical}|H_0 \text{ is true}) = \alpha$, where α is the probability of a Type I Error, or the significance level. A Type I Error is rejecting H_0 when H_0 is actually true, i.e. claiming something important is going on when actually nothing important is going on. Usually we take $\alpha = .05$ which forces $z_{critical}$=1.96.

The significance level is the chance of rejecting H_0 when we should not do so. If H_A is true then of course we *should* reject H_0. The probability that we correctly reject H_0 when H_A is true is defined as the **Power** of the test. Power is a lot more complicated than α, though.

The extra complication comes from two sources. First, H_A is not *simple* like H_0 is, so if H_A is true there are many possible values of μ other than μ_0. If the actual μ is a long way from μ_0 then it should be fairly easy to tell that H_0 is not true and the power should be high. If the actual μ is close to μ_0, though, it should be pretty hard to tell that we are not sampling from the situation described by H_0, and the power may be low. Second, the sample size has a lot to do with the power. If n is large then we have a lot of information and it should be easy to tell if H_A is true, but if n is small it may be very difficult to tell that H_0 is not true. By contrast, the test is structured so that α does not depend upon the sample size – it is always the fixed number we choose (usually .05).

We can actually tell exactly how the Z-statistic above behaves if we specify exactly which of the values of μ is true. $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is standard normal if we used the right μ, and we are using $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ (which is standard normal if we used the right μ, i.e. if H_0 is true). If H_A is true write

$$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{\overline{X}-\mu+(\mu-\mu_0)}{\sigma/\sqrt{n}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} + \frac{\mu-\mu_0}{\sigma/\sqrt{n}}$$

which tells us that Z is a normal random variable with standard deviation 1 and mean $\sqrt{n}(\frac{\mu-\mu_0}{\sigma})$ if H_A is true.

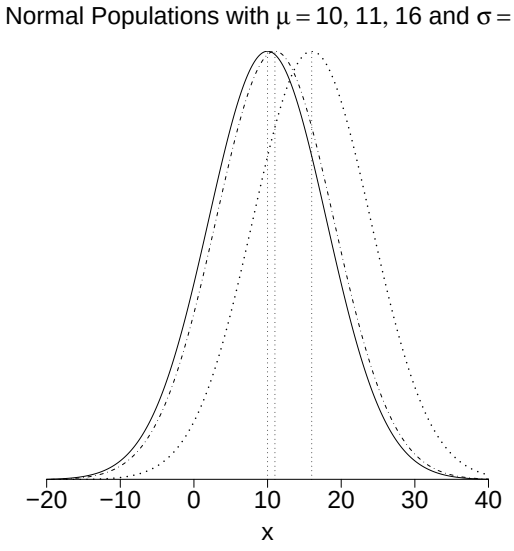


Figure 1: Three normal populations with $\sigma = 8$.

Example: Let $\sigma = 8$ and test $H_0 : \mu = 10$ vs. $H_A : \mu \neq 10$. There are infinitely many possible values of μ under H_A, but for purposes of illustration let us consider only $\mu = 11$ and $\mu = 16$. Figure 1 graphs all three populations. Clearly it is going to be fairly hard to tell whether we sampled from the population with $\mu = 11$ or the population with $\mu = 10$, since they are so little different, but it should be quite a bit easier to tell if we sampled from the population with $\mu = 16$ or the population with $\mu = 10$ (although that is not trivial either).

Consider a fixed sample size of $n = 16$. $P(|Z| > 1.96|\mu = 10) = \alpha = .05$. The power when $\mu = 11$ is $P(|Z| > 1.96|\mu = 11)$ while the power when $\mu = 16$ is $P(|Z| > 1.96|\mu = 16)$. Figure 2 (a) shows the distribution of the Z-statistic for sampling from each of these populations, (b) shows the calculation of α, (c) shows the power (.08) for $\mu = 11$ and (d) shows the power (.85) for $\mu = 16$. As expected, we have a good chance with this sample size of telling $\mu = 16$ from $\mu = 10$ but a slim chance of telling $\mu = 11$ from $\mu = 10$.

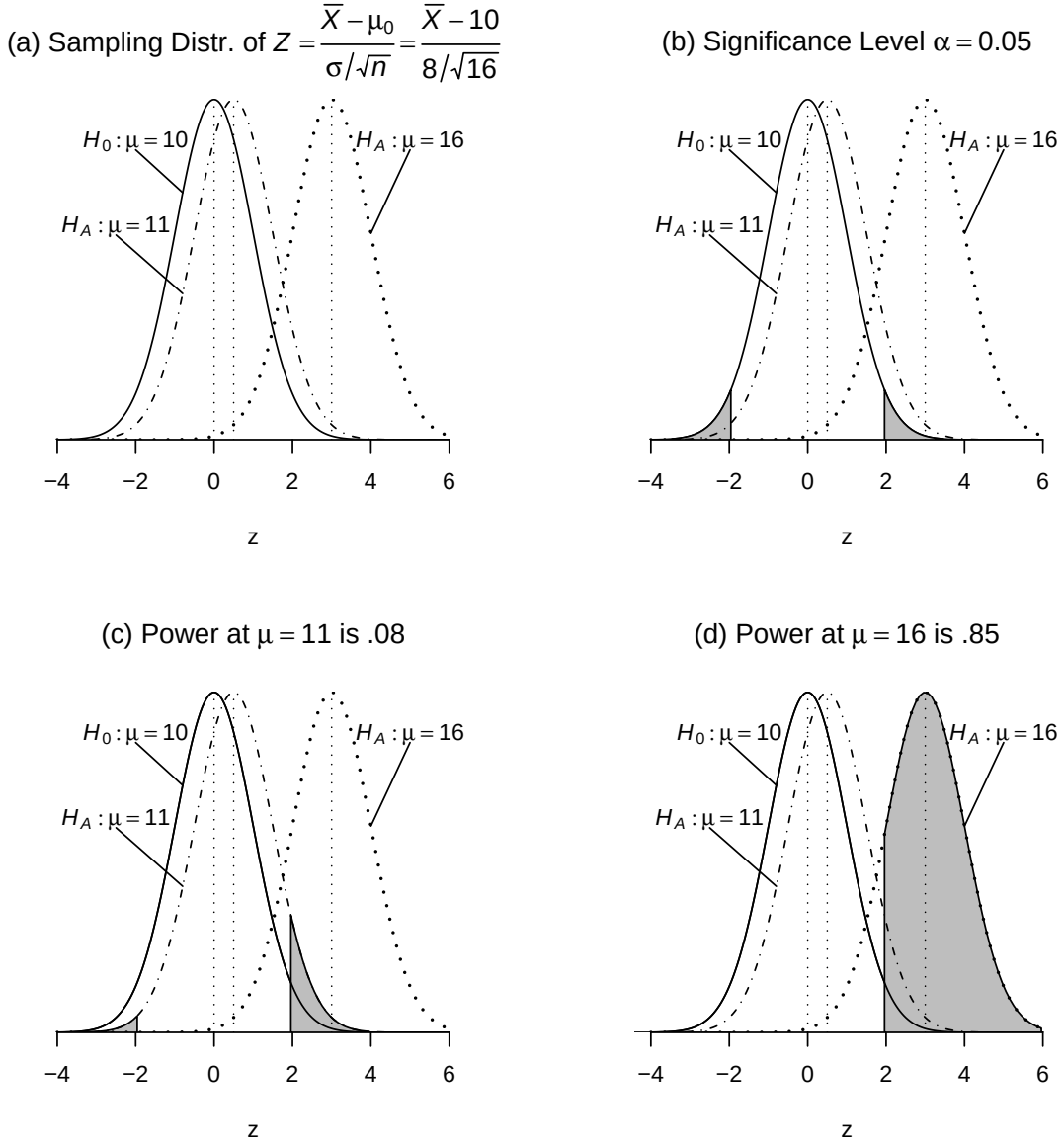


Figure 2: Power for a test of $H_0 : \mu = 10$ vs. $H_A : \mu \neq 10$ for a random sample of 16 from a normal population with $\sigma = 8$.

For a given alternative μ, the power also increases with n. Figure 3 demonstrates this behavior for the example using $\mu = 16$ and sample sizes of 5, 15, and 25.

Stata will do calculations like these for you. Follow the menu path

```
Summaries, tables, & tests
     → Classical tests of hypotheses
          → Sample size and power determination
```

and fill in the boxes. Here we are doing "One-sample comparison of mean to hypothesized value" so check that, give hypothesized value of 10, Std. deviation one of 8, and Postulated mean of 16. Next click on the Options box, ask to compute power, specify significance level of .05 and a two sided test with sample size 15. **Stata** returns the following:

```
. sampsi 10 16, alpha(.05) n1(15) sd1(8) onesample
Estimated power for one-sample comparison of mean
  to hypothesized value
Test Ho: m =     10, where m is the mean in the population
Assumptions:
         alpha =   0.0500  (two-sided)
 alternative m =       16
            sd =        8
 sample size n =       15
Estimated power:
         power =   0.8276
```

This is exactly what you find on Figure 3 for n = 15.

Much more common is the inverse of this problem, where we specify the power and ask for the sample size. Conceptually this is not different from what we have been doing, but there are some guidelines. Generally we specify some reasonable alternative, make a good guess based on published literature or preliminary data of the standard deviation, specify a two-tailed procedure at $\alpha = .05$, and target power of .8. The goal is to find n to yield that power. If we do that for the earlier problem with $\mu = 11$ and $\sigma = 8$ where $\mu_0 = 10$, we get

```
. sampsi 10 11, alpha(.05) power(.80) sd1(8) onesample
Estimated sample size for one-sample comparison of mean
  to hypothesized value
Test Ho: m =     10, where m is the mean in the population
Assumptions:
         alpha =   0.0500  (two-sided)
         power =   0.8000
 alternative m =       11
            sd =        8
Estimated required sample size:
             n =      503
```

which says we need over 500 observations to have an 80% chance of telling populations as close as the two closest in Figure 1 apart.

Stata does power analysis like the preceding on two-sample tests for means and both one- and two-sample tests for proportions. There is not much conceptual difference, but you will get a chance to experiment a bit in lab. There are specialized packages for more complex/complete calculation (I usually use PASS in NCSS because I have it), but there is free software as well. Check out UCLA's nice little calculator (`Power Calculator` at `http://calculators.stat.ucla.edu/`) for a wider variety of procedures.

The paper by Cohen on the web site is standard reading on this topic. His approach can be very useful.

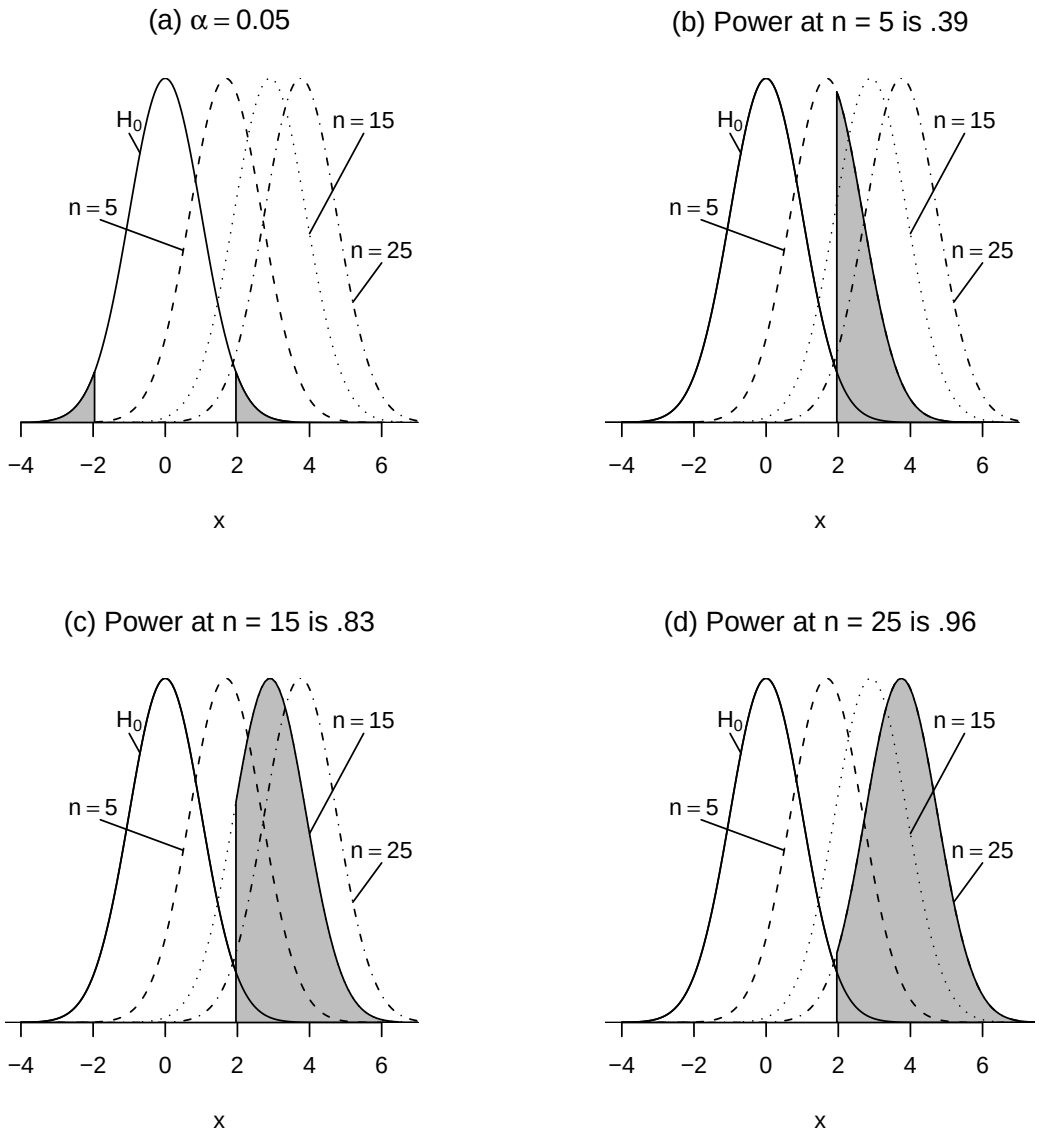


Figure 3: Power for a test of $H_0 : \mu = 10$ vs. $H_A : \mu = 16$ for random samples of 5, 15, 25 from a normal population with $\sigma = 8$.