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).

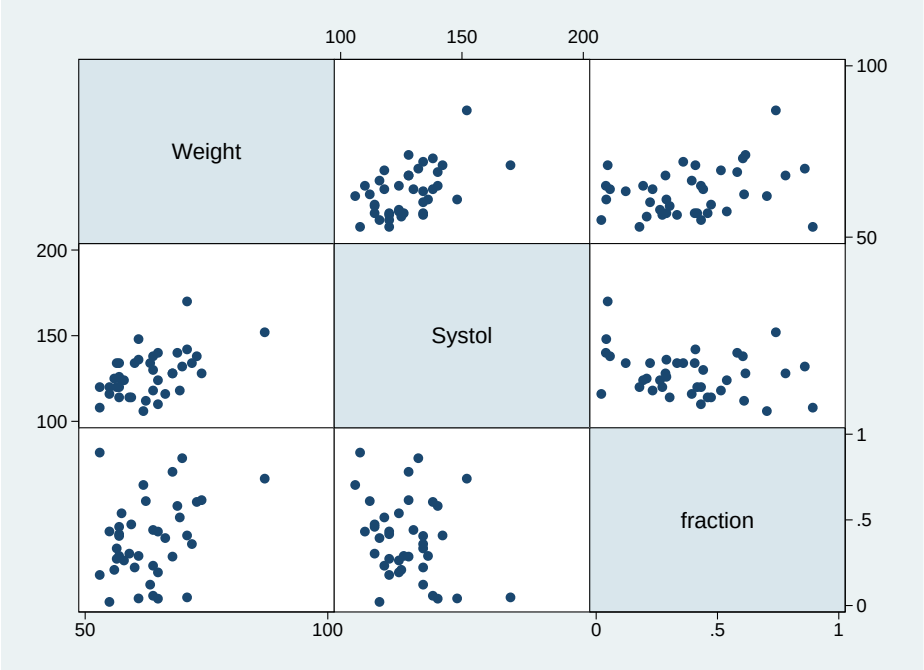
A plot of systolic blood pressure against fraction suggests a weak linear relationship (from `graph matrix weight systol fraction`). Nonetheless, consider fitting the regression model

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

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.

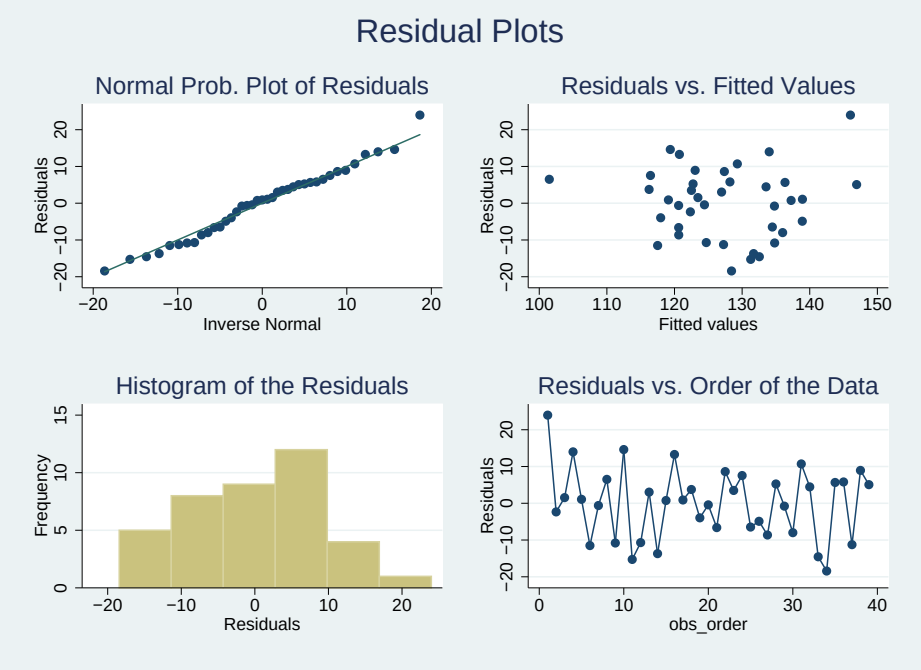
The implications of this analysis are enormous! Essentially, the correlation between a predictor and a response says very little about the importance of the predictor in a regression model with one or more additional predictors. This conclusion also holds in situations where the correlation is high, 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.

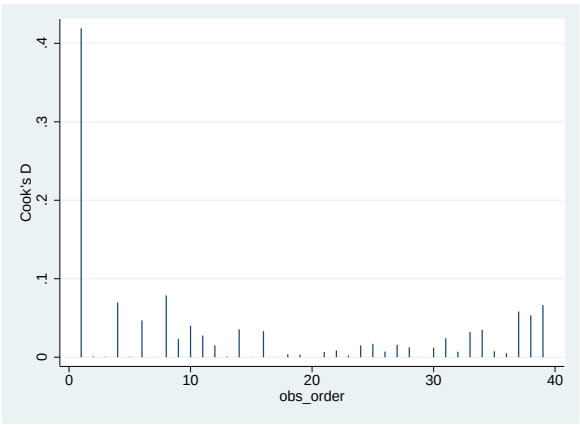
Another issue that I wish to address concerns the interpretation of the regression coefficients in a multiple regression model. For our problem, let us first focus on the fraction coefficient in the fitted model

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

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.

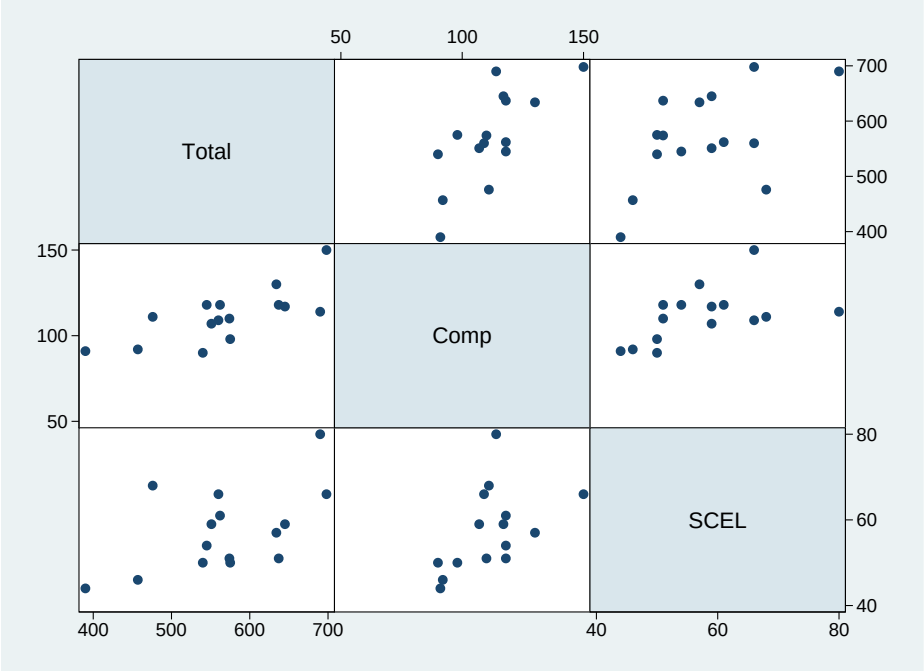
I will lead you through a number of steps to help you answer this question. Let us answer the following straightforward questions based on the **Stata** output.

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.

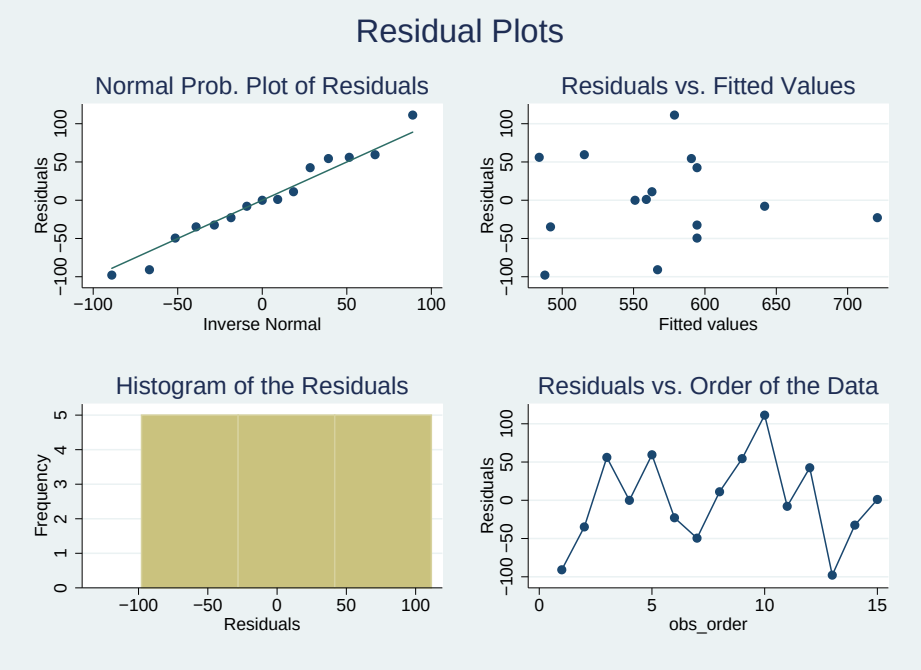
Although SCEL appears to be useful as a predictor of TOTAL on its own, the multiple regression output indicates that SCEL does not explain a significant amount of the variation in TOTAL, once the effect of COMP has been taken into account. In particular, the SCEL effect in the multiple regression model 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.

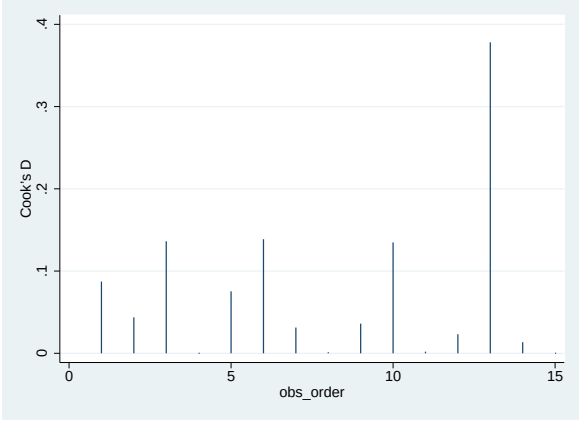
What are my conclusions? Given that SCEL is not a useful predictor in the multiple regression model, I would propose a simple linear regression model to predict TOTAL from COMP:

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

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

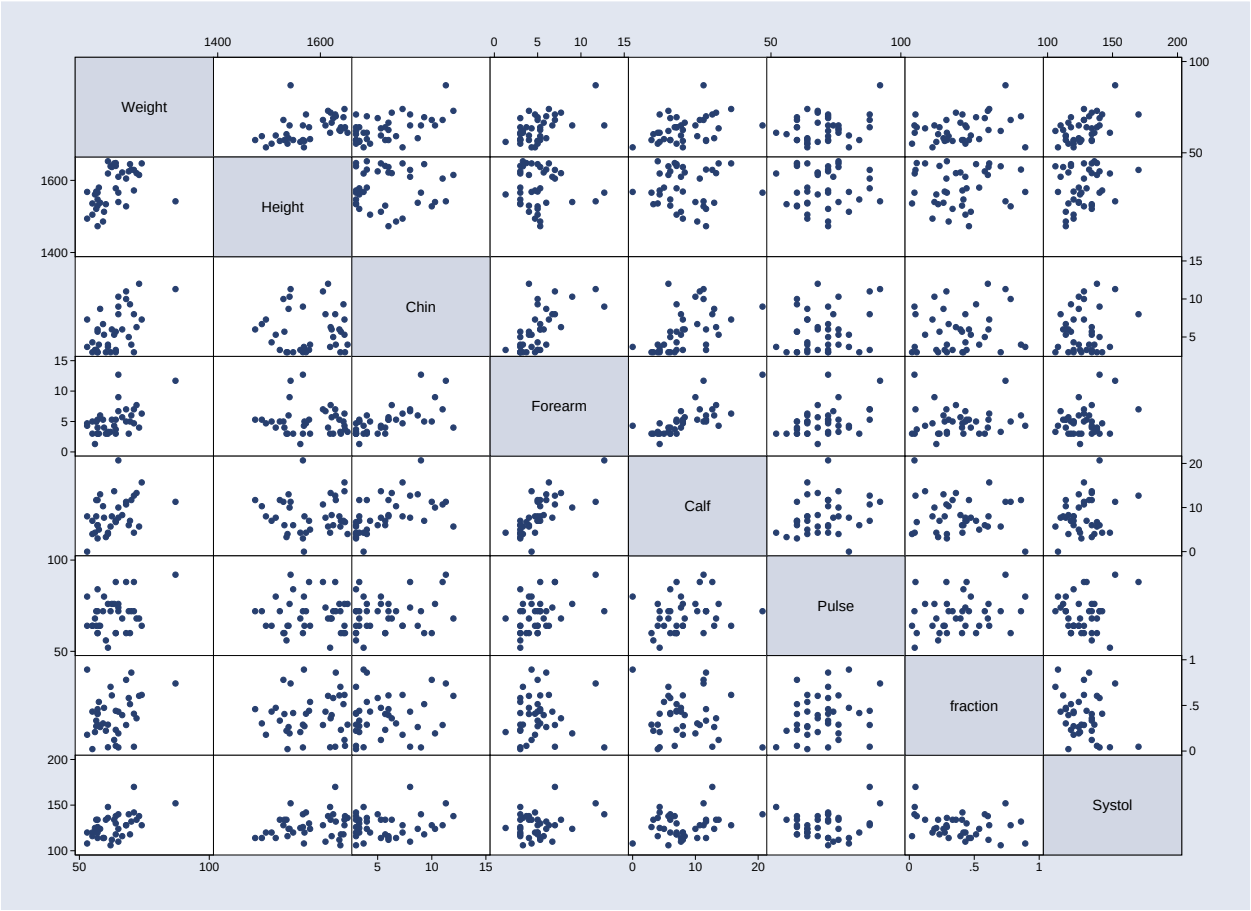
2. Find the variable which when omitted from the full model (1) reduces R^2 the least, or equivalently, increases the Residual SS the least. This is the variable that gives the largest p-value for testing an individual regression coefficient $H_0 : \beta_i = 0$ for $i > 0$. Suppose this variable is X_k. 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.