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

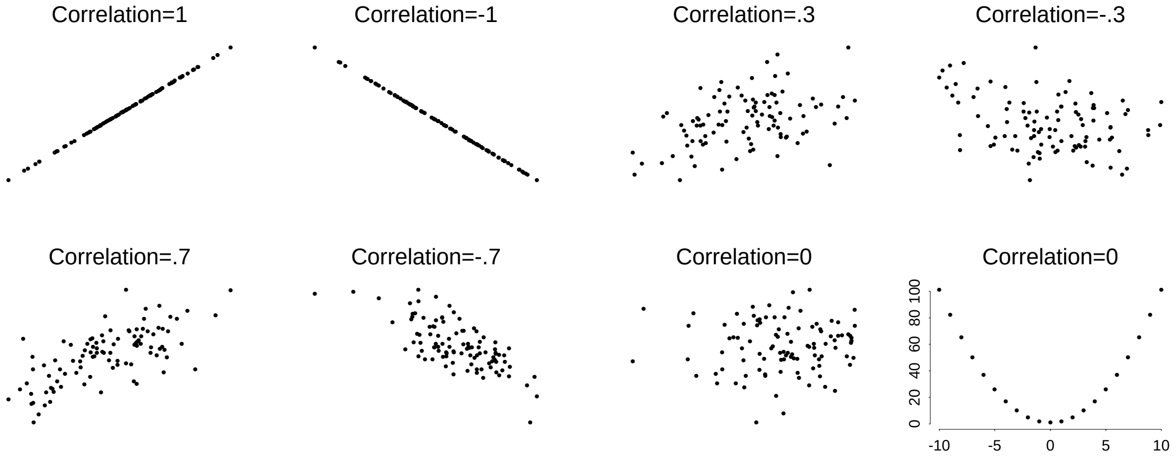
$$S_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

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

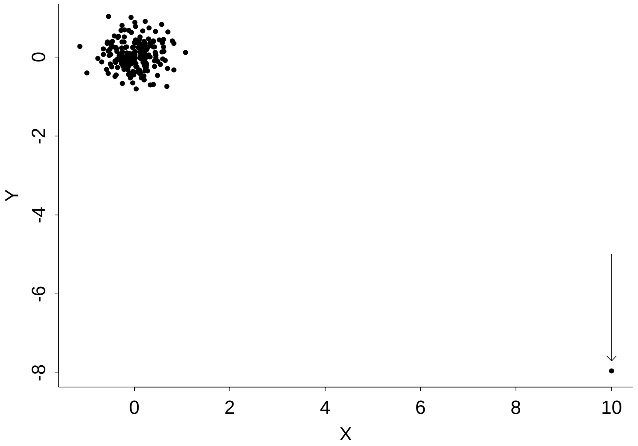
$$t_s = r\sqrt{\frac{n-2}{1-r^2}},$$

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:

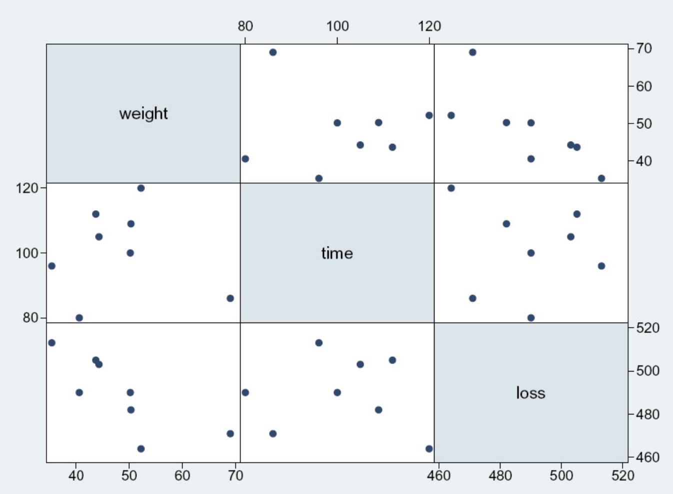
> **Direct:** Follow the menu path in Stata of `Data -> Data Editor` to bring up the spreadsheet-like Stata Editor. You have to enter data before you can name columns.
>
> **Flat Text File:** Put the data into a flat text file (just the numbers), named say `C:\biostat\bloodloss.txt`. In the Stata Command 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).

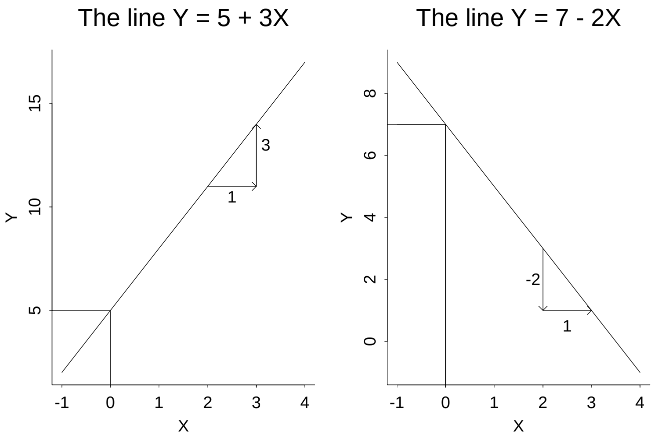
3. Another measure of association available in Stata is Kendall's τ (tau, not given here).

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.



The line Y = 5 + 3X

The line Y = 7 - 2X

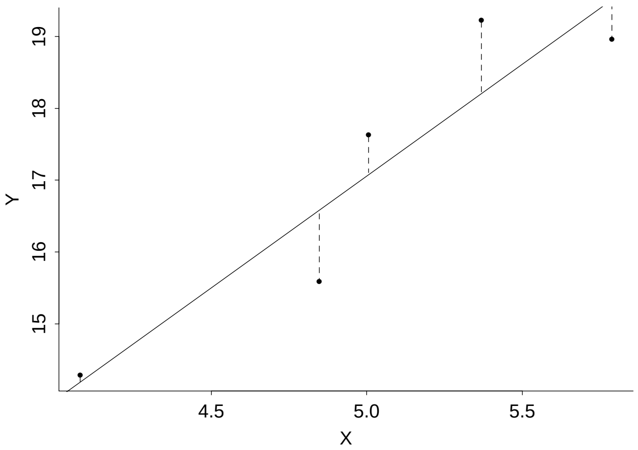
Least Squares

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

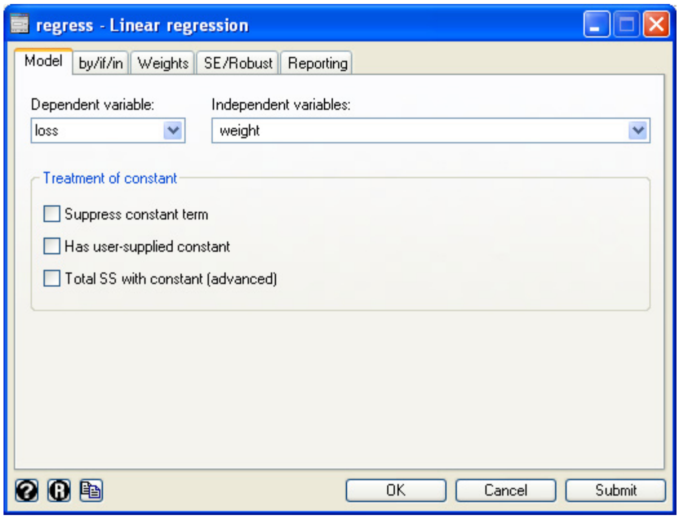
$$b_1 = \frac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} = r \frac{S_Y}{S_X}.$$

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



Dependent variable: loss

Independent variables: weight

Treatment of constant

□ Suppress constant term

□ Has user-supplied constant

□ Total SS with constant (advanced)

```
. 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
------------------------------------------------------------------------------
```

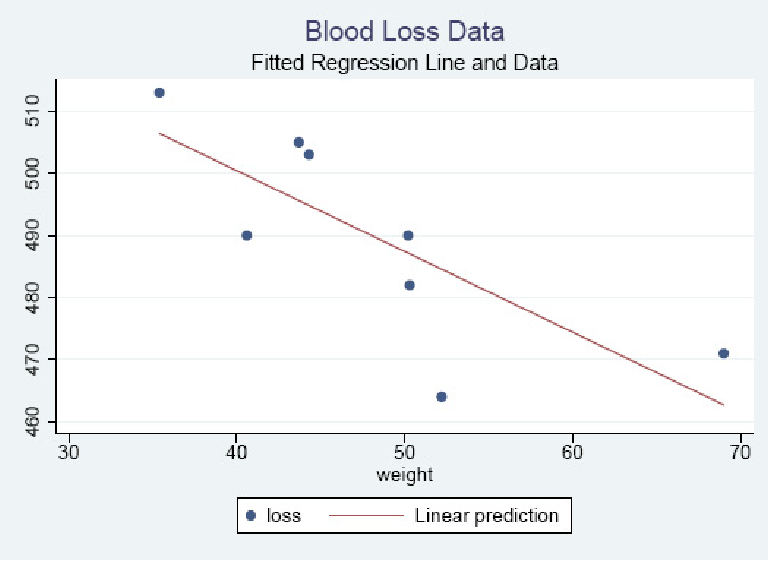
For the **thyroid operation data** with Y = Blood loss in ml and X = Weight in kg, the LS line is $\hat{Y} = 552.44 - 1.30X$, or Predicted Blood Loss = 552.44 − 1.30 Weight. For an 86kg individual, the Predicted Blood Loss = $552.44 - 1.30 * 86 = 440.64ml$. The LS regression coefficients for this model 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

$$\sum_{i=1}^{n}\{Y_i-(\beta_0+\beta_1X_i)\}^2$$

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

$$\text{Residual Sums of Squares} = \sum_{i=1}^{n}\{Y_i-(b_0+b_1X_i)\}^2.$$

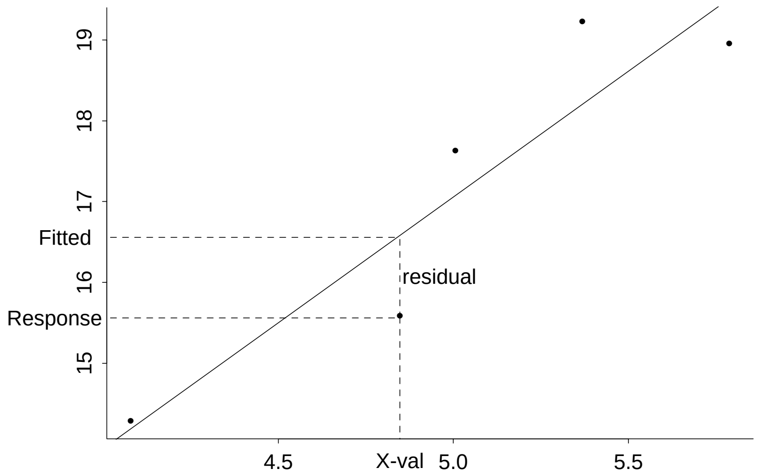
Several bits of notation are needed. Let

$$\hat{Y}_i = b_0 + b_1X_i$$

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

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

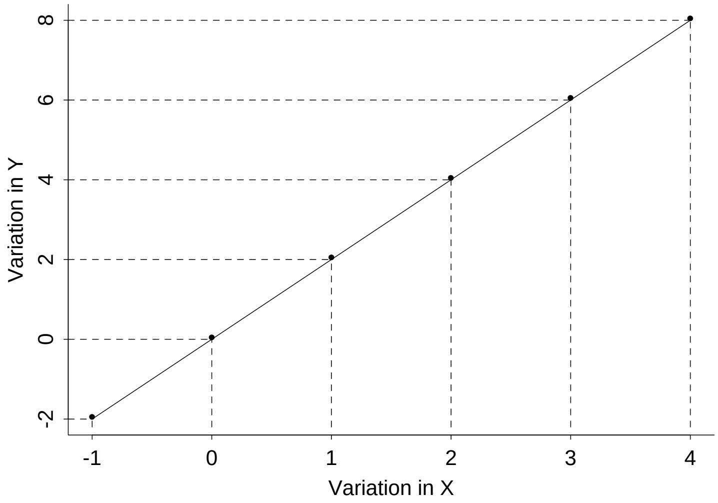
To understand the interpretation of R^2, at least in two extreme cases, note that

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

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

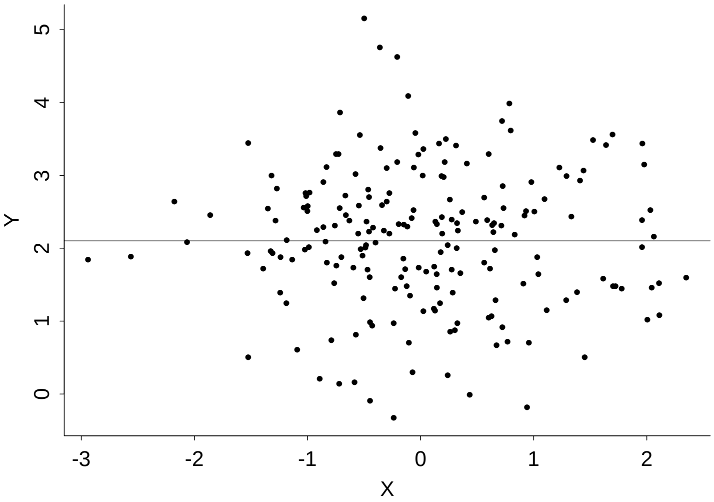
$\Leftrightarrow$ all the variability in Y is explained by the linear relationship with X (which has variation)

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



Furthermore,

$$\begin{aligned} \text{Reg SS} = 0 \quad &\Leftrightarrow \quad \text{Total SS} = \text{Res SS} \\ &\Leftrightarrow \quad b_1 = 0 \\ &\Leftrightarrow \quad \text{LS line is } \hat{Y} = \bar{Y} \\ &\Leftrightarrow \quad \text{none of the variability in } Y \text{ is explained by a linear relationship} \\ &\Leftrightarrow \quad R^2 = 0. \end{aligned}$$



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 $R^2 = .5967 = (-.77247)^2 = r^2$.