# STAT2008 – REGRESSION MODELLING
# LECTURE NOTES - CHAPTER 1: SIMPLE LINEAR REGRESSION

## I. Introduction

The basic aims of this chapter are:

- Review of the simple linear regression material covered in Statistical Techniques II;
- An introduction to some new notation, including matrices;
- A more detailed study of the properties of the regression estimates; and,
- An investigation of diagnostic procedures to check the credibility of the underlying assumptions of our regression model.

We will, as much as possible, demonstrate concepts through the use of example data. This will also give us opportunity to see how to use *S-Plus* to perform our fitting and diagnostic procedures.

When formulating a suitable model for a set of data, we should always take into account:

1. Background scientific theory which may suggest a specific structure for our model;
2. Scatterplots of the data; and,
3. Statistical model output and diagnostic procedures.

## II. The Model and Assumptions

If our dataset consists of a sample of $n$ pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$, where the $Y_i$'s are considered to be the values of a "response" or "dependent" variable (i.e., the variable whose characteristics we are most interested in examining and explaining) and the $x_i$'s are the values of a "predictor" or "independent" variable (i.e., a variable whose value may potentially influence the value of the response or dependent variable), then the simplest possible regression structure has the linear form:

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where $\epsilon$ is a mean-zero random variable having variance $\sigma^2$. Specifically, this means that we believe that each data value $Y_i$ can be expressed as:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad (i = 1, \ldots, n)$$

where the $\epsilon_i$'s are the "errors" or "noise" in the model; that is, they are the *stochastic* or *random* component of $Y_i$, and they measure the amount by which the observed value differs from what the "deterministic" part of the model would have predicted for the value of $Y_i$, namely $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$. We use the notation of conditional expectation to denote that the value we expect for the dependent variable depends on the associated value of the predictor. Since we are thus effectively conditioning on the values of the predictor variables, we generally make the simplifying assumption that these values are non-random and measured without, or with only a negligible, error. However, we shall investigate briefly what happens when this is not the case a little later on.

i. Parameter Interpretation

The constants $\beta_0$, $\beta_1$ and $\sigma^2$ are called *parameters*, and they determine the basic underlying relationship between the two variables under study. The value of $\beta_1$ is called the *slope* of the regression line, $\beta_0$ is referred to as the *intercept* of the regression line and $\sigma$ is called the *scale* of the regression errors. The parameter $\beta_1$ measures the change in the deterministic portion of the response variable, $E(Y|x)$, associated with a unit change in the value of the predictor variable, $x$. At first glance, the parameter $\beta_0$ would seem to have the interpretation of being the value of $E(Y|x)$ when $x = 0$ and indeed the name "intercept" would tend to confirm this notion. However, if the range of the predictor variables is far from the origin, this interpretation can lead to misleading and even nonsensical answers. This illustrates the danger of "extrapolating" our regression model outside the range of our data. For this reason, it is usually best to think of $\beta_0$ as simply a structural baseline component of the model from which comparisons within the range of the predictor values can be made.

Note that the interpretation of $\beta_1$ (and also of $\beta_0$) implies that its value depends on the units in which the predictor and response variables are measured. If we were to change the units of measurement for one or both of the variables, we would necessarily change the quantitative (though not the qualitative) interpretation of the parameters. This dependency on the units of measurement is taken into account by the scale parameter $\sigma$. If we change our units of measurement then we will necessarily change our interpretation of $\beta_1$ and this change will be exactly compensated for by a corresponding change in the scale parameter, so that overall the conclusions which we draw from our regression model are unaffected by the units chosen for measurement of the variables.

The dependence of the values of the parameters on the units of measurement can still be problematic, however, not least because of computational considerations which generally imply that numerical algorithms tend to work best for moderate size data values. In addition, the lack of real interpretability of the intercept, $\beta_0$, is a nuisance. To this end, it is often useful to consider an alternate, centered form of the linear model:

$$Y_i = \beta_0^\star + \beta_1^\star\left(\frac{x_i - \overline{x}}{s_x}\right) + \epsilon_i \qquad (i = 1, \ldots, n),$$

where $\overline{x}$ is the average and $s_x$ is the standard deviation of the predictor values. Thus, for this alternate form, the parameter $\beta_0^\star$ is the value of $E(Y|x)$ when the value of the predictor is equal to the average of the observed values of the predictor variable. So, we now see that the value of this new "intercept" term has a more sensible meaning, as it is the value of the deterministic part of the model in the "center" of the data. Similarly, the parameter $\beta_1^\star$ indicates the change in $E(Y|x)$ corresponding to a single *standard* unit change in the predictor variable (i.e., a change of size $s_x$, the sample standard deviation), and thus, its units are simply dependent on the measurement scale of the response variable and not that of the predictor. Simple algebra shows that these two forms of the model are in fact identical since:

$$\beta_0^\star = \beta_0 + \beta_1\overline{x}; \qquad \beta_1^\star = s_x\beta_1.$$

In fact, we can similarly center and scale the response variables and write the model as:

$$\frac{Y_i - \overline{Y}}{s_Y} = \beta_0^{\star\star} + \beta_1^{\star\star}\left(\frac{x_i - \overline{x}}{s_x}\right) + \epsilon_i^\star \qquad (i = 1, \ldots, n),$$

where $\overline{Y}$ and $s_Y$ are now the sample mean and standard deviation of the response variables, and $\epsilon_i^\star$ is a new mean-zero random variable having scale $\sigma^\star$. Again, simple algebra shows that:

$$\beta_0^{\star\star} = \frac{\beta_0 - \overline{Y} + \beta_1\overline{x}}{s_Y}; \qquad \beta_1^{\star\star} = \frac{s_x\beta_1}{s_Y}; \qquad \sigma^\star = \frac{\sigma}{s_Y}.$$

The parameter $\beta_1^{\star\star}$ is now the number of standard units (i.e., the number of $s_Y$'s) by which the expectation of the predictor value is changed when the predictor value is shifted by a single standard unit, $s_x$, and is thus a unitless quantity. The new interpretation of the parameter $\beta_0^{\star\star}$ is slightly less obvious, but can be seen to be the vertical distance from the regression line to the point $(\overline{x}, \overline{Y})$, often called the "centroid" of the data. As we shall see shortly, the method of least-squares fitting will guarantee that the regression line will pass through the centroid and thus $\beta_0^{\star\star}$ will necessarily be zero.

Each of these three formulations of the simple linear model are equivalent, and thus it makes no theoretical difference which one we choose. However, the quantitative interpretations of the parameters does change as we have seen. Also, the final form of the model is generally easier to deal with computationally, a fact which will surface again when we talk about models with more than one predictor variable.

ii. Model Assumptions

As we noted above, we shall generally assume that the values of the predictor variable are fixed and measured without any discernible error. In addition, we assume that the $\epsilon_i$'s are random variables with expectation $E(\epsilon_i) = 0$ and variance $Var(\epsilon_i) = \sigma^2$, regardless of the value $x_i$ of the predictor variable with which they are associated. More precisely, we might write:

$$E(\epsilon_i|x_i) = 0; \qquad E(\epsilon_i^2|x_i) = \sigma^2 \qquad (i = 1, \ldots, n).$$

This assumption that the scale of the errors is the same regardless of the value of the predictor value is called the assumption of *homoscedasticity*, and is crucial to the subsequent theory. It is this assumption which allows us to estimate the scale of the data, since if the scale changed at each of the values of the predictor (i.e., the data were *heteroscedastic*), then we would effectively have only one observation to estimate each of these different scales and thus would be unable to fit any regression. Of course, if we had multiple observations at each value of the predictor (a so-called "repeated measures" design) or if we were to assume some form for how the variability changed as the value of the predictor changed, then we might still be able to fit our model. Such ideas, however, are somewhat outside the scope of this course, and will not pursued in any detail. In addition, we shall also assume that the $\epsilon_i$'s are *uncorrelated*. Loosely speaking, this simply means that the value of one $\epsilon_i$ does not effect the values of the others. Mathematically, we will write this assumption as:

$$E(\epsilon_i \epsilon_j | x_i, x_j) = 0 \qquad \forall i \neq j.$$

Again, this assumption will be critical to our subsequent theory.

For the most part, we will also assume that the $\epsilon_i$'s are normally distributed. This assumption will allow us to make precise confidence statements and construct confidence intervals for the parameters of our models. However, for many issues, such as the bias and variability of our estimates, the assumption of normality will not be critical, and we will endeavor to point out where the normal assumptions have been used and where they have not.

Finally, we also note that we have made the assumption that the chosen form of the model does in fact capture the structural relationship between the response and predictor variable. This assumption may seem rather obvious, however, it is essential to any meaningful subsequent interpretation of our modelling results.

iii. Matrix Notation

Before we continue with our discussion of the details of the simple linear model, we should introduce some new notation which will facilitate both the calculations for the model itself as well as the preparation for later generalizations of the model to situations with more that one predictor variable.

First, we shall denote the values of the response variable in a column vector of length $n$, $Y = (Y_1, \ldots, Y_n)^T$. Similarly, we will place the values of the predictor variable in a column of the so-called *design matrix* (the name indicating that we are assuming the values of the predictor variable are fixed and thus have in some sense been "chosen" by the investigator by design). For a simple linear model, the design matrix has dimension $n \times 2$, the first column composed entirely of ones (to account for the intercept in the model as well as a slope):

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

The errors will also be denoted by a column vector, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. So, $\epsilon$ is a mean-zero random vector with variance-covariance matrix $\sigma^2 I$, where $I$ is the $n \times n$ identity matrix.

The *variance-covariance* matrix of a random vector is simply a two-dimensional array containing all the covariances of each possible pair of the components of the vector. Thus, if $Z = (Z_1, \ldots, Z_n)^T$ is a random vector, then its variance-covariance matrix is defined by

$$Var(Z) = V = \begin{pmatrix} v_{11} & v_{12} & \ldots & v_{1n} \\ v_{21} & v_{22} & \ldots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \ldots & v_{nn} \end{pmatrix}$$

where the components of $V$ are simply defined by the relationship: $v_{ij} = Cov(Z_i, Z_j)$.

Note that this implies that a variance-covariance matrix is symmetric (i.e., $V^T = V$), since $Cov(Z_i, Z_j) = Cov(Z_j, Z_i)$, and that its diagonal elements are the variances of the individual components of $Z$, since $v_{ii} = Cov(Z_i, Z_i) = Var(Z_i)$. Thus, $Var(\epsilon) = \sigma^2 I$ simply states that the components of $\epsilon$ (the $\epsilon_i$'s) are uncorrelated and have common variance $\sigma^2$.

Now, under the assumption of normality of the errors, $\epsilon$ will have a multivariate normal distribution. Lastly, we will denote the slope and intercept by a column vector of length 2, $\beta = (\beta_0, \beta_1)^T$, so that the model may be written as:

$$Y = X\beta + \epsilon.$$

## III. Parameter Estimation

Once we have settled upon a model, we must decide how to estimate the parameters of that model. Generally, the method of *least-squares* is employed for this purpose. The general concept behind this approach is to choose that member of the model class (i.e., the values of the slope and intercept for a particular line from among the set of all possible lines) which has the minimum "total distance" from the data.

i. Least Squares Estimation

Specifically, the method of least-squares starts by assigning a distance function

$$d(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2,$$

which measures the discrepancy between any line and the observed dataset. The least-squares estimates of the true parameters $\beta_0$ and $\beta_1$ are then those values of $b_0$ and $b_1$ (also sometimes denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$) which minimize the distance function $d$. The motivation behind this procedure is that the resulting *fitted regression line*, $y = b_0 + b_1 x$, will be "close" to all the observed data points in some sense. More specifically, we will define the *fitted* value for each data point by $\hat{Y}_i = b_0 + b_1 x_i$, and the associated *residual* value by $e_i = Y_i - \hat{Y}_i$ (also occasionally denoted as $\hat{\epsilon}_i$). Then, the least-squares estimates are those values of the slope and intercept which make the *sum of the squared residuals* as small as possible. In other words, the total vertical distance from the data to the fitted line is minimized.

To actually calculate the minimizing values of the distance function $d$, we need simple calculus. The minimizing values will solve the *normal equations*:

$$\begin{aligned}\frac{\partial d}{\partial b_0} &= -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 x_i) = 0\\ \frac{\partial d}{\partial b_1} &= -2\sum_{i=1}^{n} x_i(Y_i - b_0 - b_1 x_i) = 0.\end{aligned}$$

A bit of algebra then shows that these two equations become:

$$\begin{aligned}\sum_{i=1}^{n} Y_i &= nb_0 + b_1\sum_{i=1}^{n} x_i = n(b_0 + b_1\overline{x})\\ \sum_{i=1}^{n} x_i Y_i &= b_0\sum_{i=1}^{n} x_i + b_1\sum_{i=1}^{n} x_i^2 = n(b_0\overline{x} + b_1\overline{x}^2) + b_1 S_{xx},\end{aligned}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i - \overline{x})x_i = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$. Thus, solving the first equation of this system yields:

$$b_0 = \overline{Y} - b_1\overline{x},$$

and substituting this solution into the second equation shows that

$$\sum_{i=1}^{n} x_i Y_i = b_0\sum_{i=1}^{n} x_i + b_1\sum_{i=1}^{n} x_i^2 = n[(\overline{Y} - b_1\overline{x})\overline{x} + b_1\overline{x}^2] + b_1 S_{xx} = n\overline{x}\overline{Y} + b_1 S_{xx}.$$

Thus,

$$b_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n}\overline{x}Y_i}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{S_{xx}} = \frac{S_{xy}}{S_{xx}},$$

where $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})Y_i = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y}$.

Note that when the predictor takes the value $\overline{x}$, then

$$\begin{aligned}\hat{Y}(\overline{x}) &= \text{Estimate of } E(Y|\overline{x})\\ &= b_0 + b_1\overline{x}\\ &= \overline{Y} - b_1\overline{x} + b_1\overline{x}\\ &= \overline{Y},\end{aligned}$$

so that the centroid of the data is a point on the least-squares regression line. This fact can also be seen by substituting $b_0$ and $b_1$ into the expression given for $\beta_0^{\star\star}$, to show that the

least-squares estimate of this quantity is $b_0^{\star\star} = 0$. [NOTE: We have introduced the notation $\hat{Y}(x) = b_0 + b_1 x$ as the point on the regression line corresponding to a predictor value of $x$, which is thus an estimate of the value $E(Y|x) = \beta_0 + \beta_1 x$. So, for the actual data values, we have $\hat{Y}(x_i) = \hat{Y}_i = b_0 + b_1 x_i$ which is an estimate of $E(Y|x_i) = \beta_0 + \beta_1 x_i$.]
Of course, all of this could be done using matrix notation, by defining $b = (b_0, b_1)^T$, so that the distance function becomes:

$$d(b) = (Y - Xb)^T (Y - Xb).$$

Taking the derivative and setting it equal to zero then yields the normal equations in the form:

$$\frac{\partial d}{\partial b} = -2X^T(Y - Xb) = 0.$$

Solving this equation shows quite easily that

$$X^T X b = X^T Y \qquad \Longrightarrow \qquad b = (X^T X)^{-1} X^T Y.$$

A bit of algebra shows that:

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}; \qquad (X^T X)^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix};$$

$$X^T Y = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix},$$

which demonstrates that the two approaches do indeed produce the same answers, since:

$$\begin{aligned}
b = (X^T X)^{-1} X^T Y &= \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} \sum_{j=1}^n Y_j \\ \sum_{j=1}^n x_j Y_j \end{pmatrix} \\
&= \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 \sum_{j=1}^n Y_j - \sum_{i=1}^n x_i \sum_{j=1}^n x_j Y_j \\ n\sum_{j=1}^n x_j Y_j - \sum_{i=1}^n x_i \sum_{j=1}^n Y_j \end{pmatrix} \\
&= \frac{1}{S_{xx}} \begin{pmatrix} \overline{Y} \sum_{i=1}^n x_i^2 - \overline{x} \sum_{j=1}^n x_j Y_j \\ \sum_{j=1}^n x_j Y_j - n\overline{x}\overline{Y} \end{pmatrix} \\
&= \frac{1}{S_{xx}} \begin{pmatrix} \overline{Y} \sum_{i=1}^n x_i^2 - n\overline{Y}\overline{x}^2 + n\overline{Y}\overline{x}^2 - \overline{x} \sum_{j=1}^n x_j Y_j \\ S_{xy} \end{pmatrix} \\
&= \frac{1}{S_{xx}} \begin{pmatrix} \overline{Y} S_{xx} - \overline{x} S_{xy} \\ S_{xy} \end{pmatrix} \\
&= \begin{pmatrix} \overline{Y} - \overline{x}\frac{S_{xy}}{S_{xx}} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix}.
\end{aligned}$$

This form of the least-squares estimator leads to another useful matrix, since we may write

$$\hat{Y} = Xb = X(X^TX)^{-1}X^TY = HY,$$

where $H = X(X^TX)^{-1}X^T$ is called the "hat matrix" since multiplying the response vector,$Y$, by $H$ yields $\hat{Y}$. Similarly, the vector of residuals can be written as

$$e = Y - \hat{Y} = Y - HY = (I - H)Y.$$

From this expression, the variance-covariance matrix of the residuals is easily calculated to be:

$$Var(e) = \sigma^2(I - H).$$

The matrix $H$ is of interest in its own right, since its diagonal elements are a measure of the *influence* of each data point. More will be said about the concept of *influence* shortly, however, a simple algebraic exercise shows that the $i^{\text{th}}$ diagonal element of $H$ is:

$$h_{ii} = \frac{\sum_{j=1}^{n}(x_j - x_i)^2}{nS_{xx}},$$

and the value $h_{ii}$ is referred to as the *leverage* of the $i^{\text{th}}$ data point.

ii. Properties of Least-Squares Estimators

It is not difficult to show that the estimates $b_0$ and $b_1$ are unbiased:

$$\begin{aligned} E(b_1) = E\left(\frac{S_{xy}}{S_{xx}}\right) &= \sum_{i=1}^{n}\frac{(x_i - \overline{x})E(Y_i)}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} \\ &= \frac{\beta_0\sum_{i=1}^{n}(x_i - \overline{x}) + \beta_1 S_{xx}}{S_{xx}} = \beta_1, \end{aligned}$$

and

$$E(b_0) = E(\overline{Y} - b_1\overline{x}) = \frac{1}{n}\sum_{i=1}^{n}E(Y_i) - \beta_1\overline{x} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \beta_1\overline{x} = \beta_0.$$

In fact, it is even easier to demonstrate these results using matrix notation, since $Y = X\beta + \epsilon$ implies that the least-squares estimator, $b = (X^TX)^{-1}X^TY$, may be written as:

$$(X^TX)^{-1}X^T(X\beta + \epsilon) = (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\epsilon = \beta + (X^TX)^{-1}X^T\epsilon,$$

which implies that $E(b) = E\{\beta + (X^TX)^{-1}X^T\epsilon\} = \beta + (X^TX)^{-1}X^TE(\epsilon) = \beta$, since $E(\epsilon) = 0$. The preceding calculations only required that the $x_i$'s were considered to be non-random (i.e., the calculations should be thought of as being done conditional on the observed values of the predictor variables) and that the $\epsilon_i$'s had zero mean, $E(\epsilon_i) = 0$.

Similarly, we can calculate the variance of each of the estimators, however, we will now need the assumptions of homoscedasticity and of uncorrelatedness for the $\epsilon_i$'s in addition to those assumptions mentioned above. So,

$$\begin{aligned} Var(b_1) &= Var\left(\sum_{i=1}^{n}\frac{(x_i - \overline{x})Y_i}{S_{xx}}\right) \\ &= \sum_{i=1}^{n}\frac{(x_i - \overline{x})^2}{S_{xx}^2}Var(Y_i) + 2\sum_{i \neq j}\frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}^2}Cov(Y_i, Y_j) \\ &= \frac{\sigma^2\sum_{i=1}^{n}(x_i - \overline{x})^2}{S_{xx}^2} \\ &= \frac{\sigma^2}{S_{xx}}. \end{aligned}$$

And for the intercept estimator, we have:

$$\begin{aligned} Var(b_0) &= Var(\overline{Y} - b_1\overline{x}) \\ &= Var(\overline{Y}) + \overline{x}^2 Var(b_1) - 2\overline{x}Cov(\overline{Y}, b_1) \\ &= \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right), \end{aligned}$$

where we have used the fact that $\overline{Y}$ and $b_1$ are uncorrelated random variables. To see this fact, first note that

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \overline{x})(\beta_0 + \beta_1 + \epsilon_i)}{S_{xx}} = \beta_1 + \sum_{i=1}^n \frac{(x_i - \overline{x})\epsilon_i}{S_{xx}},$$

and

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1\overline{x} + \frac{1}{n}\sum_{i=1}^n \epsilon_i.$$

Therefore,

$$\begin{aligned} Cov(\overline{Y}, b_1) &= Cov\left(\beta_0 + \beta_1\overline{x} + \frac{1}{n}\sum_{i=1}^n \epsilon_i, \beta_1 + \sum_{j=1}^n \frac{(x_j - \overline{x})\epsilon_j}{S_{xx}}\right) \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \frac{(x_j - \overline{x})}{S_{xx}} Cov(\epsilon_i, \epsilon_j) \\ &= \frac{1}{n}\sum_{i=1}^n \frac{(x_i - \overline{x})\sigma^2}{S_{xx}} \\ &= 0, \end{aligned}$$

since the $\epsilon_i$'s are assumed to be uncorrelated and all have variance $\sigma^2$.
Again, matrix notation simplifies the task, since

$$\begin{aligned} Var(b) &= Var\{\beta + (X^TX)^{-1}X^T\epsilon\} \\ &= (X^TX)^{-1}X^TVar(\epsilon)\{(X^TX)^{-1}X^T\}^T \\ &= (X^TX)^{-1}X^T(\sigma^2 I)X\{(X^TX)^{-1}\}^T \\ &= \sigma^2(X^TX)^{-1}X^TX\{(X^TX)^{-1}\}^T \\ &= \sigma^2(X^TX)^{-1}, \end{aligned}$$

and our previous form for $(X^TX)^{-1}$ then shows that the two procedures give identical results (once we employ the identity $\sum_{i=1}^n x_i^2 = S_{xx} + n\overline{x}^2$). [NOTE: We have made use of two important matrix facts:

· For any matrix $A$, $(A^{-1})^T = (A^T)^{-1}$; and,

· If $Z$ is a random vector and $A$ is any matrix of constants, then $Var(AZ) = AVar(Z)A^T$.]

Lastly, we need an estimate of $\sigma^2$. To do so, we need to have estimates of the $\epsilon_i$'s, and the obvious choice is clearly the residuals, $e_i$. So, since $\sigma^2$ is the variance of the $\epsilon_i$'s, it should not be surprising that the estimator that is generally used is based on the sample variance of the $e_i$'s. Specifically, we will use the *residual sum of squares* (also called the *sum of squared errors*, $SSE$) divided by its appropriate degrees of freedom, $n - 2$,

$$s^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2},$$

and in matrix notation we have

$$\begin{aligned} s^2 &= \frac{e^Te}{n-2} = \frac{(Y-\hat{Y})^T(Y-\hat{Y})}{n-2} = \frac{(Y-HY)^T(Y-HY)}{n-2} \\ &= \frac{Y^T(I-H)^T(I-H)Y}{n-2} = \frac{Y^T(I-H)Y}{n-2}, \end{aligned}$$

where the last line follows from the fact that $H$ and $I-H$ are *projection* matrices (any matrix, $A$, is called a projection matrix if it satisfies the identities, $A^T = A$ and $AA = A^2 = A$; see Section IV and the first set of tutorial exercises).

The denominator of $n-2$ is referred to as the *error degrees of freedom* (or the *residual degrees of freedom*) and its use ensures that $s^2$ is an unbiased estimator of $\sigma^2$ *under the assumption that the model is correct* (see tutorial exercises). The estimator $s^2$ is usually referred to as the *mean squared error*, $MSE$, of the regression. Once we have calculated $s^2$, we can now easily use it to calculate estimates of the standard errors (i.e., the square roots of the variances) of the parameter estimates, $b_0$ and $b_1$.

The reason that the residual sum of squares has $n-2$ residual degrees of freedom can be seen in various ways. The simplest explanation is that, as with the case of the usual degrees of freedom for a one-sample variance estimate where one out of the total $n$ degrees of freedom is lost due to the estimation of the mean, here two degrees of freedom are lost due to the estimation of the slope and the intercept. More precisely, the residuals, by their construction, must satisfy the two linear constraints:

$$\sum_{i=1}^{n} e_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} x_i e_i = 0,$$

which are just restatements of the normal equations used to calculate the estimates $b_0$ and $b_1$. [NOTE: Compare this to the case of the one-sample average problem, where the deviations $X_i - \overline{X}$ must sum to zero.]

Another, more statistical way of seeing why the denominator is $n-2$ rather than $n$ is to recall that $Var(e) = \sigma^2(I-H)$. Since $h_{ij}$ will typically be non-zero, we can see that the residuals are not uncorrelated. Indeed, the fact that they must satisfy the two constraints listed above means that once we know $n-2$ of them we can automatically determine the remaining two. Moreover, it turns out that the correlations between the residuals are always positive, meaning that they tend to be closer to each other in value than we might expect, and, more importantly, closer to each other than the corresponding $\epsilon_i$'s are, so that $e^Te$ will typically underestimate the quantity $\epsilon^T\epsilon$, which is why we must divide by a slightly smaller denominator ($n-2$ instead of $n$) to overcome this correlation and make our estimator unbiased.

*Example 1 - Protein in Pregnancy Data:* Suppose that the protein in pregnancy data has been stored in the matrix `protpreg` in *S-Plus*, the first column containing the protein concentrations and the second column containing the gestations. We can then use *S-Plus* to find the parameter estimates using the following commands:

```
> protein _ protpreg[,1]
> gestation _ protpreg[,2]
> reg.out _ lsfit(gestation,protein)
> reg.out$coef
 Intercept          X
 0.2017377 0.02284426
> reg.diag _ ls.diag(reg.out)
> reg.diag$std.dev^2
[1] 0.01324297
```

Thus, $b_0 = 0.2017$, $b_1 = 0.0228$ and $s^2 = 0.0132$. We can also have *S-Plus* calculate the standard errors of the slope and intercept estimators as:

```
> reg.diag$std.err
                  [,1]
 Intercept 0.083363149
         X 0.003294676
```

So, the estimated standard error of the intercept is $s(b_0) = 0.0834$ and the estimated standard error of the slope is $s(b_1) = 0.0033$.

### IV. ANOVA Tables and Partitioning Variability

In any regression dataset, there will be variation in the values of the response variable. One important interpretation of a regression model is that it (at least partially) "explains" some of this variation. In other words, the overall variation in the response variable can be seen as consisting of two pieces, the *systematic* variation (i.e., that variation which is "explained" by the fact that the associated values of the predictor variable are different) and the remaining *stochastic* variation (i.e., that variation which is not due to the measured predictor variable). This second source of variation is assumed to be the product of the laws of chance for the purposes of the modelling exercise, however, as we shall discuss further later, it may at least in part be "explainable" variation (i.e., due to the variation in some as yet unmeasured predictor variable). This breakdown of total variation of the response can be formalized in the following mathematical identity:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

or, symbolically:

$$SST = SSR + SSE$$

Here, $SST$ is called the *total sum of squares* of the response, and is basically just the sample variance of the response data ignoring the predictor variable, while $SSR$ is the *regression sum of squares* and measures the amount of variation which would have been expected if the response values all lay exactly along the regression line. In this sense, the $SSR$ is a measure of how much of the variation of the response is "explained" by the regression model. And of course, $SSE$ is just the residual or error sum of squares which we have already seen, and it measures the left-over, "unexplained", variation in the response. To demonstrate that the above identity is indeed true, we start with the equality:

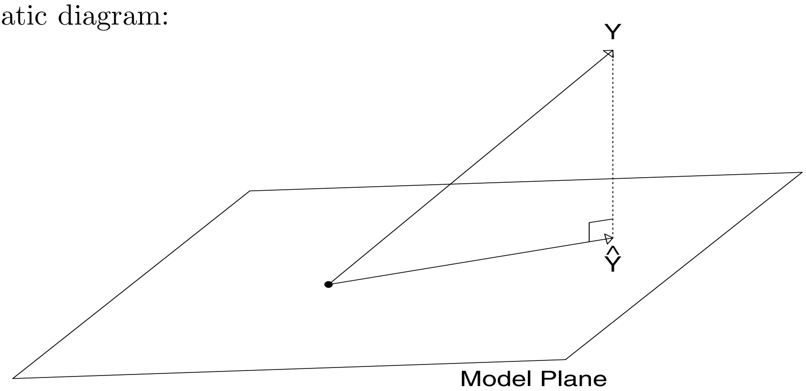
$$Y_i - \overline{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y}).$$

Then, squaring and summing both sides yields

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}).$$

However, the last term on the right-hand side of the above equation is exactly zero (see tutorial exercises).

One useful way to understand what least-squares regression is actually doing is through the use of the following schematic diagram:



Here, the main vector, $Y$, represents the data (actually, it really represents the *mean-corrected* data, $Y - \overline{Y}$), and the planar region represents the *model space*. In other words, this planar region represents all possible lines, and any vector in this place represents a dataset which would lie exactly along some line. Obviously, the actual data will never fall precisely along a line, and the degree to which a line is a good representation of the data is measured by the closeness of the vector to the plane. The projection of the data vector on the model plane produces the fitted values, $\hat{Y} = HY$, which is why the matrix $H$ is referred to as a projection matrix. Finally, the perpendicular vector connecting the data to the fitted values represents the residuals, $Y - \hat{Y}$. The residual sum of squares is just the length of the residual vector in the diagram, and the regression sum of squares is the length of the fitted vector, so the sum of squares breakdown described above can now be seen as a simple application of the Pythagorean theorem.

Now, under the normal theory assumption, it can be shown that the quantity $SSE/\sigma^2$ has a chi-squared distribution with $n - 2$ degrees of freedom. Similarly, under the condition that $\beta_1 = 0$, then $SSR/\sigma^2$ has a chi-squared distribution with a single degree of freedom. Also, if we ignore the predictor variable, it is not difficult to recognize the relationship between $SST$ and the sample variance of the observed response data, so it should not be surprising that $SST/\sigma^2$ has a chi-squared distribution with $n - 1$ degrees of freedom. So, we see that there is a breakdown for the *total degrees of freedom* into the sum of the *regression* or *model degrees of freedom* and the *residual* or *error degrees of freedom*:

$$\begin{array}{rcccl} SST & = & SSR & + & SSE \qquad \text{(Sums of Squares)} \\ (n-1) & = & 1 & + & (n-2) \qquad \text{(Degrees of Freedom)} \end{array}$$

We can now use these results to construct confidence intervals and perform hypothesis tests regarding the regression parameters, $\beta_0$ and $\beta_1$, as well as other aspects of the regression such as predictions.

i. Significance of the Regression

   One obvious question concerning the two variables under study is whether the predictor actually has any influence on the associated values of the response. Such a question can generally be answered through the use of an appropriate hypothesis test of whether the data are consistent with the idea that $\beta_1 = 0$, i.e., that the slope of the true underlying regression is zero. Mathematically, this would indicate that the quantity $E(Y|x)$ would not depend on the specific value of $x$, so that $E(Y|x) = E(Y)$ for all $x$ values. We will signify this hypothesis test as:

$$H_0 : \ \beta_1 = 0 \qquad \text{vs.} \qquad H_A : \ \beta_1 \neq 0.$$

We will shortly discuss a procedure which will allow us to determine whether we should "reject" the so-called *null hypothesis*, $H_0$, in favor of the *alternative* hypothesis, $H_A$. However, it should be pointed out that rejection of the null hypothesis must be carefully interpreted. The rejection of $H_0$ merely indicates that a trend has been detected in the dataset, but does not imply anything regarding the actually suitability of the chosen linear model, its usefulness for the purposes of future predictions, or any causal relationship between the variables.

We have seen that $SSE/\sigma^2$ has a chi-squared distribution with $n-2$ degrees of freedom, and that when $\beta_1 = 0$ (i.e., if the null hypothesis is actually true) that $SSR/\sigma^2$ has a chi-squared distribution with one degree of freedom. It turns out that $SSE$ and $SSR$ are independent as well, so that the statistic:

$$F = \frac{SSR/\sigma^2}{SSE/\{\sigma^2(n-2)\}} = \frac{SSR}{SSE/(n-2)} = \frac{MSR}{MSE} = \frac{MSR}{s^2},$$

has an $F$-distribution with one numerator degree of freedom and $n-2$ denominator degrees of freedom under $H_0$. Here, the quantity $MSR = SSR/1$ is called the *mean square for regression* and is just the regression sum of squares divided by its appropriate degrees of freedom (this may seem silly at this point since the appropriate degrees of freedom for the $SSR$ in a simple linear regression is just one, however, it will be more important when we discuss models with more than one predictor variable). The $F$-statistic (also called the $F$-ratio) can be seen as the ratio of the variability "explained" by the model to the variability left "unexplained". So, if $F$ is large, this indicates that there is a large proportion of the overall variability in the response variable which is accounted for by the model. Thus if $F$ is large, we should favor the alternative hypothesis that $\beta_1 \neq 0$ since this is the hypothesis that there is some relationship between the response and the predictor variable. Specifically, we will reject $H_0$ at significance level $\alpha$ if the observed $F$-ratio is larger than the critical value $F_{1,n-2}(1-\alpha)$, which is the $(1-\alpha)$-quantile of an $F$-distribution with one numerator degree of freedom and $n-2$ denominator degrees of freedom. Mathematically, we know that $s^2$ is unbiased, so that $E(s^2) = \sigma^2$, and it is not hard to show (see tutorial exercises) that

$$E(MSR) = \sigma^2 + \beta_1^2 S_{xx}.$$

Thus, if $\beta_1$ is not zero, then we would expect the $F$-ratio to be larger than one since $\beta_1^2 S_{xx}$ will always be a non-negative number. This is why we reject the null hypothesis for large values of the $F$-ratio but not for small ones.

*Example 1 (cont'd) - Protein in Pregnancy Data:* Suppose that we have now read the protein in pregnancy data into an *S-Plus* dataframe called `protpreg.df`, with the variables named `protein` and `gestation`. We can then use the *S-Plus* command `lm()` to calculate each of the above noted quantities and display them in a standard format, namely the so-called *analysis of variance table*:

```
> protpreg.lm _ lm(protein ~ gestation, data=protpreg.df)
> protpreg.lm
Call:
lm(formula = protein ~ gestation, data = protdf)

Coefficients:
 (Intercept) gestation
```

```
  0.2017377 0.02284426

Degrees of freedom:  19 total; 17 residual
Residual standard error:  0.1150781
> anova(protpreg.lm)
Analysis of Variance Table

Response:  protein

Terms added sequentially (first to last)
          Df Sum of Sq   Mean Sq  F Value        Pr(F)
gestation  1 0.6366696 0.6366696 48.07606 2.415694e-06
Residuals 17 0.2251304 0.0132430
> qf(0.95,1,17)
[1] 4.45132
> qf(0.99,1,17)
[1] 8.39974
```

So, we see that the observed value of the $F$-ratio is 48.08, which is much greater than either $F_{1,17}(0.95) = 4.45$ or $F_{1,17}(0.99) = 8.40$, implying that we should reject the null hypothesis and conclude that there is some relationship between protein levels and gestation. Indeed, the column in the analysis of variance (ANOVA) table headed "*Pr(F)*" provides us with the $p$-value for this significance test, and it is extremely small.

ii. The Coefficient of Determination

Another way of measuring the relationship between our response and predictor variables is to assess the degree to which the model "explains" the variability of the response. Such a measure can be easily constructed from the breakdown of the total sum of squares, $SST$. The *coefficient of determination*, denoted by $R^2$, is defined as:

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$

The coefficient of determination can thus be seen to be a measure of the proportion of the total variability in the response which is "attributable" to the predictor variable. In other words, the values of the response are seen to vary about their sample average for at least two reasons: purely random fluctuations and the fact that their associated values of the predictor variable are different. However, if we take the predictor variable into account, and thus measure the variability of the values of the response about the regression line, we are left with only that variability which is purely random. The resulting drop in our measure of the variability of the response values is an indication of how much of the original variability of the response has been "explained" by the predictor variable.

A little algebra (see tutorial exercises) shows that the coefficient of determination is equivalent to the square of the sample correlation coefficient,

$$r = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot SST}}.$$

This fact accounts for the choice of notation $R^2$. Now, if we consider our data as a random sample of pairs, so that the $x_i$'s are no longer fixed, then the sample correlation coefficient, $r$, can be seen as an estimate of the true population correlation, $\rho$. We might then test

$$H_0 : \rho = 0 \qquad \text{versus} \qquad H_A : \rho \neq 0$$

to determine whether there is a significant linear relationship between the two variables. Fortunately, as we shall see a little later, such a test is exactly equivalent to the $F$-ratio test discussed above.

*Example 1 (cont'd) - Protein in Pregnancy Data:* For the protein in pregnancy data, we saw that $SST = 0.6366696 + 0.2251304 = 0.8618$ and the $SSR = 0.6367$. Thus, the coefficient of determination is $R^2 = 0.6366696/0.8618 = 0.7387672$, so that nearly 74% of the variability of the observed responses is explained by the predictor variable. A calculation of the sample correlation coefficient shows that

```
> cor(protdf)
             protein gestation
  protein 1.00000012 0.85951567
gestation 0.85951567 1.00000012
```

and thus $r = 0.85951567 = \sqrt{0.7387672} = \sqrt{R^2}$.

iii. Hypothesis Test and Confidence Interval for the Slope

More than likely, you have seen a different test for the null hypothesis of $\beta_1 = 0$ based on the test statistic:

$$T = \frac{b_1}{s(b_1)} = \frac{b_1\sqrt{S_{xx}}}{s}$$

which has a Student's $t$-distribution with $n-2$ degrees of freedom when the null hypothesis is true (and assuming that the errors are independent and identically normally distributed). This test rejects $H_0$ at the $\alpha$ level of significance if the observed value of $T$ has an absolute value which exceeds $t_{n-2}(1-\alpha/2)$, the $(1-\alpha/2)$-quantile of a Student's $t$-distribution with $n-2$ degrees of freedom. It may seem, then, that we have two different procedures for testing $H_0$. However, a little algebra shows that the regression sum of squares, $SSR$, can be written as:

$$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}(b_0 + b_1 x_i - \overline{Y})^2 = \sum_{i=1}^{n}\{(\overline{Y} - b_1\overline{x}) + b_1 x_i - \overline{Y}\}^2 = \sum_{i=1}^{n} b_1^2(x_i - \overline{x})^2 = b_1^2 S_{xx},$$

which implies that

$$T^2 = \frac{b_1^2 S_{xx}}{s^2} = \frac{SSR/1}{s^2} = \frac{MSR}{s^2} = F.$$

Thus, since it can easily be shown that $\{t_{n-2}(1-\alpha/2)\}^2 = F_{1,n-2}(1-\alpha)$, we see that the two testing procedures are in fact identical.

*Example 1 (cont'd) - Protein in Pregnancy Data:* Previously we saw that for the protein in pregnancy data, $b_1 = 0.02284426$ and $s(b_1) = 0.003294676$, so that

$$T^2 = \frac{(0.02284426)^2}{(0.003294676)^2} = 48.07605 = F.$$

Also, *S-Plus* shows us that:

```
> qt(0.975,17)^2
```

```
[1] 4.451322
> qt(0.995,17)^2
[1] 8.39974
```

which correspond to the appropriate quantiles of the $F_{1,17}$ distribution calculated previously. One might ask why we need two different tests for the same null hypothesis. The usefulness of studying both approaches comes from the fact that each method has certain advantages. The $F$-ratio test is extremely flexible and we will focus on it primarily, since it is the easiest method to generalize to the case of multiple predictor variables. However, the $t$-test is still important, since it provides us with the easiest method of obtaining confidence intervals for the true slope parameter $\beta_1$. Specifically, recall that under the normal theory assumptions, the estimator $b_1$ has a normal distribution:

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

This in turn implies that the *pivot*

$$\frac{(b_1 - \beta_1)}{s/\sqrt{S_{xx}}} \sim t_{n-2}.$$

Therefore, a $100(1-\alpha)\%$ confidence interval for $\beta_1$ can be calculated as:

$$b_1 \pm t_{n-2}(1-\alpha/2)s(b_1) = \left(b_1 - t_{n-2}(1-\alpha/2)\frac{s}{\sqrt{S_{xx}}}, b_1 + t_{n-2}(1-\alpha/2)\frac{s}{\sqrt{S_{xx}}}\right).$$

iv. Regression through the Origin

In addition to the pivot for the slope noted above, recall that, under normal theory assumptions, the intercept estimator, $b_0$, satisfies:

$$b_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]\right).$$

This in turn implies that the pivot

$$\frac{(b_0 - \beta_0)}{s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}} \sim t_{n-2}.$$

Therefore, a $100(1-\alpha)\%$ confidence interval for $\beta_0$ can be calculated as:

$$b_0 \pm t_{n-2}(1-\alpha/2)s(b_0) = \left(b_0 - t_{n-2}(1-\alpha/2)s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}, b_0 + t_{n-2}(1-\alpha/2)s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}\right).$$

If the value zero is contained in the above interval, then we might be willing to assume that the true regression line goes through the origin. Indeed, there may be many situations in which it makes perfect sense to restrict attention to only those lines which pass through the origin. Recall, however, that this assumption implies more than the fact that the origin is a reasonable intercept for the data, but also that the linear relationship between the two variables under study holds not only in the range of the collected data, but also all the way back to the origin. If such an extrapolation is justified (either scientifically, or because the range of the data is very close to or includes the origin) then, we might want to model our data as:

$$Y = \beta_1 x + \epsilon,$$

with $\epsilon$ again a mean-zero random variable with variance $\sigma^2$.
We can then repeat our entire discussion, with the modification that we will start our least-squares estimation from the distance function:

$$d^\star(b_1) = d(0, b_1) = \sum_{i=1}^{n}(Y_i - b_1 x_i)^2.$$

Taking derivatives shows that the least-squares estimator of the slope is now

$$b_1 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

Under the assumption that the model is correct (i.e., that the true regression line does indeed pass through the origin), $b_1$ is still unbiased. Further, assuming homoscedasticity and uncorrelatedness of the errors, the variance of this estimator is:

$$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.$$

An estimate of this quantity can be calculated by substituting $s^2$ for $\sigma^2$, where the scale estimate $s^2$ is still the $MSE$ of the model, however, this is now defined by:

$$s^2 = MSE = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - b_1 x_i)^2.$$

Notice that this estimator now has $n-1$ degrees of freedom, since we no longer need to estimate an intercept term. Unfortunately, the usual partitioning of the total sum of squares, $\sum_{i=1}^n (Y_i - \overline{Y})^2$, no longer holds. However, a similar breakdown of the quantity $\sum_{i=1}^n Y_i^2$ into the sum of the $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ and the quantity $\sum_{i=1} \hat{Y}_i^2 = b_1^2 \sum_{i=1}^n x_i^2$ does hold. This last term takes the place of the regression sum of squares and, under the usual normal theory assumptions and provided that $\beta_1 = 0$, will have a scaled chi-squared distribution with a single degree of freedom. Thus, we can test the null hypothesis, $H_0 : \beta_1 = 0$, using the $F$-statistic

$$F = \frac{b_1^2 \sum_{i=1}^n x_i^2}{s^2} \sim F_{1,n-1},$$

and we can construct confidence intervals using the pivot:

$$T = \frac{b_1 \sqrt{\sum_{i=1}^n x_i^2}}{s} \sim t_{n-1},$$

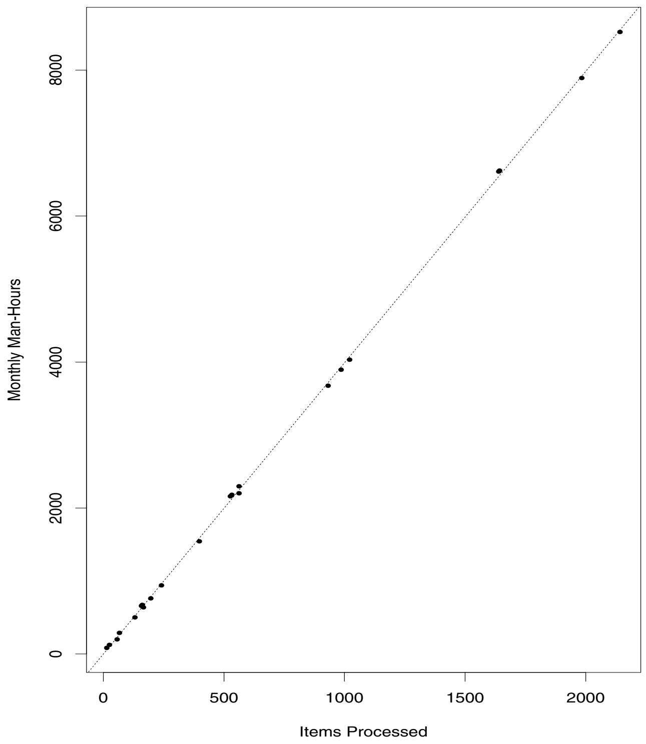
so that a $100(1-\alpha)\%$ confidence interval is given by:

$$\left( b_1 - t_{n-1}(1-\alpha/2)\frac{s}{\sqrt{\sum_{i=1}^n x_i^2}}, b_1 + t_{n-1}(1-\alpha/2)\frac{s}{\sqrt{\sum_{i=1}^n x_i^2}} \right).$$

Note that it is still the case that $T^2 = F$.

*Example 2 - Navy Manpower Data:* In this example, 22 specific US Naval institutions were sampled, and the number of man-hours worked per month were measured, as well as the total number of “items processed” (which was presumably some standard measure of product output). The data and a scatter plot are shown below:

| Items Processed | Man-hours per Month |
|---|---|
| 15 | 85 |
| 25 | 125 |
| 57 | 203 |
| 67 | 293 |
| 197 | 763 |
| 166 | 639 |
| 162 | 673 |
| 131 | 499 |
| 158 | 657 |
| 241 | 939 |
| 399 | 1546 |
| 527 | 2158 |
| 533 | 2182 |
| 563 | 2302 |
| 563 | 2202 |
| 932 | 3678 |
| 986 | 3894 |
| 1021 | 4034 |
| 1643 | 6622 |
| 1985 | 7890 |
| 1640 | 6610 |
| 2143 | 8522 |



For such data, it seems reasonable that the intercept should be the origin, since zero man-hours worked should produce zero items. Also, the plot indicates that the linear relationship is likely to hold down to the region of the origin, so a regression through the origin seems warranted in this instance. To fit such a regression in *S-Plus*, we could use the commands:

```
> navy <- as.data.frame(navy)
> names(navy)
[1] "items" "manhours"
> attach(navy)
> nvy.reg <- lsfit(items,manhours,intercept=F)
> nvy.reg$coef
        X
 3.990999
> nvy.sum <- ls.diag(nvy.reg)
> nvy.sum$std.err
          [,1]
X 0.009494946
> nvy.sum$std.dev^2
[1] 1655.349
```

However, we could also use the commands:

```
> nvy.lm <- lm(manhours ~ items - 1, data=navy)
> nvy.lm
Call:
lm(formula = manhours ~ items - 1, data = navy)

Coefficients:
    items
 3.990999

Degrees of freedom:  22 total; 21 residual
Residual standard error:  40.68597
> anova(nvy.lm)
Analysis of Variance Table

Response:  manhours

Terms added sequentially (first to last)
          Df Sum of Sq   Mean Sq  F Value Pr(F)
items      1 292460792 292460792 176676.3     0
Residuals 21     34762      1655
```

Note that the "`- 1`" in the model formula tells the `lm()` function to leave out the intercept term. Recall that the regression sum of squares can be thought of as a measure of how much of the variability of the response variable can be "explained" by the model. Suppose that we had fit a simple linear regression (i.e., including an intercept term) to this data. We could then compare the amount of variability in the response explained by this model by examining the appropriate regression sum of squares. Finally, the difference in the two regression sums of squares could be interpreted as the amount of variability explained by the inclusion of the intercept term. If this amount were small, we might feel justified in saying that the intercept term was not necessary in the model. This is an example of a pair of *nested* models, and we shall discuss this idea in more detail in later sections (see also the tutorial exercises).

v. Confidence Intervals for Prediction

Quite often, we will not be directly interested in estimating the parameters of our regression model or how well our model fits the present set of observed, but rather in predicting the response value for future observations at particular values of the predictor variable. Typically, there are two distinct types of predictions which are of interest. The first is the prediction for the mean or expected response value for observations at a particular predictor value, $x_0$; in other words, we want to estimate the value $Y(x_0) = E(Y|x_0) = \beta_0 + \beta_1 x_0$. Obviously, the simplest estimate of this quantity is just:

$$\hat{Y}(x_0) = b_0 + b_1 x_0.$$

However, we would like to know how good an estimate this is, and thus we need to provide an associated standard error. Now, since we have seen that $\overline{Y}$ and $b_1$ are uncorrelated, we have:

$$\begin{aligned} Var\{\hat{Y}(x_0)\} &= Var(b_0 + b_1 x_0) \\ &= Var\{\overline{Y} + b_1(x_0 - \overline{x})\} \\ &= Var(\overline{Y}) + (x_0 - \overline{x})^2 Var(b_1) \\ &= \sigma^2 \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}. \end{aligned}$$

[NOTE: When $x_0 = 0$, this formula reduces to $s(b_0)$ as it should.]

If we substitute our estimator, $s$, for the scale parameter, $\sigma$, we then have an estimate of the *standard error of prediction* or *standard error of the fit*, given by:

$$s\{\hat{Y}(x_0)\} = s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}.$$

Thus, assuming normal errors, we can write a $100(1-\alpha)\%$ confidence interval for the expected response $E(Y|x_0)$ as:

$$\hat{Y}(x_0) \pm t_{n-2}(1-\alpha/2)s\{\hat{Y}(x_0)\}.$$

The above interval is indeed a *confidence* interval, since it gives a plausible range for a *fixed, population quantity*. The second type of prediction we might wish to make is for the value of a single future response value, $Y_0$, at a particular value of the predictor variable, $x_0$. Note that this is a slightly different type of problem, since we want to find a likely range for a random quantity, and as such we shall refer to the resulting range as a *prediction* interval instead of a confidence interval. Obviously, the simplest estimate for the value of such an observation would again be the value $\hat{Y}(x_0) = b_0 + b_1 x_0$, since $E(Y_0|x_0) = \beta_0 + \beta_1 x_0$. However, when we try to assess the variability of this estimator now, we must take into account the fact that there are now *two* sources of variation. To see this, note that if we knew the values of the parameters $\beta_0$ and $\beta_1$ then we would know $Y(x_0)$ exactly, but the actual observed value of $Y_0$ would still differ from its exact expectation, $\beta_0 + \beta_1 x_0$, since this value is merely the center of the distribution of $Y_0$. However, we do not know the true values of the parameters, and thus we must estimate the true center of the distribution of $Y_0$ and this provides us with our second source of variability. To find out how to construct an appropriate prediction interval, we note that if $Y_0$ is a future value, it should be independent of the observed data used to estimate $b_0$ and $b_1$. Thus, $Y_0$ and $\hat{Y}(x_0) = b_0 + b_1 x_0$ will be independent random variables. So,

$$Var\{Y_0 - \hat{Y}(x_0)\} = Var(Y_0) + Var\{\hat{Y}(x_0)\} = \sigma^2 + \sigma^2\left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}.$$

Now, since $E\{Y_0 - \hat{Y}(x_0)\} = 0$, the standard normal theory assumptions show that

$$\frac{Y_0 - \hat{Y}(x_0)}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

This in turn implies that:

$$\begin{aligned}
& Pr\left\{ -t_{n-2}(1-\alpha/2) < \frac{Y_0 - \hat{Y}(x_0)}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}}} < t_{n-2}(1-\alpha/2)\right\} = 1-\alpha \\
\Longrightarrow \quad & Pr\left\{|Y_0 - \hat{Y}(x_0)| < t_{n-2}(1-\alpha/2)s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\right\} = 1-\alpha \\
\Longrightarrow \quad & Pr\left\{Y_0 \in \hat{Y}(x_0) \pm t_{n-2}(1-\alpha/2)s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\right\} = 1-\alpha,
\end{aligned}$$

which provides us with our $100(1-\alpha)\%$ prediction interval for a single future response observation associated with a predictor value of $x_0$.

*Example 1 (cont'd) - Protein in Pregnancy Data:* Previously we saw that for the protein in pregnancy data, the least-squares parameter estimates were $b_0 = 0.2017377$ and $b_1 = 0.02284426$, and the estimate of the regression scale was $s = 0.1150781$. In addition, we can calculate $\overline{x} = 24$ and $S_{xx} = 1220$. Suppose we want to predict the expected protein level of a pregnant woman who has just come to full term, which is typically 38 weeks. Our estimate of the expected protein concentration for such women would be $\hat{Y}(38) = b_0 + 38b_1 = 1.06982$, and the standard error is

$$s\{\hat{Y}(38)\} = s\sqrt{\frac{1}{n} + \frac{(38-\overline{x})^2}{S_{xx}}} = 0.1150781\sqrt{\frac{1}{19} + \frac{(38-24)^2}{1220}} = 0.05314656.$$

Thus, a 95% confidence interval for $E(Y|x_0 = 38)$ would be $1.06982 \pm 0.05314656t_{17}(0.975) = (0.9576902, 1.181949)$ since $t_{17}(0.975) = 2.109816$. The *S-Plus* commands necessary to calculate this interval are provided below, assuming that the appropriate model has already been fit and its output stored in the object `protpreg.lm` (see previous discussion of this example):

```
> gestation <- 38
> pr.pred <- predict(protpreg.lm,as.data.frame(gestation),se.fit=T)
> pr.pred$fit
        1
 1.06982
> pr.pred$se.fit
          1
 0.05314656
> qt(0.975,17)
[1] 2.109816
```

Now, suppose that we have just taken a protein sample from a pregnant woman who is just about to give birth (i.e., she is at 38 weeks gestation), what would we guess for the value of her protein level? Here, we are asked for a prediction interval and not a confidence interval. So, a 95% prediction interval in this case would be:

$$\begin{aligned}\hat{Y}(x_0) \pm t_{n-2}(1-\alpha/2)s\sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}} &= \hat{Y}(x_0) \pm t_{n-2}(1-\alpha/2)\sqrt{s^2 + [s\{\hat{Y}(x_0)\}]^2} \\ &= 1.06982 \pm 2.109816\sqrt{(0.1150781)^2 + (0.05314656)^2} \\ &= 1.06982 \pm 0.2674354 \\ \Longrightarrow \quad & (0.8023846, 1.337255).\end{aligned}$$

Note that the prediction interval is much wider than the confidence interval.

## IV. Model Diagnostics

Most of the preceding inferential theory regarding simple linear regression relied heavily on our basic assumptions:

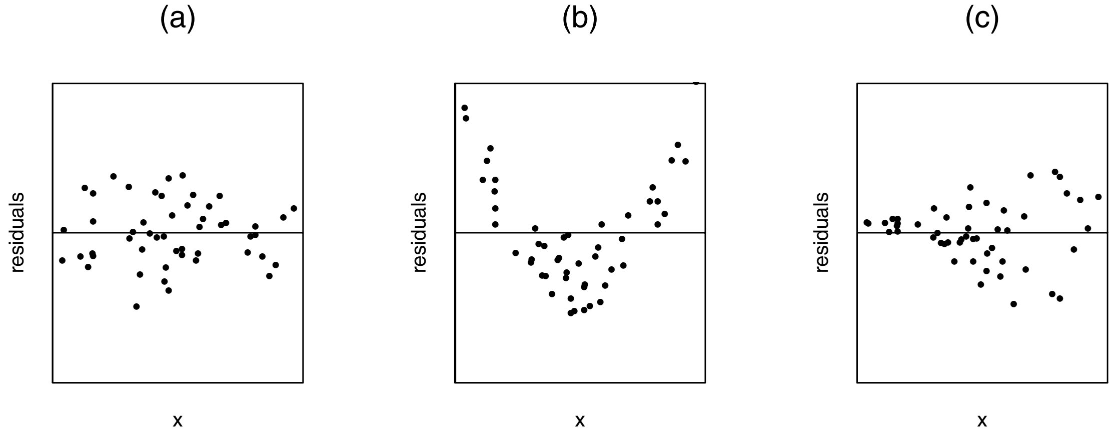
- The underlying relationship between the variables is linear;
- The error structure is homoscedastic; and,
- The errors are normally distributed.

Each of these assumptions can be investigated within the dataset under study to determine if they can be reasonably believed to hold. Specifically, we will undertake a detailed investigation of the regression residuals, $e_i = Y_i - \hat{Y}_i$, in an effort to determine if their observed behavior is consistent with what would be expected of them under the basic assumptions of the simple linear model.

In general, the residuals ought to behave in the same way that the $\epsilon_i$'s would behave; namely, as observations from a mean-zero multivariate normal distribution. However, some concession must be made to the fact that the residuals are only estimates of the true errors, and as such will depart somewhat from such behavior [e.g., we have already seen that the residuals will be slightly correlated, since $Var(e) = \sigma^2(I - H)$]. We will see more detailed methods for taking these differences into account in our subsequent investigations of models with several predictors variables. Nonetheless, no modelling exercise should be considered complete without a strong diagnostic study, as the results from a model whose underlying assumptions are not supported by the observed data are at best highly suspect, and generally can lead to highly erroneous conclusions. In particular, even in situations where plots of the data are not possible (e.g., situations with more than two predictor variables), plots of the subsequent residuals from the model are always available and should be examined thoroughly before any conclusions regarding the model output are drawn.

i. Residuals and Data Transformations

Severe departures of the probabilistic structure of the underlying $\epsilon_i$'s from the assumptions can often be ascertained by a simple plot of the observed residuals versus the predictor values. Below are three depictions of canonical residual plots:



Plot $(a)$ is a depiction of what one would expect to see if the assumptions of linearity and homoscedasticity are indeed satisfied. That is, under the assumptions of the model, the residuals should tend for form a basically rectangular (or more precisely, a slightly elliptical) pattern around the zero-line, having basically uniform spread and a nice random scatter both above and below. The fact that the plot is basically rectangular (elliptical) indicates that homoscedasticity is a reasonable assumption, since the uniform scale of the errors, $\epsilon_i$, should be reflected in the uniform scatter of the observed residuals. [NOTE: The reason that we might expect an elliptical, rather than rectangular, plot is that, while the scale of the residuals should be the same throughout the range of the predictor variable, the fact that there are more datapoints in the center of the predictor range indicates that there will naturally appear to be a larger spread in the observed residuals. This is simply because more data values means more chance for data points to stray a bit farther from their mean. Statistically speaking, the idea is that if we have only a few observations from a normal distribution with mean 0 and variance $\sigma^2$, then we would expect almost all the values to be within $2\sigma$ of 0; however, if we have a larger number of observations from the same distribution, we would expect to see the occasional value rather far from 0. In addition, the data points associated with predictor values on the ends of the predictor range tend to have high *influence* which will cause their observed residual values to be smaller than expected (see the following section on Influential Points).]

Plot ($b$) shows a typical situation in which the true relationship between the two variables under study is not a linear one. One of the first things to look for in a residual plot is some sort of definite pattern. If the linear model is the correct one, then we have noted above that the residuals should look like pure random scatter. Thus, any noticable pattern or structure in the residuals is an indication that the underlying assumptions of the simple linear regression model do not hold. In graph ($b$) above, it seems quite clear that there is a quadratic structure to the relationship which needs to be taken into account. One way to proceed would be to try and fit the more complicated model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon,$$

and we will see how to fit such a *polynomial regression* shortly. Another potential approach might be to *transform* the data by letting $Z = f(Y)$ for some chosen function $f$ and/or $w = g(x)$ for some chosen function $g$, and then fit the model:

$$Z = \beta_0 + \beta_1 w + \epsilon.$$

Some very common choices for the functions $f$ and $g$ are the natural logarithm function, and the square root function (of course, the data must be all strictly positive to use such functions directly, however, a simple location shift can be included to facilitate this). The choice of the functions $f$ and $g$ can often be based on scientific or mathematical reasoning. For example, in the dataset relating the volume of a cherry tree to its height and diameter, we were naturally led to the logarithmic transformation by geometric considerations, and we arrived at the sensible model:

$$\ln(\text{Volume}) = \beta_0 + \beta_1 \ln(\text{Height}) + \beta_2 \ln(\text{Diameter}) + \epsilon.$$

While for the Forbes' data, we had scientific theory to lead us to the model:

$$\ln(\text{Pressure}) = \beta_0 + \beta_1 \text{Boiling Point} + \epsilon.$$

In other situations, we might simply have to try out several possible transformations and see which one yields the most suitable residual plot. However, we must be aware that transformations have consequences. First, the units of the parameters, and therefore their interpretations, are changed by transformation. This means that the interpretability of the model is not as straightforward as it might otherwise have been. For example, if we use a logarithmic transformation on our data, then any predictions which we make using this transformed model will need to be exponentiated in order to put the results on the proper scale. Specifically, suppose we fit the model:

$$Z = \beta_0 + \beta_1 w + \epsilon,$$

where $w = \ln(x)$ and $Z = \sqrt{Y}$. If we want to predict the value of the original response variable $Y$ for a particular value, $x_0$, of the original predictor, we must note that:

$$\begin{aligned} w_0 &= \ln(x_0) \\ \hat{Y}(x_0) &= \{\hat{Z}(w_0)\}^2 = \{\beta_0 + \beta_1 w_0\}^2 = \left\{\beta_0 + \beta_1 \ln(x_0)\right\}^2. \end{aligned}$$

Moreover, if plot $(b)$ above represents the residual plot for a simple linear regression of the response variable, $Y$, on the predictor variable, $x$, and we transform the variables so that $Z = \sqrt{Y}$ and then fit the model:

$$Z = \beta_0 + \beta_1 x + \epsilon,$$

The resulting residual plot would tend to look like a mirror image of plot $(c)$, in other words, it would have a "funnel" shape with the "wide end" to the left rather than the right. This is because a transformation of the response data (i.e., of the model structure) inherently affects the associated error structure. The idea here is that the square root of a very small number (i.e., much less than one) is only a moderately small number, while the square roots of moderate numbers remain moderate. Thus, taking the square root of the response data can change the nature of the scale of the data in a differential way across the range of the predictors, and thus possibly induce heteroscedasticity in the resultant transformed model.

However, this very feature of transformations can also be seen as a positive, since if we have a data set whose residual plot looks like plot $(c)$ above, it may be that a transformation of the response will fix the apparent heteroscedasticity. Of course, it may do this at the expense of the linearity of the data. Remarkably, however, it is often the case that a single transformation may solve both a nonlinearity and a heteroscedasticity problem simultaneously. In any case, it is highly important to examine the resulting residual plots of any models we choose to fit to the data in order to verify that the underlying assumptions are plausibly valid.

As a final note, we point out that it is possible to handle heteroscedasticity without data transformation, however, the required techniques (which amount to specifying how the variability changes over the range of the predictor) are, for the most part, outside the scope of this course.

ii. Normal Probability Plots

The examination of a residual plot was shown to be an extremely useful device in investigating the linearity and homoscedasticity of a dataset. It can also provide some amount of insight into the normality of the data. The idea is that if the $\epsilon_i$'s really come from a multivariate normal distribution, then we would expect to observe almost all their values within $2\sigma$ of zero. Thus, if we see a large number of residuals with values either larger than $2s$ or smaller than $-2s$ we might become suspicious as to whether the errors were truly normally distributed. There is a slight problem here, which results from the fact that, unlike the $\epsilon_i$'s, the variance of the $e_i$'s is given by $Var(e) = \sigma^2(I - H)$ and thus, technically, the residuals are correlated with each other, as well as each having slighty different individual variances. We will take up this point more fully in our discussion of multiple regression, where we will attempt to account for this problem using the so-called *Studentized* residuals. For now, though, it suffices that we can examine a residual plot and take the existence of several large residual values as evidence against the normality of the underlying errors (see the section below on Outliers).

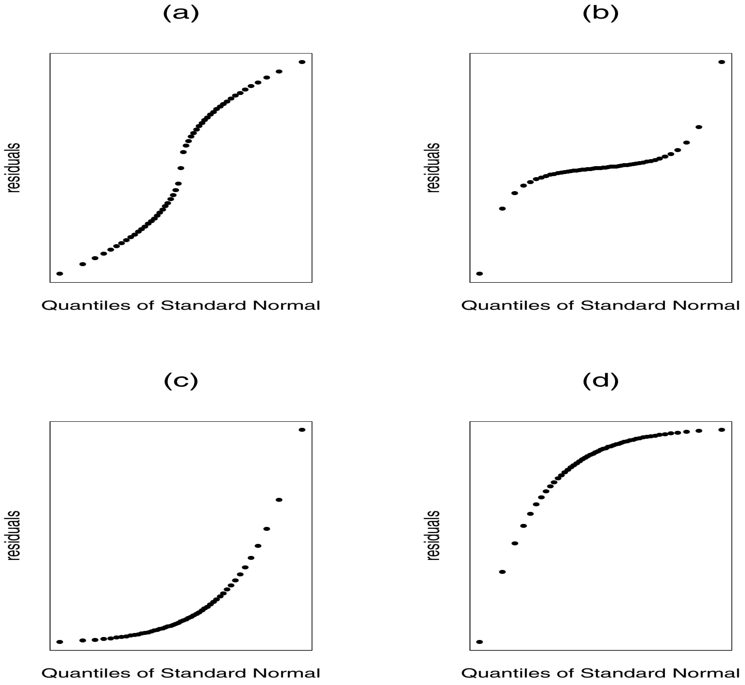
A more useful method of examining the normality of the data is through the use of a *normal q-q plot*. The idea behind such a plot is rather straightforward. A normal q-q plot simply displays the ordered observed residuals versus what we would have expected for these values if the errors in the data were indeed normally distributed. Thus, if a normal q-q plot appears as an approximate straight line, we can conclude that the errors are likely to be normally distributed.

First, we must calculate what we would expect the residuals to be if they were indeed from a normal distribution. It turns out that, if we write $e_{[i]}$ for the $i^{\text{th}}$ smallest observed residual (i.e., $e_{[1]}$ being the smallest, meaning the most negative, residual, $e_{[2]}$ the next smallest and so on up to $e_{[n]}$ being the largest residual) then

$$E\left(\frac{e_{[i]}}{s}\right) \approx \Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right),$$

the $\left(\frac{i-0.375}{n+0.25}\right)$-quantile of the Standard Normal Distribution. These expectations are generally referred to as the *normal scores* for the data, and if we plot the normal scores versus their associated observed residual value, then we have a normal q-q plot. In *S-Plus*, the function `qqnorm()` takes a vector of residuals as its input and produces a normal q-q plot. As discussed above, technically, we need to consider the fact that the observed residuals have an inherently unequal scale, and thus we should be using the Studentized residuals mentioned above, however, we will withhold discussion of this until later.

As pointed out, if the data are truly from a normal distribution, the q-q plot should look approximately like a straight line, and the relationship between the normal scores and the expectations of the residuals indicates that this line should have an intercept of zero and an approximate slope of $s$, the estimated residual scale. There are typically four basic departures from a straight line which are common:

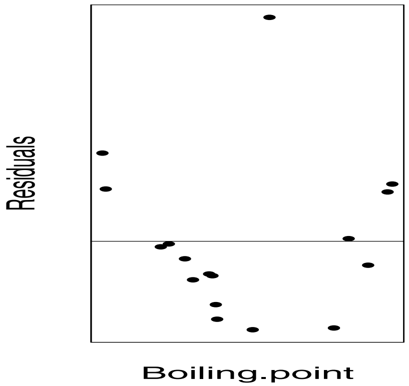


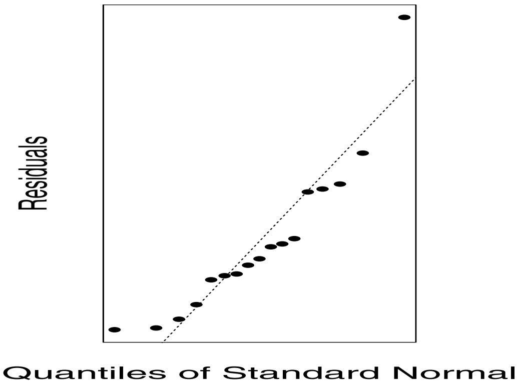
Plot ($a$) depicts the situation in which the residuals are “light-tailed”. In other words, they are less spread out than would have been expected if they were truly normally distributed. This is seen from the characteristic “S”-shape, showing that the largest and smallest observed residuals tend to be closer to zero than their associated normal scores. Plot ($b$) depicts the opposite situation in which the residuals are “heavy-tailed”. In other words, they are prone to extreme values, so that they are more spread out than would have been expected under normality. This is seen from the characteristic reverse “S”-shape, showing that the largest and smallest observed residuals tend to be farther from zero than their associated normal scores. Plot ($c$) depicts a situation in which the residuals are positively skewed. This is seen from the characteristic concave-up shape, indicating that the largest residuals tend to be larger than would have been expected. Similarly, plot ($d$) depicts a situation in which the residuals are negatively skewed. This is shown by the characteristic concave-down shape, indicating that the smallest residuals are even smaller (i.e., farther negative) than would have been expected.

If we have a dataset where the normal q-q plot shows that the assumption of normality is not valid, we have several options. It is possible that a data transformation may make the data more normally distributed. Of course, we must keep in mind that this transformation may also affect the linearity and homoscedasticity of the data. Another option is to move to a maximum likelihood approach using some other distributional assumption. Of course, this requires us to have some other distribution in mind, and it is rare that we have good reason to believe in specific distributions other than the normal (though such situations do exist, and that is the starting point for the study of *generalized linear models*, but that is outside the scope of this course). Another option is to switch to so-called *robust* methods, for which the assumption of normality is less critical. Again, the details of this topic are outside the scope of this course. However, the basic notions are centered around developing methods which use distance measurements which are not as adversely affected by large residuals as is the sum of squares measure.

*Example 3 - The Forbes' Data:* Suppose that we were to examine a simple linear regression for the Forbes' Data, using a linear model relating pressure to boiling point, and we examined the residual plot and a normal q-q plot:

```
> forbes.df <- as.data.frame(forbes)
> attach(forbes.df)
> names(forbes.df)
[1] "Boiling.point" "Pressure"
> Residuals <- residuals(lm(Pressure ~ Boiling.point))
> s2 <- sum(Residuals^2)/(length(Residuals)-2)
> plot(Boiling.point,Residuals)
> abline(0,0)
> qqnorm(Residuals)
> abline(0,sqrt(s2),lty=2)
```

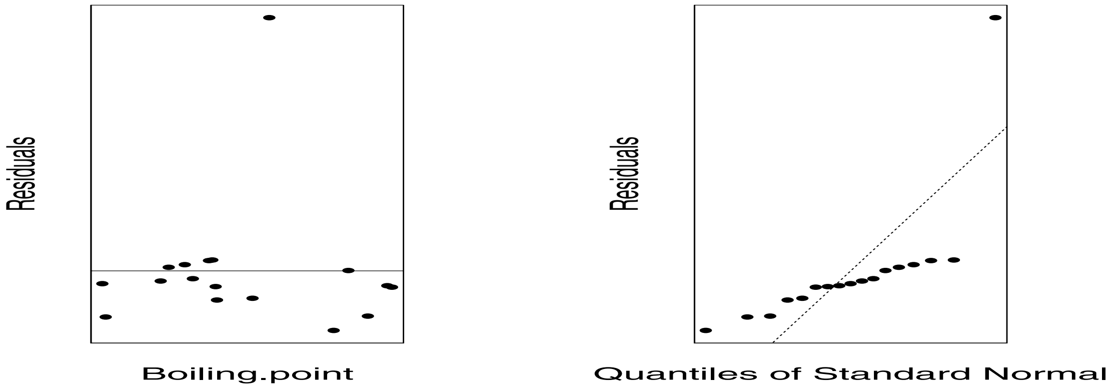




The residual plot clearly indicates that a strictly linear relationship is not very appropriate. In addition, the q-q plot indicates that the residuals from such a model do not appear to be from a normal distribution, but instead are somewhat positively skewed (though, admittedly, this is a borderline decision). Taking logarithms has the effect of dramatically reducing very large observations while only moderately reducing smaller values, and is thus an effective way of reducing positive skew. In addition, it has the effect of "straightening" some curvature.

Thus, if we now fit a linear model relating the logarithm of pressure to the boiling point, and examining the associated residual and normal q-q plots, we have:

```
> Residuals <- residuals(lm(log(Pressure) ~ Boiling.point))
> s2 <- sum(Residuals^2)/(length(Residuals)-2)
> plot(Boiling.point,Residuals)
> abline(0,0)
> qqnorm(Residuals)
> abline(0,sqrt(s2),lty=2)
```



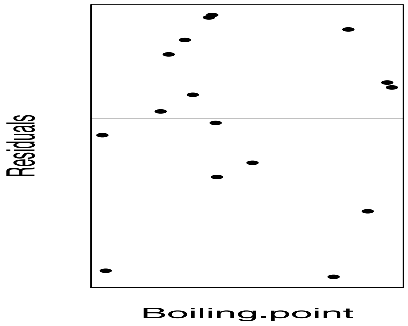
These plots now show that there is an obviously discrepant point, but that otherwise this linear model appears to be an acceptable one (i.e., the residuals other than the very large one are nicely scattered and the normal q-q plot looks linear except for the strange point, which is adversely affecting the measure of scale, $s$, which is why the dotted line on this plot does not follow the bulk of the data). This leads us to the discussion of outliers and influential points.

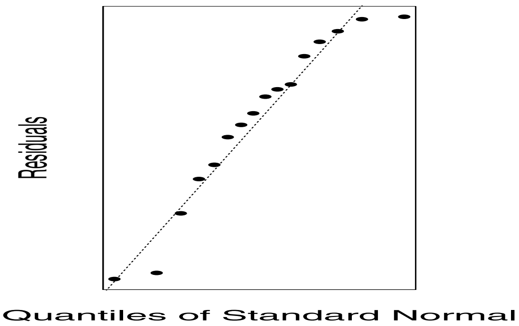
iii. Outliers and Influential Points.

Another important use of the residual plot is to spot potentially aberrant data points. *Outliers*, as they are called, are points with associated residuals which are drastically vertically removed from the main mass of the residuals. Note that this is not the same as over-dispersion of the residuals which leads to a "heavy-tailed" looking q-q plot. Here we are talking about a single, or only a few, data points which are markedly different from the rest. It is important to identify such points, because they can have a dramatic effect on our estimate of scale, $s^2$, and thus we can lose precision in our tests and confidence intervals. Generally, we will want to remove outliers from our dataset and refit the model. Of course, we have to remember that we should not needlessly throw away information without good reason, and thus an investigation should be made to see if there is some explanation for the discrepancy of the outlying points. Often, these points will lead us to important realizations about the structure of the population from which we have drawn our data. On the other hand, they can simply be the product of a data entry error. Whatever the reason though, outliers should certainly be noted, and their removal should be contemplated.

*Example 3 (cont'd) - Forbes' Data:* If we remove the obvious outlier from the Forbes' data and re-fit the model relating the logaritm of pressure to the boiling point, the resulting residual and normal q-q plots are:

```
> Residuals <- residuals(lm(log(Pressure[-12]) ~ Boiling.point[-12]))
> s2 <- sum(Residuals^2)/(length(Residuals)-2)
> plot(Boiling.point[-12],Residuals,xlab="Boiling.point")
> abline(0,0)
> qqnorm(Residuals)
> abline(0,sqrt(s2),lty=2)
```





We now see that the residual plot and the normal q-q plot show that the assumptions of the linear model are plausible (though, perhaps the q-q plot shows that the logarithmic transformation was a bit over-aggressive in attacking the positive skew in the residuals, since it now appears that the residuals have a slight negative skew, but it is not really severe enough to be a real cause for concern). The resulting fitted relationship is for the reduced dataset is:

$$\ln(\text{Pressure}) = -0.9518 + 0.0205\text{Boiling Point}$$

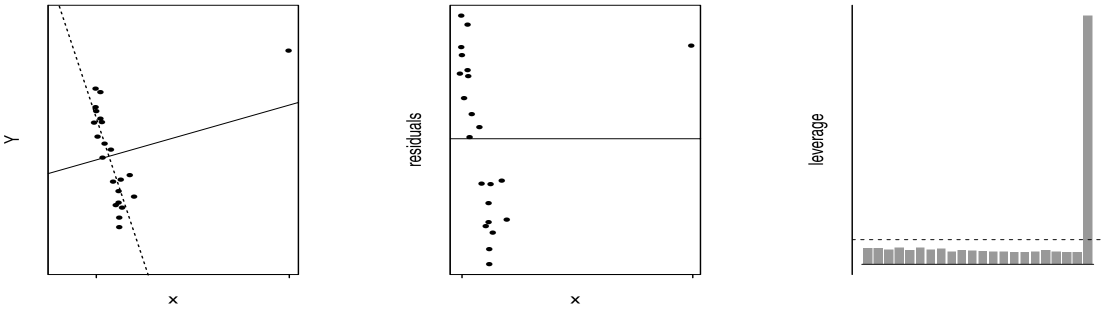
or,

$$\text{Pressure} = e^{-0.9518+0.0205\text{Boiling Point}} = 0.3861e^{0.0205\text{Boiling Point}}.$$

As pointed out, outliers are points with extreme residual values, and are thus characterized by being far from the main body of the data in the vertical direction. Points which are far from the main body of the data in the horizontal direction can also cause problems. Such points have a potential for high *influence* on the fitted regression line. The concept of influence is a somewhat nebulous one. Basically, however, the idea is that a point is highly influential if its removal from the dataset causes a dramatic change in the estimated parameters of the regression line. One useful way of flagging the potentially influential data points is through the use of the *leverages*, $h_{ii}$, which, as we saw, are the diagonal elements of the hat matrix, $H$. Since the hat matrix involved only the values of the predictor variable, we see that this is a measure of how far from the main body of the data each point is in the horizontal (or predictor) direction. It turns out (see tutorial exercises) that the sum of all the leverages in a simple linear regression is equal to 2:

$$\sum_{i=1}^{n} h_{ii} = 2.$$

If all the data points had the same leverage, each of the $h_{ii}$'s should be equal to $2/n$. So, any point which exceeds this value, and more particularly, any point which exceeds twice this value is a potentially influential point. Of course, in order to see whether such points are truly influential, we must see how the fitted regression line would change once the point is deleted from the dataset. As with outliers, all influential points should be carefully examined. As a final note, we point out that, as the following picture shows, points with high influence may appear on the data plot as an "outlying point", however, their influence on the fitted line causes their residuals to actually be relatively small. The following plot illustrates (somewhat over-dramatically) what can happen with a highly influential point:



The first plot demonstrates how the influential point has "dragged" the fitted line up towards it (the dotted line indicates the fitted regression line for the dataset which ignores the influential point). Note that the residual associated with the influential point is not exceptionally large compared to the other residuals (of course, in this simple example there is a distinct pattern in the residuals which we would have easily found). Finally, the leverages clearly show how much potential for high influence the last point has.

To summarize the general points of this section:

- Outliers are points with extreme residuals;
- High leverage points are potentially (though not necessarily) influential; and
- Influential points are generally not outliers in the sense of having large residuals.

The use of residuals and leverages will be discussed further in the framework of multiple regression where they take on an even more important role.

## V. Random Predictors

In all of the above discussion, we have assumed that the values of the predictor variable, $x_i$, have been fixed. Of course, in many situations, we will not truly have control over the predictor values. Suppose that we now assume that the pairs $(X_i, Y_i)$ are randomly sampled from a population in such a way that they have a *bivariate normal* distribution with means and variances:

$$E(X) = \mu_x; \qquad E(Y) = \mu_y; \qquad Var(X) = \sigma_x^2; \qquad Var(Y) = \sigma_y^2; \qquad Cov(X,Y) = \rho\sigma_x\sigma_y.$$

If this is the case, then it can be shown that the conditional distribution of $Y$ given $X$ is also normal with:

$$E(Y|X) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(X - \mu_x) = \beta_0 + \beta_1 x; \qquad Var(Y|X) = \sigma_y^2(1-\rho^2) = \sigma^2.$$

So, we can see that the values of $Y$ and $X$ are linearly related with

$$\begin{aligned}\beta_0 &= \mu_y - \beta_1\mu_x \\ \beta_1 &= \rho\frac{\sigma_y}{\sigma_x},\end{aligned}$$

and the expression for the variance shows that the conditional variability of $Y$ does not depend on $X$, so that the data will be homoscedastic. In this situation, interest generally centers on the parameter $\rho$, though the above identities show that this is in some sense equivalent to interest in $\beta_1$. Further, we can see that

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_y^2} = \beta_1^2 \frac{\sigma_x^2}{\sigma_y^2}.$$

A comparison of these formulae with those for the fixed predictor case shows the strong similarity between the two situations. Note that since $\rho$ is a correlation, it must lie between -1 and 1, and can only equal these values if $\sigma^2 = 0$, i.e., if their is perfect linear association between the random variables $X$ and $Y$. Thus, $\rho$ measures the degree of linear association between the two random variables.

Estimation of the parameters proceeds along nearly identical lines to the previous fixed predictor case, and the formulae for the estimates are indeed identical, along with the estimate for $\rho$, which is $r$, the sample correlation coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = b_1\sqrt{\frac{S_{xx}}{S_{yy}}},$$

where $S_{yy} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = SST$.

As mentioned above, typically interest centers on testing the hypotheses

$$H_0 : \rho = 0 \qquad \text{versus} \qquad H_A : \rho \neq 0$$

which amounts to testing whether there is any linear association between the two variables. It turns out that, if $H_0$ is true, then the test statistic

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

has a Student's $t$-distribution with $n-2$ degrees of freedom, and thus we can test the null hypothesis by comparing the observed value of $T$ to the appropriate $t$-quantiles, $t_{n-2}(1-\alpha/2)$. We have noted previously that all tests of the significance of a regression are actually identical, and we now show that:

$$\begin{aligned}
T &= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \\
&= \frac{\left(b_1\sqrt{\frac{S_{xx}}{S_{yy}}}\right)\sqrt{n-2}}{\sqrt{1-\frac{SSR}{SST}}} \\
&= \frac{b_1\sqrt{S_{xx}}}{\sqrt{\frac{SST-SSR}{n-2}}} \\
&= \frac{b_1\sqrt{S_{xx}}}{\sqrt{\frac{SSE}{n-2}}} \\
&= \frac{b_1\sqrt{S_{xx}}}{\sqrt{MSE}} \\
&= \frac{b_1\sqrt{S_{xx}}}{s} \\
&= \frac{b_1}{s(b_1)}.
\end{aligned}$$

Thus, this test is the same as the $t$-test and the $F$-test for the null hypothesis that $\beta_1 = 0$. (Note that $\beta_1 = 0$ if and only if $\rho = 0$.)
Of course, the reason for the similarity between the two approaches derives from the fact that even in the case of random predictors, we focus on the conditional distribution of the response given the predictor values. This is a perfectly reasonable approach if it is truly the variation in the response that we are trying to explain, since the way in which the values of the predictor variables are arrived at is then secondary information. Because of this, we shall always make the assumption that the $x_i$'s are fixed.

### VI. Summary

So, what have we learned so far? The general procedure for examining the relationship between two variables proceeds as follows:

1. Examine a plot of the data to see if a linear association seems a plausible explanation of the scatterplot;
2. Using scientific or other background information, transform the data appropriately;
3. Fit a linear regression model to the two variables using least-squares estimates;
4. Test the significance of the regression and make required predictions;
5. Use residual plots and normal q-q plots to examine the plausibility of the basic assumptions of the model;
6. If necessary (based on the plots of step 5) transform the data again (or move to a more complicated model structure);
7. Re-fit the regression and again examine the residuals thoroughly;
8. When satisfactory residual analyses have been reached, re-test and re-predict as required (remembering to transform back to appropriate scales if necessary).

We will now take up the point mentioned parenthetically in step 6, and examine more complicated, though still linear, models, where we will allow for the possibility of more than one predictor variable.

STAT2008 – REGRESSION MODELLING
LECTURE NOTES - CHAPTER 2: MULTIPLE LINEAR REGRESSION

## I. Introduction

Typically, a response variable of interest will be associated with the effects of several predictor variables. Even in the case where there is only one predictor of prime consideration, there will generally be several others which should obviously be taken into account. Indeed, most often a set of predictors will break itself naturally into two groups: those variables which have an accepted relationship with the response variable and for which any analysis must therefore be "adjusted"; and, those variables with a suspected relationship which the current analysis is attempting to confirm or refute. For example, in any study of the potential relationship between low infant birthweight and fetal exposure to passive cigarette smoke, we must certainly take into account the age of the mother of each infant in the study. Similarly, if we are examining the effect of a food additive on the amount and quality of marketable beef produced per head of cattle, we must take into account the breed, age and other factors associated with each animal included in the study.
If we measure a response variable, $Y$, and $k$ predictor values, $x_1, x_2 \ldots, x_k$, on $n$ individual "items" or "subjects", then our data will consist of the $n$ ordered $(k+1)$-tuples:

$$(Y_i, x_{1i}, x_{2i}, \ldots, x_{ki}) \qquad i = 1, \ldots, n$$

The easiest and most straightforward way to accomodate more than one predictor in a model is to simply extend the "linear" structure we began with in the simple linear regression setting:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i$$

Note that we must now impose the condition that $n > k$ to ensure that there is enough data to estimate all the parameters (actually, this condition was trivially required even in the simple linear regression case, since we clearly could not fit a line to a scatterplot with only a single data point). While this new model may look more complicated (and indeed, it does provide a great deal of flexibility as we shall see), much of our work has already been done for us once we notice that we can write this model as

$$Y = X\beta + \epsilon,$$

where $\beta$ is now a column vector of length $k + 1 = p$,

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

and $X$ is the $n \times p$ design matrix:

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}$$

the initial column of ones again indicating that an "intercept" (i.e., the constant term $\beta_0$) has been included in the model.

### II. Model Assumptions

As before, we shall consider the values of the predictor variables, $x_{ij}$, to be fixed and measured without error (either by design or through arguments of conditioning). Similarly, we shall assume that the random error vector, $\epsilon$, satisfies:

$$E(\epsilon) = 0 \qquad \text{and} \qquad Var(\epsilon) = \sigma^2 I.$$

That is, we assume uncorrelated and homoscedastic errors. Indeed, as we did in the case of simple linear regression, we will generally assume that the $\epsilon_i$'s are normally distributed in order to make precise inferential statements.

Lastly, we assume that the underlying true relationship between the response and the predictors is a linear one. By this we mean "linear in the parameters". As we saw in simple linear regression, it was often useful to transform the response and/or the predictor and this did not change the underlying "linearity" of our model structure (though, it inevitably changed the "linearity" of our resulting picture!). So, for example, the models:

$$\begin{aligned} Y &= \beta_0 + \beta_1 x + \beta_2 x^2 \\ Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2, \end{aligned}$$

are both linear, and their associated design matrices are:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}; \qquad \text{and} \qquad X = \begin{pmatrix} 1 & x_{11} & x_{21} & x_{11} \cdot x_{21} \\ 1 & x_{12} & x_{22} & x_{12} \cdot x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{1n} \cdot x_{2n} \end{pmatrix},$$

respectively. These two models are examples of what is often referred to as a *polynomial regression*, since the model has the structure of a polynomial in the predictor variables. Occasionally, polynomial regression is presented as a separate topic, but we can see that it is simply a special case of the overall multiple linear model.

Even some models which do not at first appear to be linear, are "linearizable". For example, the cherry tree data demonstrated how the model:

$$Y = \beta_0 x_1^{\beta_1} x_2^{\beta_2}$$

could be written as:

$$\ln(Y) = \ln(\beta_0) + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) = \beta_0' + \beta_1 \ln(x_1) + \beta_2 \ln(x_2).$$

Similarly, the model

$$Y = \frac{\beta_0}{1 + \beta_1 x_1 + \beta_2 x_2}$$

can be linearized by taking reciprocals and writing the new model as

$$\frac{1}{Y} = \frac{1}{\beta_0} + \frac{\beta_1}{\beta_0} x_1 + \frac{\beta_2}{\beta_0} x_2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2.$$

Once the $\gamma$ parameters have been estimated, the relationships

$$\beta_0 = \frac{1}{\gamma_0}; \qquad \beta_1 = \frac{\gamma_1}{\gamma_0}; \qquad \beta_2 = \frac{\gamma_2}{\gamma_0}$$

can be used to find corresponding estimates of the $\beta$ parameters.

It must be remembered, however, that such linearization transformations have an inherent effect on the nature of the error variable. As we noted in simple linear regression, this can often have the desirable effect of attaining homoscedasticity or normality. However, such fortunes should not be relied upon and care should be taken. We will shortly discuss diagnostic procedures to examine these aspects of our model in detail.

There are, of course, many models which are not linear and not even linearizable. For example, the model

$$Y = \beta_0 + \beta_1 x_1^{\beta_2} + \beta_3 x_2^{\beta_4}.$$

No amount of tranformation will turn this into a model which is linear in its parameters (despite that fact that the model $Y = \beta_1 x_1^{\beta_2}$ was seen to be easily linearized using logarithms). For these models, we must employ non-linear regression techniques which are beyond the scope of this course.

### III. Interpretation of Model Parameters

The $\beta$ parameters in the multiple regression model are obviously very similar in nature to the corresponding parameter in the simple regression model, in that they occupy positions corresponding to "slopes" of some kind. However, the added complexity of the multiple regression model requires us to be a bit more careful in our interpretations. If we were to naively keep the interpretation that the $\beta_i$'s represented the change in the response per a unit change in the value of the associated predictor, then we should reasonably expect that the estimates we arrived at for, say, $\beta_1$ in the two models:

$$Y = \beta_0 + \beta_1 x_1$$
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

would be the same. However, as we shall see, this is rarely ever the case. And indeed, a closer inspection indicates why this is so. In the multiple linear regression setting, the $\beta_i$'s are often referred to as *partial* regression coefficients, since their proper interpretation is the change in the response per a unit change in the associated predictor *with all other predictors held constant*. This last clause is the crux of why the $\beta_i$'s are called partial regression coefficients and also why we arrive at different estimates for the "slope" associated with a predictor in the simple versus multiple regression settings. The notion is that in theory it is easy to conceive of simply changing the value of one predictor variable without changing the values of the others, but in practice, the predictor variables themselves may be interrelated in such a way that no member of the population exists with such a set of predictor values. Thus, in the simple regression case, when we say "a unit change in the predictor", we may implicitly be including an associated change in another variable which was not included in the model. Once this variable is included in the model, then the new "slope" parameter becomes a somewhat more abstract entity which may not really be measurable in the population. We will investigate these notions further when we discuss the problem of multicollinearity later.

Another difficulty with the interpretation of the multiple regression model is that pictures become much more difficult to draw, and indeed are nearly impossible when $p > 3$. This means that we must simply take the entire regression equation as just an algebraic description of the pieces which make up the value of the response variable. Also, as a final comment, we point out that the hazards of the interpretation of the intercept are the same as they were for the case of simple linear regression, and it is generally wise to simply consider $\beta_0$ as a structural and not a directly interpretable component of the model.

### IV. Least-Squares Estimation

i. The regression coefficients

We will again use the least-squares method to estimate the $\beta_i$'s. To do so, we define the distance function

$$d(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n}(Y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})^2.$$

The least-squares estimators are those values which minimize this distance function. In other words, they are the values such that the total vertical distance from the observed data to the fitted regression surface is minimized. (As a structural note, we point out that keeping $b_2 = b_3 = \ldots = b_k = 0$ in the distance function and minimizing the distance over the $b_0$ and $b_1$ will of course yield the simple linear regression least-squares estimates. The added freedom of allowing the other estimates to range freely shows that we can now minimize the distance much more effectively, and thus will naturally arrive at a different $b_0$ and $b_1$ then we did in the simple regression case.)

In order to actually calculate the least-squares estimates, we could proceed directly by differentiating and then solving the resulting set of $p$ equations. However, an appeal to matrix notation shows that we may write the distance function as:

$$d(b) = (Y - Xb)^T(Y - Xb),$$

where $b$ is now a vector of length $p$. From here we can simply mimic our development for the simple regression case and note that

$$\frac{\partial d}{\partial b} = -2X^T Y + 2(X^T X)b = 0 \qquad \Longrightarrow \qquad b = (X^T X)^{-1} X^T Y.$$

As a result, we also have the corresponding definitions for both the fitted values, $\hat{Y} = Xb = X(X^T X)^{-1} X^T Y = HY$, and the residuals $e = Y - \hat{Y} = Y - HY = (I - H)Y$.

This direct analogy with our development in the simple linear model allows us to immediately see that the least-squares estimators are unbiased and have variance covariance matrix

$$Var(b) = \sigma^2 (X^T X)^{-1}.$$

In fact, under the normal error assumption, we know that

$$b \sim N\{\beta, \sigma^2 (X^T X)^{-1}\}.$$

So, once we have an estimator for $\sigma$, we can easily construct tests and confidence intervals for the $\beta_i$'s, since the joint distribution of $b$ implies that the marginal distributions of each $b_j$ is

$$b_j \sim N(\beta_j, \sigma^2 c_{jj}),$$

where $c_{jj}$ is the $j^{\text{th}}$ diagonal element of the matrix $(X^T X)^{-1}$.

ii. Estimation of Regression Scale

As in the case of simple linear regression, we will estimate $\sigma^2$ by starting from the sum of squared errors,

$$SSE = e^T e = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

In this instance, however, the appropriate number of degrees of freedom by which we divide the $SSE$ in order to arrive at our unbiased estimate of $\sigma^2$ is now $n-p$, so

$$s^2 = \frac{SSE}{n-p}.$$

As before, the number of degrees of freedom arises from the fact that we are now estimating $p$ parameters rather than just 2, and had we carried out the individual differentiations in the previous section we would have seen that the residuals must now satisfy $p$ linear constraints; namely,

$$\sum_{i=1}^{n} e_i = 0; \qquad \sum_{i=1}^{n} e_i x_{1i} = 0; \qquad \ldots \qquad \sum_{i=1}^{n} e_i x_{ki} = 0.$$

*Example 1 - Squid Data:* A study was conducted to examine the size of squid. Measurements and weight were taken on a sample of 22 squid. The response variable was the weight in pounds, and the 5 predictor variables were structural measurements of the squid's beak or mouth:

$x1$ : Rostral length in inches

$x2$ : Wing length in inches

$x3$ : Rostral to notch length

$x4$ : Notch to wing length

$x5$ : Width in inches

To fit the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

in *S-Plus* we would use the commands:

```
> squid <- as.data.frame(squid)
> names(squid)
[1] "x1"    "x2"    "x3"    "x4"    "x5"    "weight"
> attach(squid)
> reg.lm <- lm(weight ~ x1 + x2 + x3 + x4 + x5)
> reg.lm
Call:
lm(formula = weight ~ x1 + x2 + x3 + x4 + x5)

Coefficients:
 (Intercept)       x1        x2       x3       x4       x5
   -6.512215 1.999413 -3.675096 2.524486 5.158082 14.40116

Degrees of freedom:  22 total; 16 residual
Residual standard error:  0.7034523
```

Alternatively, we could get the same results using the commands:

```
> reg.out <- lsfit(cbind(x1,x2,x3,x4,x5),weight)
> reg.out$coef
 Intercept       x1        x2       x3       x4       x5
 -6.512215 1.999413 -3.675096 2.524486 5.158082 14.40116
> reg.sum <- ls.diag(reg.out)
> reg.sum$std.dev
[1] 0.7034523
```

Note that this is clearly a case where the $\beta_i$'s are more of a structural than an interpretable component to the model, since it is not likely that one could find a squid with any given set of beak measurements, so that holding four measurements fixed and changing the fifth measurement is not a very physically meaningful idea. Nonetheless, taken as a group they can tell us something about the relationship of the weight of a squid to its beak dimensions.

## V. Hypothesis Tests for The Parameters

i. Partitioning of Variability

As in the case of simple linear regression, the fundamental identity:

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ (SST) \quad &= \quad (SSR) \quad + \quad (SSE)\end{aligned}$$

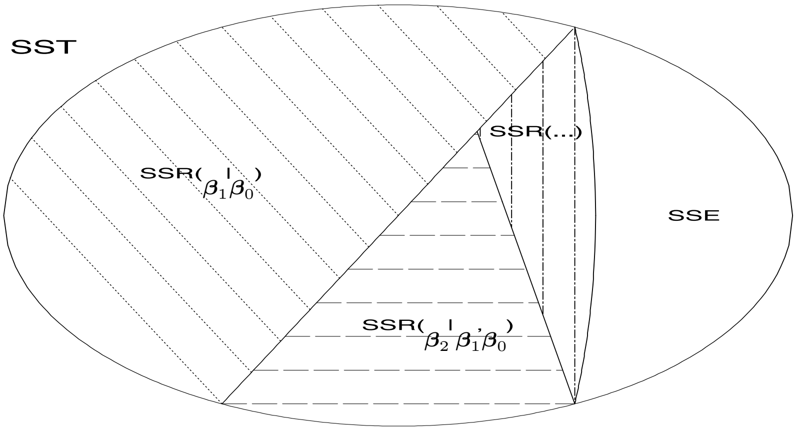
continues to hold. Similarly, the coefficient of determination, $R^2$, has the same interpretation as in simple linear regression and is again defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

In the case of multiple regression however, the associated degrees of freedom for error was seen to be $n - p$, and since the total degrees of freedom remains $n - 1$, we see that the regression degrees of freedom is now $p - 1$. Since the $SSR$ now has more than one degree of freedom, it is possible to break it down further into one degree of freedom pieces to help assess the individual usefulness of each of the predictors. The $SSR$ can be apportioned into so-called *sequential* sums of squares as follows:

$$\begin{aligned}SSR = SSR(\beta_1, \beta_2, \ldots, \beta_k|\beta_0) =& SSR(\beta_1|\beta_0) + SSR(\beta_2|\beta_1, \beta_0) + SSR(\beta_3|\beta_2, \beta_1, \beta_0) \\ &+ \ldots + SSR(\beta_k|\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-2}, \beta_{k-1}).\end{aligned}$$

The notation $SSR(\cdot|\cdot)$ indicates the amount of variation attributable to predictors associated with the parameters listed prior to the vertical line from among the variability left unexplained by the predictors associated with the parameters listed after the vertical line. For example, $SSR(\beta_2|\beta_1, \beta_0)$ is the amount of the unexplained variability from a simple linear regression on $x_1$ which is subsequently explained by $x_2$. In other words, it is simply the increase in the sum of squares obtained by adding the predictor $x_2$ to the model. This idea is best explained pictorially:


SST
SSR( $\beta_1 | \beta_0$ )
SSR(...)
SSE
SSR( $\beta_2 | \beta_1, \beta_0$ )

The ellipse represents the total sum of squares for a particular set of response values, and as regressions with respect to various predictors are performed, the ellipse is carved up into corresponding pieces. The right-most unshaded portion represents the sum of squared errors associated with the "full model" (sometimes denoted as $SSE_{full}$), that is, the model where all available predictors have been included. As each individual regressor is added sequentially (thus, the name "sequential sums of squares") it explains a bit of the remaining unexplained variation. So, the left-most, largest shaded region represents $SSR(\beta_1|\beta_0)$, i.e., the regression sum of squares from a simple linear regression on the first predictor. The lower central region then represents $SSR(\beta_2|\beta_1, \beta_0)$, the amount of the variation left unexplained by the first predictor which is subsequently explained by the second predictor, and so on. Note that the sum of the left and lower central region is

$$SSR(\beta_1|\beta_0) + SSR(\beta_2|\beta_1, \beta_0) = SSR(\beta_1, \beta_2|\beta_0),$$

the overall regression sum of squares for the regression using only the first two predictors. If we imagine fitting this model in the opposite order, then the breakdown of this overall region would look somewhat different, but the overall result is the same, since it is always the case that:

$$SSR(\beta_1|\beta_0) + SSR(\beta_2|\beta_1, \beta_0) = SSR(\beta_2|\beta_0) + SSR(\beta_1|\beta_2, \beta_0).$$

This fact about the sequential sums of squares can help us assess the worth of a particular subset of predictors. For example, if we have a model with 4 predictors, and we wish to investigate whether the last two are adding anything to the explanation of the response, we note that the overall regression sum of squares may be broken down as:

$$SSR(\beta_1, \beta_2, \beta_3, \beta_4|\beta_0) = SSR(\beta_1, \beta_2|\beta_0) + SSR(\beta_3, \beta_4|\beta_0, \beta_1, \beta_2).$$

If the second term on the right-hand side of this equality is only a small proportion of the overall total, then we might wish to conclude that the last two predictors are not adding much. We shall see how to formalize this in the next section.

Before going on to this, however, we bring to light an important aspect of these sequential regression sums of squares. As the diagram above shows, each new sequential sum of squares, $SSR(\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1})$ carves out its piece of the ellipse from what would have been the $SSE$ of the regression with only the first $j-1$ predictors. Thus, adding a predictor to a regression will *always* reduce the $SSE$ of the regression relative to the original regression. This does not necessarily mean that $s^2$ will always drop, since by adding a predictor we increase $p$, thereby reducing $n-p$, the denominator in calculating the $MSE$. However, it does mean that the coefficient of determination will always increase. Even if we add a completely superfluous predictor to our model! For this reason, we must be careful in our interpretation of an $R^2$ from a regression with a large number of predictors. We shall take up this topic again when we address the issue of model selection.

ii. Tests on Subsets of Parameters

As we noted previously, it is often the case that the predictor variables naturally partition themselves into two groups: those predictors with an already accepted relationship with the response, and those whose potential relationship is under investigation. If we suppose that the predictors are listed in an appropriate order, we may then partition both the parameter vector, $\beta$, and the design matrix, $X$, as:

$$X = [X_{(1)}\ X_{(2)}]; \qquad \beta = \begin{bmatrix} \beta_{(1)} \\ \beta_{(2)} \end{bmatrix},$$

where $X_{(1)}$ is an $n \times p_1$ matrix containing columns for each of the predictors associated with the parameters in the sub-vector $\beta_{(1)}$, and thus $X_{(2)}$ and $\beta_{(2)}$ contain the remaining columns of predictors and parameters, respectively. Our multiple linear regression model can then be re-written as:

$$Y = X\beta + \epsilon = X_{(1)}\beta_{(1)} + X_{(2)}\beta_{(2)} + \epsilon.$$

If $\beta_{(1)}$ consists of the parameters associated with the predictors under investigation, then we might wish to test the hypotheses:

$$H_0 : \beta_{(1)} = 0 \qquad \text{versus} \qquad H_A : \beta_{(1)} \neq 0.$$

As we noted in the previous section, testing this null hypothesis will be based on examining how much of the overall regression sum of squares the predictors associated with $\beta_{(1)}$ are contributing. More precisely, we note that if the errors are assumed to be normal, then under $H_0$, the $F$-statistic

$$F = \frac{SSR(\beta_{(1)}|\beta_{(2)})/p_1}{s^2}$$

has an $F$-distribution with $p_1$ numerator and $n-p$ denominator degrees of freedom. Thus, an $\alpha$ level test of $H_0$ is easily conducted. Also, note that if we want to test the overall significance of the regression; that is, test the null hypothesis that the response variable is not related to any of the predictors in the model, then we are in the case where $\beta_{(1)} = (\beta_1, \beta_2, \ldots, \beta_k)$, $\beta_{(2)} = \beta_0$, $p_1 = k = p-1$, and the test statistic becomes the familiar:

$$F = \frac{SSR(\beta_1, \beta_2, \ldots, \beta_k|\beta_0)/(p-1)}{s^2} = \frac{SSR/(p-1)}{MSE} = \frac{MSR}{MSE},$$

where $MSR = SSR/(p-1)$ is the overall regression sum of squares divided by its appropriate degrees of freedom, also called the mean square for regression. Similarly, if we want to test the significance of a single $b_j$, then we have $\beta_{(1)} = \beta_j$, $p_1 = 1$, and the test statistic becomes

$$F = \frac{SSR(\beta_j|\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_{k-1}, \beta_k)}{s^2}.$$

It can be shown that this test statistic is equal to the square of the $t$-statistic:

$$F = T^2 = \frac{b_j^2}{s^2 c_{jj}}.$$

As expected, the two testing procedures are equivalent.

Finally, we note that, in general, we can write:

$$\begin{aligned} SSR(\beta_{(1)}|\beta_{(2)}) &= SSR(\beta) - SSR(\beta_{(2)}) \\ &= SSR(\text{full model}) - SSR(\text{reduced model}) \\ &= SSE_{reduced} - SSE_{full}. \end{aligned}$$

Thus, our test of $H_0 : \beta_{(1)} = 0$ can be seen to be based on the drop in the residual sum of squares (i.e., the unexplained variation) from the model without the predictors associated with the parameters being tested to the full model with all the predictors. If this drop is sufficiently large, then we will reject the null hypothesis, otherwise, we will decide that the predictors associated with these parameters are not significantly adding to the explanatory power of the model and can reasonably be dropped.

*Example 1 (cont'd) - Squid Data:* To produce an ANOVA table with sequential sums of squares for the squid data model fit previously we would simply issue the *S-Plus* command:

```
> anova(reg.lm)
Analysis of Variance Table
Response:  weight
Terms added sequentially (first to last)
           Df Sum of Sq  Mean Sq  F Value     Pr(F)
x1          1  199.1453 199.1453 402.4397 0.0000000
x2          1    0.1267   0.1267   0.2560 0.6198036
x3          1    4.1195   4.1195   8.3249 0.0107645
x4          1    0.2635   0.2635   0.5325 0.4761139
x5          1    4.3522   4.3522   8.7951 0.0091089
Residuals 16    7.9175   0.4948
```

Note that *S-Plus* prints the sequential sums of squares in the order that the predictors were given in the `lm()` command. So, to test the overall significance of this regression, the ANOVA table can be used to calculate the overall regression sum of squares as $SSR = 199.1453 + 0.1267 + 4.1195 + 0.2635 + 4.3522 = 208.0072$, so that the desired $F$-statistic is:

$$F = \frac{208.0072/5}{0.4948} = 84.07728,$$

which must be compared to the quantiles of the $F_{5,16}$ distribution. Clearly, the regression is significant, meaning that at least one of the predictors has a strong relationship with the response, but we cannot say which ones from just this test. Suppose we decided to see if the last two predictors could be reasonably dropped from the model. In other words, we want to test the null hypothesis $H_0 : \beta_4 = \beta_5 = 0$. We can do this using the above ANOVA table as well since the required test statistic is

$$\begin{aligned} F &= \frac{SSR(\beta_4, \beta_5|\beta_0, \beta_1, \beta_2, \beta_3)/2}{s^2} \\ &= \frac{\{SSR(\beta_4|\beta_0, \beta_1, \beta_2, \beta_3) + SSR(\beta_5|\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)\}/2}{s^2} \\ &= \frac{(0.2635 + 4.3522)/2}{0.4948} = 4.664208. \end{aligned}$$

We compare this to the values $F_{2,16}(0.95) = 3.633723$ and $F_{2,16}(0.99) = 6.226235$ and see that we would reject $H_0$ at the $\alpha = 0.05$ level of significance, but not at the $\alpha = 0.01$ level.
We were able to perform this test using just the ANOVA table provided because the hypothesis in which we were interested happened to pertain to the predictors which were fit last into the model. However, if we had been interested in testing the hypothesis $H_0 : \beta_1 = \beta_4 = 0$, then we could not have used the ANOVA table provided to calculate the required regression sum of squares, $SSR(\beta_1, \beta_4|\beta_0, \beta_2, \beta_3, \beta_5)$. This is not a problem, however, since we merely need to refit the model with the predictors entered in a different order. Nonetheless, knowing which specific tests we are interested in beforehand will allow us to put the predictors into the `lm()` function in the most useful initial ordering. For this latest example, we could use:

```
> reg.lm <- lm(weight ~ x2 + x3 + x5 + x4 + x1)
> anova(reg.lm)
Analysis of Variance Table

Response:  weight

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
x2         1  190.2797 190.2797 384.5237 0.0000000
x3         1    7.6548   7.6548  15.4691 0.0011881
x5         1    6.9421   6.9421  14.0288 0.0017644
x4         1    2.8319   2.8319   5.7228 0.0293692
x1         1    0.2987   0.2987   0.6037 0.4485105
Residuals 16    7.9175   0.4948
```

So, the required $F$-statistic is:

$$\begin{aligned} F &= \frac{SSR(\beta_4, \beta_1|\beta_0, \beta_2, \beta_3, \beta_5)/2}{s^2} \\ &= \frac{\{SSR(\beta_4|\beta_0, \beta_2, \beta_3, \beta_5) + SSR(\beta_1|\beta_0, \beta_2, \beta_3, \beta_4, \beta_5)\}/2}{s^2} \\ &= \frac{(2.8319 + 0.2987)/2}{0.4948} = 3.164. \end{aligned}$$

Therefore, the $p$-value for the test of $H_0 : \beta_1 = \beta_4 = 0$ is $Pr\{F_{2,16} > 3.164\} = 0.0695$ and we would probably not reject the null hypothesis and would therefore consider dropping these terms from the model.

iii. Tests for individual parameters
The preceding ANOVA table could also be used to test the hypothesis $H_0 : \beta_1 = 0$. The appropriate $F$-statistic would be $F = 0.2987/0.4948 = 0.6037$ and we clearly would not reject $H_0$. Of course, we could have used the more standard $t$-test approach to this problem. The required *S-Plus* command is:

```
> summary(reg.lm)

Call:  lm(formula = weight ~ x2 + x3 + x5 + x4 + x1)
Residuals:
    Min      1Q Median    3Q    Max
-1.261 -0.5373 0.1355 0.512 0.8611

Coefficients:
               Value Std.  Error  t value Pr(>|t|)
```

```
(Intercept)  -6.5122   0.9336    -6.9757   0.0000
         x2  -3.6751   2.7737    -1.3250   0.2038
         x3   2.5245   6.3475     0.3977   0.6961
         x5  14.4012   4.8560     2.9656   0.0091
         x4   5.1581   3.6603     1.4092   0.1779
         x1   1.9994   2.5733     0.7770   0.4485

Residual standard error:  0.7035 on 16 degrees of freedom
Multiple R-Squared:  0.9633
F-statistic:  84.07 on 5 and 16 degrees of freedom,
the p-value is 6.575e-11

Correlation of Coefficients:
   (Intercept)      x2      x3      x5      x4
x2 -0.0091
x3 -0.4411     -0.4616
x5  0.6917      0.0283 -0.6377
x4  0.3394     -0.4660 -0.0074  0.2361
x1 -0.5383     -0.1518  0.0452 -0.4286 -0.6147
```

Amongst all the other information, this command displays the standard errors for each of the parameter estimates, and the $t$-statistics as well. For the predictor `x1`, the $t$-statistic is 0.777, which shows that

$$T^2 = (0.777)^2 = 0.6037 = F,$$

confirming the relationship between the two procedures for testing a single parameter.

There is an important sidelight to this test. If we were to fit a *simple* linear model of weight on the first predictor, the associated regression sum of squares would be $SSR(\beta_1|\beta_0) = 199.145$ and the residual sum of squares would be $SSE = SSE_{full} + SSR(\beta_2|\beta_0,\beta_1) + SSR(\beta_3|\beta_0,\beta_1,\beta_2) + SSR(\beta_4|\beta_0,\beta_1,\beta_2,\beta_3) + SSR(\beta_5|\beta_0,\beta_1,\beta_2,\beta_3,\beta_4) = 7.9175 + 0.1267 + 4.1195 + 0.2635 + 4.3522 = 16.78$. Thus, the $F$-statistic for the test of $H_0 : \beta_1 = 0$ would now be:

$$F = \frac{199.145}{16.78/20} = 237.360,$$

and we would reject $H_0$ very strongly. This fact points out the difference between the sequential $F$-test approach and the so-called *partial* $F$-tests (i.e., testing each of the parameters individually from the full model) and demonstrates that the importance and significance of a predictor variable in a regression depends heavily on what other predictors are in the model. The proper interpretation of the preceding example is that, on its own the first predictor has a significant relationship to the response, however once the other four variables have been included, the first predictor is no longer necessary and does not significantly add to the explanation of the variation in the response. In fact, it can happen that the overall regression is highly significant, but that each of the individual parameters are not significantly different from zero. This is a disconcerting situation, and we will investigate it in further detail when we discuss the idea of multicollinearity of the predictor variables.

iv. Confidence and Prediction Intervals

As in the simple linear model, we will often be interested in predicting the value of the response associated with a particular set of predictor values, $x_0 = (1, x_{01}, \ldots, x_{0k})^T$, where we have included the leading one associated with the intercept for convenience of notation. Note that

$x_0^T$ now looks like a new row of the design matrix. Clearly, our best guess for the value of the response at $x_0$ is:

$$\hat{Y}(x_0) = b_0 + b_1 x_{01} + \ldots + b_k x_{0k} = x_0^T b.$$

Similarly, following the development of the simple linear model case, we see that

$$Var\{\hat{Y}(x_0)\} = Var(x_0^T b) = x_0^T Var(b)\{x_0^T\}^T = \sigma^2 x_0^T (X^T X)^{-1} x_0.$$

Therefore, we can see that a $100(1-\alpha)\%$ confidence interval for the expected response associated with the set of predictor values, $x_0$, is given by:

$$\hat{Y}(x_0) \pm t_{n-p}(1-\alpha/2) s \sqrt{x_0^T (X^T X)^{-1} x_0}.$$

Again using arguments analogous to those for the case of the simple linear model, we can see that a $100(1-\alpha)\%$ prediction interval can also be calculated as:

$$\hat{Y}(x_0) \pm t_{n-p}(1-\alpha/2) s \sqrt{1 + x_0^T (X^T X)^{-1} x_0}.$$

*Example 2 - Forestry Data:* Twenty stands of pine trees were measured in an effort to assess the amount and quality of wood which would be obtained. Each stand's age (`AGE`), average height of the dominant trees (`HD`), number of trees (`N`) and average diameter at 4.5 feet above the ground (`MDBH`) were obtained. Theory suggests that `MDBH` may be effectively modelled as:

$$MDBH = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

where $x_1 = HD$, $x_2 = AGE \cdot N$ and $x_3 = HD/N$. Suppose that we wanted to predict the expected value of all pine stands as well as the value of an individual pine stand's `MDBH` for three different types of stands. The first is a 5 year old stand having 500 trees of an average height of 10 feet. The second is a 10 year old stand having 600 trees of an average height of 80 feet. The third is a 25 year old stand having 1000 trees of an average height of 75 feet. Using the `predict()` command in *S-Plus*:

```
> pine <- as.data.frame(pine)
> x1 <- HD
> x2 <- AGE*N
> x3 <- HD/N
> pine.lm <- lm(MDBH ~ x1 + x2 + x3)
> anova(pine.lm)
Analysis of Variance Table

Response:  MDBH

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
x1         1  6.207045 6.207045 72.01171 0.0000003
x2         1  2.784561 2.784561 32.30539 0.0000339
x3         1  0.014774 0.014774  0.17140 0.6843652
Residuals 16  1.379119 0.086195
> x1 <- c(10,80,75)
> x2 <- c(5*500,10*600,25*1000)
> x3 <- c(10/500,80/600,75/1000)
> prd <- predict(pine.lm,as.data.frame(cbind(x1,x2,x3)),se.fit=T)
> CIup <- prd$fit+(qt(0.975,16)*prd$se.fit)
> CLlw <- prd$fit-(qt(0.975,16)*prd$se.fit)
> cbind(CIlw,prd$fit,CIup)
      CIlw              CIup
1 3.279259  3.85699  4.434721
2 9.027290 10.47730 11.927312
3 5.848304  6.57982  7.311336
> pi.std <- sqrt(prd$residual.scale^2+prd$se.fit^2)
> PIup <- prd$fit+(qt(0.975,16)*pi.std)
> PIlw <- prd$fit-(qt(0.975,16)*pi.std)
> cbind(PIlw,prd$fit,PIup)
      PIlw              PIup
1 3.007795  3.85699  4.706186
2 8.899362 10.47730 12.055240
3 5.619364  6.57982  7.540276
```

Now, examination of the ANOVA table shows that once the first two predictors are included in the model, the third predictor does not add substantially to the model. So, if we were to re-fit the model and re-predict without this predictor we would have:

```
> x1 <- HD
> x2 <- AGE*N
> pine.lm <- lm(MDBH ~ x1 + x2)
> anova(pine.lm)
Analysis of Variance Table

Response:  MDBH

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
x1         1  6.207045 6.207045 75.70148 1.142600e-07
x2         1  2.784561 2.784561 33.96067 2.018102e-05
Residuals 17  1.393893 0.081994
> x1 <- c(10,80,75)
> x2 <- c(5*500,10*600,25*1000)
> prd <- predict(pine.lm,as.data.frame(cbind(x1,x2)),se.fit=T)
> CIup <- prd$fit+(qt(0.975,17)*prd$se.fit)
> CLlw <- prd$fit-(qt(0.975,17)*prd$se.fit)
> cbind(CIlw,prd$fit,CIup)
      CIlw              CIup
1 3.294539  3.855267  4.415995
2 9.667724 10.675075 11.682427
3 5.860646  6.535209  7.209772
> pi.std <- sqrt(prd$residual.scale^2+prd$se.fit^2)
> PIup <- prd$fit+(qt(0.975,17)*pi.std)
> PIlw <- prd$fit-(qt(0.975,17)*pi.std)
> cbind(PIlw,prd$fit,PIup)
```

```
       PIlw                PIup
1 3.031012  3.855267  4.6796522
2 9.500453 10.675075 11.849698
3 5.629662  6.535209  7.440756
```

As a final note, we point out that the confidence and prediction intervals calculated in the preceding example are *individual* intervals. In other words, their coverage percentage pertains to only whether they themselves will cover the true value they are estimating. If we wanted to have *simultaneous* coverage of at least 95% for all three intervals, we could use the Bonferroni procedure. To do so, we would merely need to expand each of the individual coverage percentages from 95% up to $100(1-0.05/3) = 98\frac{1}{3}\%$ intervals. Alternatively, we could use the so-called Working-Hotelling procedure, which simply amounts to replacing the usual $t_{n-p}(1-\alpha/2)$ multiplier in each of the individual intervals with the new multiplier $\sqrt{pF_{p,n-p}(1-\alpha)}$. Typically, if $p$ is large and we want joint confidence intervals for only a few expected responses, then the Bonferroni procedure will be considerably more efficient, in the sense that it will provide a much narrower set of valid joint intervals. Finally, we could also create an elliptical joint confidence region. To do this, we would appeal to the ideas in the very next section.

## VI. Model Diagnostics

We will now explore the methods of investigating whether our model is properly specified and whether the underlying assumptions appear plausible for the dataset at hand.

i. Residual Analysis

As for simple linear regression, we must examine the plausibility of the underlying assumptions of our multiple regression model, and again the most useful diagnostic tool for such investigations will be the residuals, $e = Y - \hat{Y}$.

The first and most important diagnostic tool is, as it was for simple linear regression, a plot of the residuals. Of course, in the case of multiple regression, we have several predictor variables and thus a plot of the residuals versus the predictors is not generally possible (though plots against each of the predictor variables individually certainly are). So, for multiple regression, we will generally plot the residuals versus the fitted values. The principle behind such a plot is the same as it was for simple linear regression. We are generally looking for patterns, which would indicate a non-linearity in the data, or for a funnelling shape, which would indicate heteroscedasticity.

Why have we plotted the residuals versus the fitted values? What we are generally looking for in our residual plot is some pattern to indicate that either the central tendency of the residuals or their variability is changing as we move through the vector space of predictor values. In effect, the fitted value is the most useful way to determine where in the range of the predictor values the residual in question is located. And thus, a plot of residuals versus the fitted values does indeed give an indication as to whether the response variable is systematically trending away from what the regression surface would predict for it, or whether the variability of the residuals is changing in relation to where the data point lies within the range of the predictor values. Of course, it may still be of interest to examine the residuals versus each of the predictors individually, particularly in the case of a noticable non-linearity in the residuals versus fitted values plot. If an overall non-linearity is detected, the plots of the residuals versus each of the predictor variables may give guidance into which variables are at the root of the non-linearity and should potentially be transformed (though, they are not generally as useful as the so-called *added variable plots* which we shall discuss shortly).

As we noted in the case of simple linear regression, while $Var(\epsilon) = \sigma^2 I$, the variance-covariance matrix of the residuals is given by $Var(e) = \sigma^2(I - H)$, where $H = X(X^T X)^{-1} X^T$ is the hat matrix. This means that the behavior of the residuals would not be expected to exactly mirror that of the true errors. And thus, simply examining the *ordinary* residuals, $e_i$, may not be the optimal way of investigating the underlying behavior of the $\epsilon_i$'s, upon which the assumptions are actually made. The variance-covariance matrix of the residuals demonstrates that they do not behave like the true errors in two important ways. First, they do not all have the same variability, since $Var(e_i) = \sigma^2(1 - h_{ii})$, and the leverage values, $h_{ii}$, are typically different for each data point. Second, since the matrix $H$ is generally not diagonal, the residuals are not uncorrelated.

The second problem is generally not a large problem, since (at least when $n$ is moderately large) it is rare that the correlations between the residuals are large (i.e., the $h_{ij}$'s for $i \neq j$ are typically very near zero). We can, however, remedy the first problem by noting that

$$Var\left(\frac{e_i}{\sigma\sqrt{1 - h_{ii}}}\right) = \frac{Var(e_i)}{\sigma^2(1 - h_{ii})} = \frac{\sigma^2(1 - h_{ii})}{\sigma^2(1 - h_{ii})} = 1.$$

In other words, the scaled residual values $e_i/(\sigma\sqrt{1 - h_{ii}})$ now at least all have the same variability. Of course, we do not generally have the value of $\sigma^2$ available to us, and thus we will define the *Studentized* residuals as:

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}} = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}.$$

A residual plot of these values versus the fitted values should look like a random (rectangular or elliptical) scatter of points. As with the ordinary residual plot, patterns and funnelling indicate potential non-linearity or heteroscedasticity, respectively. In addition, note that the Studentized residuals have now been scaled so that they have unit variance, meaning that outliers are somewhat easier to detect, since if the errors are truly normally distributed, then the Studentized residual values should generally lie between -2 and 2, regardless of the ordinary residual scale, $s$.

Similarly, a normal q-q plot of the Studentized residuals should look like a straight line, this time with a slope of 1, rather than the ordinary residual scale, $s$. As in the case of the ordinary residuals, the basic patterns which describe heavy- or light-tailed residuals or skewed residuals are still the standard departures for which we should look.

ii. Outlier detection

We will now embark on a more detailed discussion of the detection of "outlying" data points. Before we begin, however, two important points should be made. First, the $i^{\text{th}}$ data point may be considered an outlier if any of the model assumptions break down at this point. Specifically, the two most common sources of outliers are:

- There is a location shift at the $i^{\text{th}}$ data point, so that $E(\epsilon_i) = \Delta_i \neq 0$.
- There is a scale shift at the $i^{\text{th}}$ data point, so that $Var(\epsilon_i) > \sigma^2$.

Of course, these discrepancies may be caused by something "real" (i.e., related to the underlying nature of the populations under study) or simply by a data collection error.

The second point which needs to be made is that the outlier detection schemes which we are about to study should be thought of as primarily diagnostic and not truly formal statistical procedures. It is very tempting to automatically eliminate any data point whose removal will subsequently improve some measure of the quality of the fit of the model. However, we must take care to ensure that our model does not solely represent the central majority of our observed data (i.e., an "overfit" of the data), and instead gives a useful description of the overall population of interest. The decision to remove a data point should never be a purely statistical one, but should be made only in the proper context of the field of study under investigation. Often, an apparently outlying point can lead to further investigation and help to uncover a flaw in the initial model structure. Thus, points which simply "do not follow the trend" should not be blithely discarded without any thought to the potential reasons for their discrepancy.

Now, it might initially be argued that, since every data point has some influence on the resulting fitted regression equation, we should measure the degree to which data points are outliers by using the so-called *PRESS (PREdiction Sum of Squares) residual*:

$$e_{i,-i} = Y_i - \hat{Y}_{i,-i},$$

where $\hat{Y}_{i,-i}$ is the predicted value at $x_i$ from a regression fit to the dataset with the $i^{\text{th}}$ point removed. In other words, $e_{i,-i}$ measures how far the $i^{\text{th}}$ response is from a prediction over which it has no influence. Note that, if there is a location shift, $\Delta_i$, at the $i^{\text{th}}$ data point then

$$E(e_{i,-i}) = \Delta_i \neq E(e_i).$$

It might at first seem time-consuming to calculate all the $e_{i,-i}$'s since they apparently each require fitting a new regression. However, an interesting relation exists between the PRESS residuals and the ordinary residuals:

$$e_{i,-i} = \frac{e_i}{1-h_{ii}},$$

where $h_{ii}$ is just the usual leverage value for the $i^{\text{th}}$ data point in the original regression with all the data points included. Again, though, we should standardize the PRESS residuals, since they do not all have the same variance:

$$\begin{aligned}
Var(e_{i,-i}) &= Var\left(\frac{e_i}{1-h_{ii}}\right)\\
&= \frac{1}{(1-h_{ii})^2}Var(e_i)\\
&= \frac{\sigma^2(1-h_{ii})}{(1-h_{ii})^2}\\
&= \frac{\sigma^2}{1-h_{ii}}.
\end{aligned}$$

So, standardizing the PRESS residuals yields:

$$\begin{aligned}
\frac{e_{i,-i}}{\sqrt{Var(e_{i,-i})}} &= \frac{e_i}{(1-h_{ii})\sqrt{Var(e_{i,-i})}}\\
&= \frac{e_i}{(1-h_{ii})\sqrt{\sigma^2/(1-h_{ii})}}\\
&= \frac{e_i}{\sigma\sqrt{1-h_{ii}}}.
\end{aligned}$$

So, once we replace $\sigma$ by $s$, we see that the standardized PRESS residuals are the same as the Studentized residuals, adding credibility to the use of the $r_i$'s as outlier diagnostics. Note, however, that we have glossed over an important point here. In standardizing the PRESS residuals, we used $s$, the regression scale from the original regression on the entire dataset, to estimate $\sigma$, and doing so led us to the (*internally*) Studentized residual. However, perhaps a better estimate of $\sigma$ in this case might be the residual scale from the regression calculated without the $i^{\text{th}}$ data point, denoted by $s_{-i}$, which can be calculated as:

$$s_{-i} = \sqrt{\frac{(n-p)s^2 - \{e_i^2/(1-h_{ii})\}}{n-p-1}}.$$

If we now use this estimate for $\sigma$ we arrive at the *externally Studentized* residual:

$$t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} = \frac{e_{i,-i}}{s_{-i}/\sqrt{1-h_{ii}}}.$$

Typically, $r_i$ and $t_i$ give very similar values, however, the size of the discrepancy between them will depend on the influence of the $i^{\text{th}}$ data point. Our formula for $s_{-i}$ shows that a point with a large ordinary residual and a large leverage will produce a sizable difference between $s$ and $s_{-i}$ and thus between $r_i$ and $t_i$. Often, the externally Studentized residual will be a more sensitive detector of outliers. In fact, if we assume that the $\epsilon_i$'s are indeed normally distributed, then each $t_i$ is distributed according to a $t$-distribution with $n-p-1$ degrees of freedom under the assumption that the $i^{\text{th}}$ data point does not suffer from a location shift, that is, under $H_0 : \Delta_i = 0$, (this is not true for the internally Studentized residual, $r_i$, since, unlike $t_i$ it cannot be written as a ratio of two independent quantities). This provides us with a formal mechanism to test for the presence of location shift outliers. In fact, the externally Studentized residual will also pick up scale shifts. Of course, we should recall our initial warning about the treatment of outlier detection as a formal statistical technique instead of a simple diagnostic tool.

*Example 5 - Body fat data:* A study of the relationship of body fat to several physical measurements was conducted on a sample of 20 healthy women aged 25-34 years. The predictor measurements were the triceps skinfold, the thigh circumference and the midarm circumference. If we perform a multiple linear regression of body fat on the three predictors, the associated internally and externally Studentized residuals are:

```
> bft.df <- as.data.frame(bft)
> attach(bft.df)
> names(bft.df)
[1] "Triceps" "Thigh" "Midarm" "BodyFat"
> bft.reg <- lsfit(cbind(Triceps,Thigh,Midarm),BodyFat)
> bft.sum <- ls.diag(bft.reg)
> ri <- bft.sum$std.res
> ti <- bft.sum$stud.res
> cbind(ri,ti)
            ri          ti
 1 -1.46802633 -1.52803951
 2  1.13326956  1.14416429
 3 -1.23262045 -1.25452990
 4 -1.29571232 -1.32606735
 5  0.57630252  0.56388572
 6 -0.23525791 -0.22818249
 7  0.62249950  0.61016668
 8  1.38022830  1.42385079
 9  0.76529502  0.75493974
10 -0.57761774 -0.56519997
11  0.34965240  0.33985038
12  0.94324119  0.93979234
13 -1.50477923 -1.57251203
14  1.84715613  2.01637183
15  0.49352568  0.48153342
16  0.07392664  0.07159138
17 -0.16107977 -0.15609143
18 -0.63614383 -0.62388391
19 -1.61308352 -1.70680191
20  0.25538013  0.24777625
```

So, we see that none of the Studentized residuals seem excessively large. Suppose that we wished to test whether the data point with the largest Studentized residual is a mean-shift outlier. To do so, we could compare its value of 2.01637183 to $t_{15}(0.975) = 2.13145$ or, to be conservative, to $t_{15}(1 - 0.05/(2 \times 20)) = t_{15}(0.99875) = 3.623918$ (note that $n - p - 1 = 20 - 4 - 1 = 15$ for this example).

Bear in mind the warning regarding interpreting the $p$-values for these sorts of tests literally, since we have intentionally sought out the largest Studentized residuals. Now, if we had some *a priori* reason for thinking that the 14th woman in the study would have a mean shift in her body fat measurement (i.e., some reason based on information which we had prior to seeing the residual value associated with her data), then we could be a bit more comfortable in the standard interpretation of the associated $p$-value.

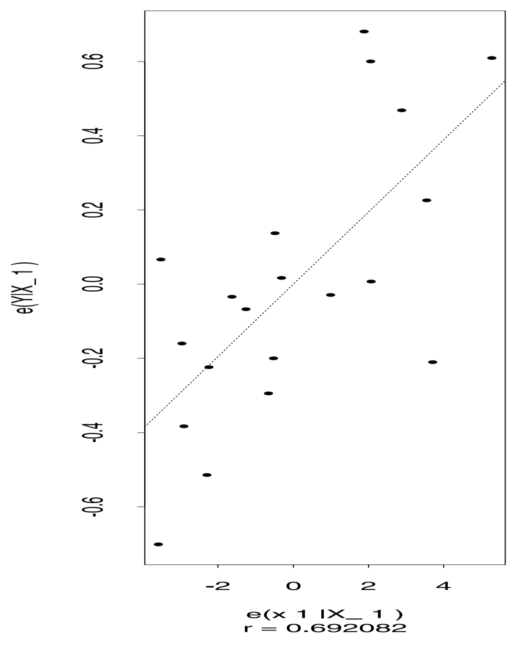
iii. Added Variable Plots (Partial Regression Plots)
As we will shortly discuss, one of the major questions which arises in multiple regression is which of the predictors to include in the model. Specifically, we have seen that the sequential sums of squares can give us a formal testing framework for deciding whether a particular variable is adding anything new to the explanation of the variation in the response variable. Similar information can be gathered pictorially through the use of *added variable plots*. Simply looking at a plot of the response variable versus an individual predictor can provide evidence of whether there is a relationship between the two variables in isolation, however, such relationships may not be "real" in the sense that there may be some *confounding* variable which is actually related to both of the plotted variables and thus is causing an apparent relationship between them (of course, confounding can also work in the reverse direction as well, that is, to disguise a real relationship). Thus, such plots will not provide reliable evidence of whether there is any relationship between the two variables after having "adjusted" for the effects of other predictors. To display such evidence pictorially, we use *added variable plots* (also called *partial regression plots*). An added variable plot for a particular predictor, $x_j$, displays the residuals from a regression of the response on all of the predictors *except* $x_j$, which we denote $e_{Y|X_{-j}}$, and the residuals from a regression using $x_j$ as the response and the remaining $x_i$'s as

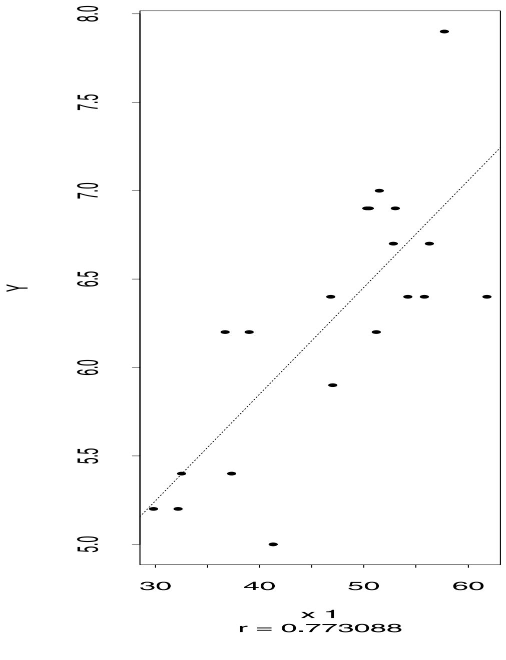
the predictors, which we denote $e_{x_j|X_{-j}}$. The idea here is to remove the effects of all the other variables from both the response and the particular predictor in question and then examine the relationship between the remaining "unexplained" portions of the two variables in question. The idea is that any relationship which is seen between $e_{Y|X_{-j}}$ and $e_{x_j|X_{-j}}$ cannot have been "contaminated" by any of the possible (linear) confounding effects of the other variables.
It can be shown that if the linear model $Y = X\beta + \epsilon$ holds, then

$$e_{Y|X_{-j}} = \beta_j e_{x_j|X_{-j}} + \epsilon^\star,$$

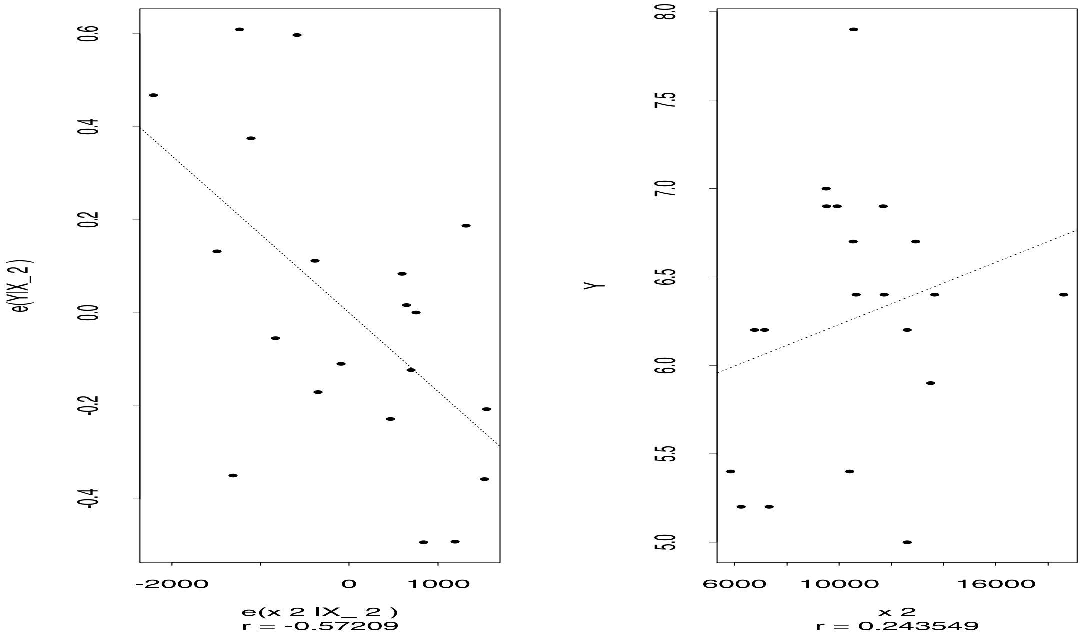
where $\epsilon^\star = (I - H_{-j})\epsilon$ and $H_{-j}$ is the hat matrix calculated from $X_{-j}$, the design matrix containing all of the predictors except $x_j$. Thus, the slope of the partial regression plot is an unbiased estimator of $\beta_j$. In fact, if we perform a least-squares regression *through the origin* of $e_{Y|X_{-j}}$ on $e_{x_j|X_{-j}}$, the resulting slope estimate will be *exactly* $b_j$, the least-squares estimate from the full multiple regression of the response on all the predictors.
Thus, if an added variable plot shows a linear structure, this is evidence that the predictor variable under investigation should indeed be included in the model. On the other hand, if the added variable plot appears to be a simple random scatter of points, then we will likely conclude that the predictor is not adding any further explanation of the response and can be dropped from the model (as long as the other variables are retained, since dropping a different predictor from the model may make the predictor under study suddenly significant, we will discuss such situations in the later subsection on multicollinearity). Also, recall that the correlation of two variables indicates the strength of their linear relationship, so we can also include the value of $r$ along with a simple linear regression line on our added variable plot to help us determine the strength of the linear association.

*Example 2 (cont'd) - Forestry Data:* Recall the dataset regarding the measurements of 20 different pine stands. For this data we fit a multiple regression of the response variable, $MDBH$, on the three predictors, $x_1 = HD$, $x_2 = AGE \cdot N$ and $x_3 = HD/N$. Below are the added variable plots along with the associated plot of the "unadjusted" response versus predictor plots. For the first predictor variable:

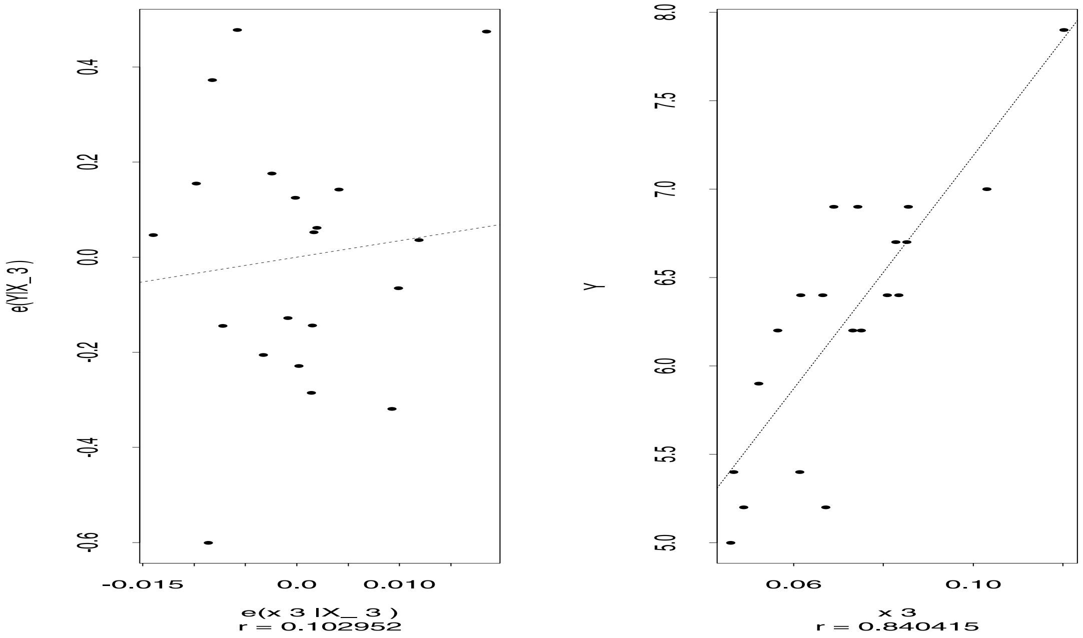




For this variable, we do not see much distinction between the two plots, and clearly $x_1$ is an important predictor to include in our model. For the second predictor:



In this case, not only do we see the added variable plot indicating a stronger relationship than the "unadjusted" plot, but the direction of the relationship it indicates is in the opposite direction. This is why it is important to examine the added variable plot, since interrelationships between the predictors can cause the simple plots of the response versus each individual predictor to be misleading. Again, by virtue of the added variable plot, we see that the second predictor is an important variable in the model. Finally, for the third predictor:



For this predictor, the added variable plot indicates that it is not very important for the regression (a fact that we were able to determine from our $F$-tests as well).

Another use of the added variable plots is in determining where any potential non-linearity is in the data. If the predictor $x_j$ is truly linearly related to the response variable, then the added variable plot should look like a straight line through the origin. Any distinguishable curvature in the added variable plot indicates a non-linearity in the relationship between the response and the predictor under study. When used in concert with the plots of the residuals versus each of the predictors, they can indicate the probable sources of any non-linearity which may have been noted (e.g., in the overall plot of the residual versus the fitted values). The added variable plots of the previous example show no noticable non-linearity, so the first two variables (which were shown to be important) should be entered into the model linearly (i.e., without any transformation or inclusion of higher-order terms, such as quadratics or cubics). It turns out that the added variable plots are not necessarily the most useful tool for discovering non-linearity, and several more complicated variations of them have been proposed, but we shall not discuss them here.

In addition to non-linearity, added variable plots may also uncover pictorially which points are having an undue influence over the results. For example, the point in the upper right corner of the last added variable plot in the preceding example may be having a strong influence on our interpretation of the plot. To investigate this potential problem further, we need to use the methods of the next subsection.

iv. Influence Diagnostics

The previous section covering outlier detection dealt with identifying data points which were discrepant in terms of their observed residual value. In other words, these are data points which are aberrant in the vertical direction. We also want to identify points which are aberrant in the horizontal direction, just as we did for simple linear regression. Of course, for multiple regression, there are now several possible "horizontal" directions. Nonetheless, we can still use the leverage values, $h_{ii}$, to assess which data points are potentially influential. However, since we are somewhat restricted in our ability to examine these data points graphically, it is not as easy to simply pick out visually whether they are having large influence over the regression fit. Thus, we will develop some numerical diagnostic tools to help us determine how and which, if any, data points are having an undo influence over our regression. We should again remember, however, that like the case for outlier detection, these numerical tools are merely diagnostic, and removal of influential points should not be made solely on the grounds of these measures (although, recall that if the predictor values are indeed under our control, then we should be careful in our initial selection of predictor values so as not to include data points with a high potential for excessive influence).

As for a simple regression, we start with the leverages, $h_{ii}$, and we note that it is not difficult to show:

$$\sum_{i=1}^{n} h_{ii} = p,$$

where $p$ is the number of parameters in the model. Thus, we again will say that any data point which is more than twice the average leverage value (i.e., any point for which $h_{ii} > 2p/n$) has the *potential* for exerting a high influence over the regression fit. Of course, just because a data point has a large leverage value does not necessarily mean that it has high influence, and we now introduce several measures designed to investigate the actual degree of influence that each data point has.

The conceptual idea behind "influence" is that it measures the amount by which the removal of a single data point will change the results of our regression analysis. All of our influence diagnostics will be based on this concept. First, we can investigate the extent to which the removal of the $i^{\text{th}}$ data point affects the associated fitted value for this point using:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}}.$$

Clearly, the $DFFITS_i$ value measures the amount by which the prediction at $x_i$ changes when the $i^{\text{th}}$ data point is removed. The denominator is based on the fact that $Var(\hat{Y}) = \sigma^2 H$, and thus it is a "hybrid" of the standard deviation of the predictions $\hat{Y}_i$ and $\hat{Y}_{i,-i}$. Thus, the $DFFITS_i$ value measures the approximate number of standard deviations by which the predicted value at $x_i$ changes if the $i^{\text{th}}$ data point is included in the analysis. It can be shown (see tutorial exercises) that the $DFFITS_i$'s can be calculated without apparent necessity of re-fitting the $n$ different regressions required for the calculation of the $\hat{Y}_{i,-i}$'s:

$$DFFITS_i = t_i\sqrt{\frac{h_{ii}}{1-h_{ii}}},$$

where $t_i$ is the externally Studentized residual for the $i^{\text{th}}$ data point from the regression including all the data points.

We can also investigate the extent to which each data point is influencing the values of the parameter estimates. To do this, we have two useful diagnostic measures. The first is:

$$DFBETAS_{j,i} = \frac{b_j - b_{j,-i}}{s_{-i}\sqrt{c_{jj}}},$$

where $b_{j,-i}$ is the least-squares estimate of $\beta_j$ from the regression fit without the $i^{\text{th}}$ data point and $c_{jj}$ is the $j^{\text{th}}$ diagonal element of $(X^TX)^{-1}$, the hat matrix from the regression including all the data points. As for the $DFFITS_i$'s, the $DFBETAS_{j,i}$ value measures the approximate number of standard deviations by which the least-squares estimator of $\beta_j$ changes upon the inclusion of the $i^{\text{th}}$ data point into the analysis (since $Var(b_j) = \sigma^2 c_{jj}$).

The second measure of the influence each data point has over the parameter estimates is Cook's distance:

$$D_i = \frac{(b-b_{-i})^T(X^TX)(b-b_{-i})}{ps^2}.$$

Note that Cook's distance measures the *overall* effect on the parameter estimates (i.e., the effect on the entire parameter vector), whereas the $DFBETAS_{j,i}$'s measure the effects on each of the parameter estimates individually. To better understand the construction of Cook's distance, we note that the vector $d_i = b - b_{-i}$ measures the change in the least-squares estimator associated with the removal of the $i^{\text{th}}$ data point. Since $Var(b) = \sigma^2(X^TX)^{-1}$, we see that Cook's distance is just an approximate standardization of the vector $d_i$. In fact, $D_i$ can indeed be thought of as a distance measure, representing the "standardized" distance between the two vectors $b$ and $b_{-i}$. As a distance, it will be the case that $D_i$ will always be a positive number (as opposed to $DFFITS_i$ and $DFBETAS_{j,i}$ which can be either positive or negative). In fact, it can be shown (see tutorial exercises) that:

$$D_i = \left(\frac{r_i^2}{p}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right),$$

where $r_i$ is the internally Studentized residual for the $i^{\text{th}}$ data point. Note the similarity between $D_i$ and the square of $DFFITS_i$.

All of these measures of influence lead inevitably to the question of how large the values of the diagnostic measures need to be before we start to worry that a data point is exerting excessive influence over our regression fit. We note that $DFFITS_i$ and $DFBETAS_{j,i}$ were constructed as "standardized discrepancies" and thus might be thought to behave in a "$t$-like" manner, indicating that values outside the range -2 to 2 would be thought of as indicating influential points. However, the quantities which make up each of these measures are not statistically independent, and thus the $t$-distribution cannot be a formal yardstick for determining influence. In addition, the size of the dataset must be considered, since values of these diagnostics are rarely ever outside the range -2 to 2 in datasets with a large number of observations (recall that the leverage values were compared to a "cut-off" value which depended on the sample size). In large samples, commonly suggested values for comparison for the two measures are:

- Any datapoint with $|DFFITS_i|$ greater than $2\sqrt{p/n}$ should be regarded as potentially influential;
- A datapoint with any $|DFBETA_{j,i}|$ greater than $2/\sqrt{n}$ should be regarded as potentially influential.

Of course, any formal rule to determine which points are influential contradicts our initial warning that these tools should be thought of as merely diagnostic in nature. As long as we are aware of what these tools are actually measuring, and we have a basic understanding of the context in which the data was gathered, then we can use this knowledge (along with our general experience with regressions in various settings) to tell us what a reasonable level is for determining which are the "highly" influential points that require further investigation.

Similarly, Cook's distance has the appearance of an $F$-ratio (e.g., the ratio used to construct simultaneous confidence regions for $\beta$), and indeed, commonly used comparison values for determining "large" values of $D_i$ are the quantiles of the $F_{p,n-p}$ distribution. Again, however, a formal cut-off value goes against the general tenor of the use of these measures as simply diagnostic tools.

Note that all three of the above measures attempt to pinpoint data values which are having a strong influence over the *results* of the regression (e.g., predictions and parameter estimates). However, this is only part of the story. For example, a point with a large $DFBETAS_{j,i}$ merely means that this data point has a large effect on the value of $b_j$, but it says nothing about whether or not the inclusion of this data point has increased the precision of the associate estimate or not. A measure of how the $i^{\text{th}}$ data point influences overall performance of the model can be encapsulated in the value:

$$COVRATIO_i = \left(\frac{s^2_{-i}}{s^2}\right)^p \left(\frac{1}{1-h_{ii}}\right).$$

Note that this measure attempts to investigate how the $i^{\text{th}}$ data point is affecting the estimate of residual scale. Clearly, points which have a extreme values of $COVRATIO_i$ are points whose inclusion in the analysis is dramatically affecting the precision of predictions and parameter estimates since $s$ is included in the standard errors of all of these quantities. Again, for large samples, a commonly suggested "rule of thumb" is that any point for which $COVRATIO_i > 1 + 3p/n$ or $COVRATIO_i < 1 - 3p/n$ should be investigated as potentially highly influential.

As a final note, we point out that each of the above measures can be seen to be a combination of the leverage and residual values associate with each data point, and thus do not comprise independent pieces of information. In other words, the different diagnostics are designed to be sensitive to influence of a particular type, and the fact that a single data point shows moderate or large values for several of these diagnostic measures should not necessarily be interpreted as additional evidence towards a conclusion of high influence.

*Example 2 (cont'd) - Forestry Data:* We can use *S-Plus* to calculate each of the above diagnostic measures:

```
> pine.reg <- lsfit(cbind(HD,AGE*N,HD/N),MDBH)
> pine.sum <- ls.diag(pine.reg)
> DFFIT <- pine.sum$dfits
> Di <- pine.sum$cooks
> pine.lm <- lm(MDBH ~ HD + I(AGE*N) + I(HD/N))
> inf <- lm.influence(pine.lm)
> sqrtcjj <- (pine.sum$std.err)/pine.sum$std.dev
> DFBETAS <- t(coefficients(pine.lm)-t(inf$coefficients))/
> (inf$sigma%*%t(sqrtcjj))
> s2 <- sum(residuals(pine.lm)^2)/(length(MDBH)-4)
> COVRATIO <- ((inf$sigma^2/s2)^4)/(1-inf$hat)
> options(digits=3)
> cbind(DFFIT,Di,COVRATIO,DFBETAS)
     DFFIT       Di COVRATIO (Intercept)       HD I(AGE * N)  I(HD/N)
 1 -0.0109 3.14e-05    1.705     0.00323  0.00429   -0.00318 -0.00749
 2 -0.5955 8.72e-02    1.196    -0.25416  0.25353   -0.31411 -0.05199
 3  0.7312 1.22e-01    0.854     0.54026  0.30812   -0.44202 -0.40773
 4 -0.8320 1.63e-01    1.041    -0.54898  0.51389   -0.28638 -0.43572
 5  0.0657 1.15e-03    1.485     0.03568 -0.00478   -0.01422  0.00839
 6  0.2807 2.07e-02    1.611     0.24324 -0.14468    0.08804  0.06620
 7 -0.1464 5.62e-03    1.304     0.03017  0.02429   -0.05175 -0.02341
 8  0.2592 1.74e-02    1.316     0.06760  0.18032   -0.16508 -0.20216
 9  0.2085 1.16e-02    2.235    -0.08646 -0.03452    0.10170  0.01760
10 -1.1093 2.35e-01    0.409     0.35882 -0.86001    0.72397  0.65839
11  0.2024 1.08e-02    1.467     0.14379 -0.07047    0.06630 -0.00179
12 -0.1715 7.77e-03    1.528     0.06608  0.10123   -0.12663 -0.11329
13 -0.4192 4.46e-02    1.319    -0.34992  0.05519    0.06957 -0.00506
14 -0.1483 5.79e-03    1.371     0.07996 -0.03954    0.00878  0.01153
15 -0.2398 1.49e-02    1.281     0.06897 -0.13047    0.12481  0.06773
16  0.2979 2.30e-02    1.371     0.05945 -0.03105    0.09089 -0.05147
17 -0.2010 1.06e-02    1.491     0.04153 -0.14759    0.14779  0.09900
18  0.5837 7.30e-02    0.589    -0.01256  0.32614   -0.24293 -0.33039
19  0.2219 1.30e-02    1.701     0.00321  0.19655   -0.19582 -0.17170
20  2.1217 9.14e-01    0.850    -0.99435 -0.92553    0.82465  1.59896
> cutoff <- c(2*sqrt(4/20),qf(0.9,4,16),1+(3*4/20),2/sqrt(20))
> cutoff
[1] 0.894 2.333 1.600 0.447
```

So, a cursory examination shows that there are no real problems (recall that the cut-off values are most appropriate for larger samples, and we have only $n = 20$). There is perhaps a suspicion about the last data point, and if we were to identify this point, it would indeed be the point in the upper right corner of the last added variable plot of the example in the preceding subsection. So, we see that the added variable plots can give us a visual idea of which points will show up as infuential (particularly in regard to $DFBETAS_{j,i}$). Note that if we were to fit the regression without this data point, the analysis results would not change dramatically, except for the parameter estimate of $\beta_3$, and the regression would likely show that the third variable is closer to being important to the model, as the external scientific theory suggested.

*Example 2 (cont'd) - Forestry Example:* We re-fit the regression without the last data point:

```
> pine.lm
Call:
lm(formula = MDBH ~ HD + I(AGE * N) + I(HD/N))

Coefficients:
 (Intercept)         HD    I(AGE * N)  I(HD/N)
    3.235732 0.09740562 -0.0001688538 3.466813

Degrees of freedom:  20 total; 16 residual
Residual standard error:  0.2935898
> anova(pine.lm)
Analysis of Variance Table

Response:  MDBH

Terms added sequentially (first to last)
           Df  Sum of Sq  Mean Sq  F Value     Pr(F)
HD          1  6.207045 6.207045 72.01171 0.0000003
I(AGE * N)  1  2.784561 2.784561 32.30539 0.0000339
I(HD/N)     1  0.014774 0.014774  0.17140 0.6843652
Residuals  16  1.379119 0.086195
> pine.lm1 <- lm(MDBH[-20] ~ HD[-20]+I(AGE[-20]*N[-20])+I(HD[-20]/N[-20]))
> pine.lm1
Call:
lm(formula = MDBH[-20] ~ HD[-20] + I(AGE[-20] * N[-20]) + I(HD[-20]/N[-20]))

Coefficients:
 (Intercept)   HD[-20] I(AGE[-20] * N[-20]) I(HD[-20]/N[-20])
    3.546324 0.1185862        -0.0002138238         -8.597763

Degrees of freedom:  19 total; 15 residual
Residual standard error:  0.2645416
> anova(pine.lm1)
Analysis of Variance Table

Response:  MDBH[-20]

Terms added sequentially (first to last)
                     Df  Sum of Sq  Mean Sq  F Value     Pr(F)
HD[-20]               1  4.486931 4.486931 64.11528 0.0000009
I(AGE[-20] * N[-20])  1  1.976031 1.976031 28.23618 0.0000867
I(HD[-20]/N[-20])     1  0.058883 0.058883  0.84139 0.3735119
Residuals            15  1.049734 0.069982
```

So, the third variable is still not significant (at least with respect to a partial $F$-test), but it is indeed closer to significant ($p$-value down to 0.37 from 0.68 for full regression). Also, the sign of the third coefficient has changed, which we would have expected from the added variable plot.

v. Multicollinearity

In the Forestry data example of the preceding section, we saw that the second added variable plot showed us a negatively sloping relationship, while the ordinary plot of the response versus the second predictor indicated a positively sloping relationship (albeit a somewhat weak one). How can such a thing occur? The answer is *multicollinearity*. The term multicollinearity refers to a condition in the *predictor* variables. Specifically, it relates to the degree of interrelation among the predictor values. Thus, as a first diagnostic test for whether there is multicollinearity present in the data, we calculate the correlation matrix of the predictors, $R = (r_{ij})$, where

$$r_{ij} = Corr(X_i, X_j) = \frac{\sum_{k=1}^{n}(X_{ki} - \overline{x}_i)(X_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(X_{ki} - \overline{x}_i)^2 \sum_{k=1}^{n}(X_{kj} - \overline{x}_j)^2}},$$

where $X_i$ is the $i^{\text{th}}$ column of the design matrix and $\overline{x}_i$ is just the average of all the values of the $i^{\text{th}}$ predictor. Note that the term *correlation* matrix is a bit of a misnomer, since we are considering the predictor values to be fixed (i.e., non-random), however, the value $r_{ij}$ still measures the linear relationship between the values of the two predictor variables $x_i$ and $x_j$. If any of the off-diagonal $r_{ij}$'s (i.e., those for which $i \neq j$) are large, then we know that the associated two predictor variables are highly correlated and this results in multicollinearity. However, this is not the only way that multicollinearity can arise. Indeed, the word *multi*collinearity implies that the problem may arise because several of the predictors are strongly linearly related to each other. Generally speaking, a set of predictor values are said to be multicollinear if there exist any constants $c_1, \ldots, c_p$ (not all equal to zero) such that

$$\sum_{j=1}^{p} c_j X_j \approx 0.$$

Why is multicollinearity a problem? There are various answers to this question, and we shall investigate several. First, remember that the interpretation of the parameters in a multiple regression is a rate of change of the response with respect to a particular predictor while the rest of the predictors are held constant. If we have predictor values which are highly interrelated, then it is not possible to "hold the other predictors constant" within our dataset (whether or not this is possible in the population), and therefore we cannot expect to get reliable estimates of the parameter values from such data.

Next, recall the centered regression model which we introduced for the simple regression model. We can do the same thing for a multiple regression model, noting that $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$ can be re-written as

$$\begin{aligned} Y &= \beta_0^\star + \beta_1(x_1 - \overline{x}_1) + \ldots + \beta_k(x_k - \overline{x}_k) + \epsilon \\ &= \beta_0^\star + \beta_1 x_1^\star + \ldots + \beta_k x_k^\star \\ &= \beta_0^\star + X^\star \beta + \epsilon, \end{aligned}$$

where the intercept parameter is suitably transformed to $\beta_0^\star = \beta_0 + \beta_1 \overline{x}_1 + \ldots + \beta_k \overline{x}_k$ and $X^\star$ is the centered design matrix, without an initial column of ones for the intercept term. With this restructuring, it can easily be seen that

$$R = (X^\star)^T X^\star.$$

So, to estimate the $\beta$ vector for this regression, we need to invert the matrix $R$, since $b = R^{-1}(X^{\star})^T Y$. Suppose that we are in an extreme case, where the correlation between two of the predictors is exactly one. In such a case, the $R$ matrix will not be invertible! In other words, we cannot estimate the parameters at all. The idea here is that once the first of the two perfectly correlated predictors is put into the model, the least-squares fit will calculate an estimate of the associated parameter value, and then when the second predictor is incorporated, it adds nothing new to the model and therefore should be assigned an estimated parameter value of zero. On the other hand, if the entry order of the two predictors were reversed, then the first predictor would now be the one with the parameter estimate of zero. Thus, the least-squares gets confused by completely collinear predictors, and the non-invertibility of $R$ means that no parameter estimates can be calculated.

Of course, exact collinearity is rarely a problem (though we will have to deal with it in the next chapter when we discuss analysis of variance models and indicator variables). However, even though the $R$ matrix will be invertible as long as there is no exact collinearity (and thus estimates will be calculable), if the predictors are *nearly* collinear (e.g., if there are any off-diagonal elements of the $R$ matrix which are very near unity), then the parameter estimates will be highly unstable and imprecise (the idea here being similar to the arithmetic notion that 0 is not invertible, and while very small numbers are indeed invertible, their inverses are extremely large, and thus difficult to handle).

This last notion leads to the quantitative measure of the difficulties caused by multicollinearity, called the *variance inflation factors* or $VIF_i$'s. Before we directly define the $VIF_i$'s, we will examine a rather contrived, but nonetheless informative example:

*Example 6 - Multicollinearity:* Suppose that we are going to conduct a study of the relationship of a response variable, $Y$, and two predictors, $x_1$ and $x_2$, and we know that the assumptions of the linear model hold, so that $Y = X\beta + \epsilon$ and $Var(\epsilon) = \sigma^2 I$. We are asked to choose between two possible sets of predictor values:

Predictor set 1:

| $x_1$ | 10 | 10 | 10 | 10 | 15 | 15 | 15 | 15 |
|---|---|---|---|---|---|---|---|---|
| $x_2$ | 10 | 10 | 15 | 15 | 10 | 10 | 15 | 15 |

Predictor set 2:

| $x_1$ | 10.0 | 11.0 | 11.9 | 12.7 | 13.3 | 14.2 | 14.7 | 15.0 |
|---|---|---|---|---|---|---|---|---|
| $x_2$ | 10.0 | 11.4 | 12.2 | 12.5 | 13.2 | 13.9 | 14.4 | 15.0 |

So, if we were to estimate the parameters from a regression using each of these datasets, we would find that for the first set of predictor values, $r_{12} = 0$ so that:

$$(X^{\star})^T X^{\star} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Longrightarrow \qquad \left\{(X^{\star})^T X^{\star}\right\}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

while for the second predictor set, $r_{12} = 0.99215$ so that:

$$(X^{\star})^T X^{\star} = \begin{pmatrix} 1 & 0.99215 \\ 0.99215 & 1 \end{pmatrix} \qquad \Longrightarrow \qquad \left\{(X^{\star})^T X^{\star}\right\}^{-1} = \begin{pmatrix} 63.94 & -63.44 \\ -63.44 & 63.94 \end{pmatrix}.$$

Thus, we can see that the least-squares parameter estimates we get from the first set of predictors will have variances:

$$Var(b_1) = Var(b_2) = \sigma^2,$$

while the parameter estimates from the second set of predictor values will have variances:

$$Var(b_1) = Var(b_2) = 63.94\sigma^2.$$

Note that this inflation of the variance has nothing to do with the response variable, or with the intrinsic variability of the errors, $\epsilon$, in the population, it is purely an artifact of the correlation structure of the predictors.

The value 63.94 in the previous example is the *variance inflation factor*, $VIF_1 = VIF_2$, since it shows the proportion by which the variance has been increased over a set of *orthogonal* predictor values (i.e., predictor values for which $R = I$). In general, the variance inflation factor associated with the $i^{\text{th}}$ predictor is:

$$VIF_i = \frac{1}{1 - R_i^2},$$

where $R_i^2$ is the coefficient of determination for a regression of the predictor $x_i$ as the response variable on the rest of the predictors, $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$. So, the stronger the linear relationship between $x_i$ and the rest of the predictors, the larger the value of $R_i^2$ and thus the larger the value of $VIF_i$. Therefore, the parameter estimates associated with predictor variables with large $VIF_i$'s will have very low precision, which in turn implies that any quantities which rely heavily on these parameter estimates (e.g., various predicted values and general linear hypotheses) may also tend to be highly imprecise. To see why this is the case, we again make use of a contrived, but enlightening example. Suppose that the true underlying model relating a response to two predictors is $Y = 2x_1 + 3x_2 + \epsilon$, and that we select a set of predictor values such that $x_1 = x_2$ for all data points. For such a set of predictors, we note that

$$\begin{aligned} Y &= 2x_1 + 3x_2 + \epsilon \\ &= 2x_1 + 3x_1 + \epsilon \\ &= 5x_1 + 0x_2 + \epsilon \\ &= 2x_2 + 3x_2 + \epsilon \\ &= 0x_1 + 5x_2 + \epsilon \\ &= 7x_1 - 2x_2 + \epsilon \\ &\vdots \end{aligned}$$

So we see that we have no real hope of accurately predicting the parameters in such a situation (though, if we had chosen a better set of predictor values we could certainly estimate the parameters, since the above equations are no longer true without the assumption that $x_1 = x_2$ for all the data). Indeed, note that we even have a situation where the sign of the coefficient is the reverse of what it should be (note however, that in all the above equations, the two coefficients sum to 5, and thus we can be reasonably precise in our estimation of the value $\beta_1 + \beta_2$, which should come as no surprise, since this situation truly amounts to a simple linear regression in which the slope parameter is 5). Of course, we will rarely run into a case where $x_1 = x_2$ for every data point, but if $x_1 \approx x_2$ for most of the data then we will still have the above problem approximately. Clearly, we must be extremely careful of our interpretation of parameter estimate values in the presence of multicollinearity.

If we find that multicollinearity is a problem, there are several approaches to handling the situation. There are, of course, more advanced procedures, such as *ridge regression* and *principle* *components regression*, designed to reduce the effects of multicollinearity. However, it is often the case that the problem can be solved by simply removing one or a few of the predictors. Which predictor to remove can often be a difficult decision, however. The removal should ideally be based on both statistical and non-statistical grounds. Generally, a good place to start is to remove the variable with the highest $VIF_i$ and see if this solves the problem or else the predictor with the weakest simple regression with the response, however, external considerations should be taken into account, as these variables may turn out to be extremely relevant in the context in which the data was gathered.

*Example 5 (cont'd) - Body Fat Data:* Recall the body fat dataset measured on 20 women. For this dataset, the correlation between the predictors is seen to be:

```
> cor(cbind(Triceps,Thigh,Midarm))
          Triceps      Thigh     Midarm
Triceps 1.0000000 0.92384249 0.45777723
  Thigh 0.9238425 0.99999988 0.08466751
 Midarm 0.4577772 0.08466751 1.00000000
```

and thus the $VIF_i$'s are:

```
> diag(solve(cor(cbind(Triceps,Thigh,Midarm))))
[1] 708.8918 564.3823 104.6132
```

Which can be confirmed using:

```
> R2i <- summary(lm(Triceps ~ Thigh + Midarm))$r.squared
> 1/(1-R2i)
[1] 708.8429
```

If we fit the model using all three predictors, we see that the standard errors of the parameters estimates are quite large (in comparison to the estimate values themselves), and that the last variable entered into the model is not significant.

```
> lsfit(cbind(Triceps,Thigh,Midarm),BodyFat)$coef
 Intercept  Triceps     Thigh   Midarm
  117.0847 4.334092 -2.856848 -2.18606
> ls.diag(lsfit(cbind(Triceps,Thigh,Midarm),BodyFat))$std.err
             [,1]
Intercept 99.782403
  Triceps  3.015511
    Thigh  2.582015
   Midarm  1.595499
> anova(lm(BodyFat ~ Triceps + Thigh + Midarm))
Analysis of Variance Table

Response:  BodyFat

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
Triceps    1  352.2698 352.2698 57.27680 0.0000011
Thigh      1   33.1689  33.1689  5.39305 0.0337319
Midarm     1   11.5459  11.5459  1.87729 0.1895628
Residuals 16   98.4049   6.1503
```

Removing the predictor with the largest $VIF_i$ and then refitting the model yields:

```
> lsfit(cbind(Thigh,Midarm),BodyFat)$coef
```

```
 Intercept     Thigh     Midarm
 -25.99695 0.8508817 0.09602947
> ls.diag(lsfit(cbind(Thigh,Midarm),BodyFat))$std.err
              [,1]
Intercept 6.9973208
    Thigh 0.1124482
   Midarm 0.1613927
> anova(lm(BodyFat ~ Thigh + Midarm))
Analysis of Variance Table

Response:  BodyFat

Terms added sequentially (first to last)
          Df  Sum of Sq  Mean Sq  F Value     Pr(F)
Thigh      1  381.9658 381.9658 58.44147 0.0000007
Midarm     1    2.3139   2.3139  0.35403 0.5596775
Residuals 17  111.1098   6.5359
> diag(solve(cor(cbind(Thigh,Midarm))))
[1] 1.00722 1.00722
```

So, we see that we have solved the multicollinearity problem, however, the last variable in the model (`Midarm`) is still not significant. It turns out that this predictor is also the one with the weakest simple regression relationship, however, dropping it from the model and retaining the other two predictors will clearly not solve the multicollinearity problem. Alternatively, we could try dropping the thigh circumference variable from the original three variable model, since the largest correlation in the $R$ matrix was between thigh circumference and triceps skinfold:

```
> lsfit(cbind(Triceps,Midarm),BodyFat)$coef
 Intercept  Triceps    Midarm
  6.791627 1.000585 -0.431442
> ls.diag(lsfit(cbind(Triceps,Midarm),BodyFat))$std.err
              [,1]
Intercept 4.4882871
  Triceps 0.1282321
   Midarm 0.1766156

Analysis of Variance Table

Response:  BodyFat

Terms added sequentially (first to last)
          Df  Sum of Sq  Mean Sq  F Value      Pr(F)
Triceps    1  352.2698 352.2698 56.53121 0.00000084
Midarm     1   37.1855  37.1855  5.96742 0.02578645
Residuals 17  105.9342   6.2314
> diag(solve(cor(cbind(Triceps,Midarm))))
[1] 1.265118 1.265118
```

which again shows that the multicollinearity has been fixed, and now both the predictors are significant. However, we now have a model which uses only measurements on the arm of the women, and this seems rather unsound from a medical and biological perspective. We will discuss the ideas of model selection more fully in the next section.

## VII. Model Selection

Armed with our model fitting and diagnostic tools, we will now embark on a discussion of the proper methods for comparing regression models. Typically, we will be in the situation of having gathered a dataset in which we have a large collection of possible predictor variables. Thus, there will be a large selection of *candidate models* for the explanation of the variability in the response variable, and we desire to choose one, or possibly a few, “good” models for the response. Of course, what constitutes a “good” model will generally depend on the context in which the data upon which it is to be based was gathered as well as the model's intended use. For instance, we may want to:

· *Learning about some aspect of the system from which the data are taken*: We may be interested in the sign of a particular coefficient, which will tell us the nature of the (adjusted) relationship between the response and a particular predictor. Alternatively, we may be interested in the location within the predictor space of a local maximum or minimum value of the response, for example in an industrial setting, where the particular input levels which maximize the output of a certain production process are sought.

· *Variable screening:* In the absence of strong underlying scientific theory, we might wish to determine which among a large collection of possible predictors appear related to the response variable under study. Of course, we must always remember that a statistical relationship is not, in and of itself, proof of a true causative link between the variables.

· *Prediction:* We might only wish to choose a model which will provide useful and accurate future predictions. To this end, we must recall that overly complicated models will generally not perform well in this endeavor, particularly in cases where we wish to predict for values somewhat outside the range of the observed data values. Thus, parsimony is a highly desirable feature, and we must take great care not to “overfit” our data, making the resulting model nothing more than a complicated restatement of the observed response values, instead of a useful tool for prediction at values other than those actually observed.

This last task of prediction, highlights a key concept in model selection. Namely, that we need not locate the “true” model to satisfy our goals, in fact, we may never be able to locate the “true” model at all. Indeed, even the concept of a “true” model which generated our data is often simply a useful fiction to facilitate our general goal of an understanding of the relationships in the system or population from which our data was originally drawn.

For each of the above tasks, we must take into account not only statistical considerations (e.g., multicollinearity among the predictors), but also external scientific information which may lead us in the correct direction (e.g., ensuring that certain known important predictor variables are definitely retained in our model). Too heavy a reliance on either one of these aspects can lead to poor models. The statistician must yield to scientific evidence in the case of borderline statistical evidence, while the scientist must make sure that the data actually support the proposed theory, and that a “pet” theory is not preferred when it is in strong contradiction to the observed statistical evidence.

i. Standard Model Comparison Criteria.

We have already seen several simple methods of measuring the degree to which a model accurately fits the data. First, we saw how to formally test the null hypothesis $H_0 : \beta_{(1)} = 0$, indicating that a subset of the parameters were equal to zero, and thus their associated parameters might be reasonably removed from the model. This approach was based on the sequential sums of squares, however, we saw that it could equivalently be considered as a comparison of the residual sums of squares from the "full" and the "reduced" models (i.e., the model including all the predictors and the model excluding those predictors associated with the parameters in $\beta_{(1)}$). Thus, one way of comparing two different models is to compare the $MSE$ from each of them and choose the model which has the smaller value. Recall that if a model is "underspecified"; that is, we fit the model

$$Y = X_{(2)}\beta_{(2)} + \epsilon,$$

when the true model is:

$$Y = X_{(1)}\beta_{(1)} + X_{(2)}\beta_{(2)} + \epsilon,$$

then the $MSE$ of our regression will be an over-estimate of the true value of $\sigma^2$. In other words, if we have fitted a model which does not contain all the relevant predictors, then our estimate of the underlying error variance will have a positive bias. Suppose we fit a "full" regression containing $P$ parameters and a "reduced" regression containing $p < P$ parameters, and calculate the $MSE$'s for these two regressions as $s_P^2$ and $s_p^2$, respectively. If there were a large underspefication associated with the smaller model, then we would expect to see $s_P^2 < s_p^2$, since the latter estimate would have positive bias in such a case, and we would tend to prefer the full model as more appropriate. Of course, we should recall that as the number of parameters, $P$, in the full model nears the sample size, $n$, then $s_P^2$ will necessarily decrease down towards zero, regardless of the appropriateness of the model fit. However, if the two estimates of $\sigma^2$ are nearly equal, or if $s_P^2 > s_p^2$, then we can conclude that there is likely to be no underspecification, or that the "reduced" model is the more appropriate of the two models. Of course, all of the preceding discussion assumes that we are comparing two *nested* models (i.e., two models for which the predictors used in one form a proper subset of the predictors used in the other). The same comparisons are possible between two non-nested models, however, the justification for choosing the model with the smaller mean square error is less clear (unless, of course, the two models under comparison have the same number of parameters).

Another measure of the overall association between a response and a set of predictors was the coefficient of determinaton,

$$R^2 = 1 - \frac{SSE}{SST},$$

which measured the proportion of response variation "explained" by the regression, and could be used to compare nested and non-nested models alike. However, as pointed out at the time, this measure is not very useful for comparing across models, since the addition of any predictor to the model (whether or not it has any relationship to the response) will necessarily increase the value of $R^2$, and thus will tend to lead us towards overly complicated models (e.g., if we fit an $n^{\text{th}}$ degree polynomial to a dataset of $n$ pairs $(x_i, Y_i)$, the $R^2$ value will be exactly 1). The problem with the coefficient of determination as a comparitive tool is that the sums of squares used in its computation do not adequately take account of the number of parameters in the model. The *adjusted* $R^2$, or $R_a^2$, for a particular regression is defined by replacing the sums of squares in the $R^2$ definition with their corresponding mean squares, thus accounting for the number of parameters in the model:

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{s^2(n-1)}{SST}.$$

Unlike the unadjusted coefficient of determination, $R_a^2$ can, and often does, decrease when additional predictors are added to the model, and thus the choice of the model with the highest value of $R_a^2$ is often suggested as a simple model selection methodology. Clearly, however, there is a strong connection between the two measures which is demonstrated by the relationship (see tutorial exercises):

$$R_a^2 = \left(\frac{n-1}{n-p}\right) R^2 - \frac{p-1}{n-p}.$$

ii. Cross-Validation and the PRESS Statistic

A more complicated methodology for choosing between models can be developed by using the data itself to examine how well the model is performing, at least in terms of its predictive capability. Initially, we might think that simply examining the ordinary residuals, $e_i$, will indicate how well our model will perform. The problem with this approach is that the two components from which the residual is calculated, $Y_i$ and $\hat{Y}_i$, are not "independent" of each other. One general solution to this problem is to *split* our dataset, $S = M \cup V$ (with sample size $n$), into two groups, using the first (the *fitting sample* or *modelling set*, $M$, containing $n_1$ data points) to actually fit various regression models and then seeing how well they predict the values of the second (the *validation sample* or *validation set*, $V$, containing the remaining $n_2 = n - n_1$ data points).

If we fit candidate regression models on the modelling set, we can then evaluate the predictive performance by calculating a measure of discrepancy on the validation set, such as:

$$\sum_{i:\, x_i \in V} (Y_i - \hat{Y}_i)^2 \qquad \text{or} \qquad \sum_{i:\, x_i \in V} |Y_i - \hat{Y}_i|,$$

where the summation is over the indices for those data points in the validation set, $V$, and $\hat{Y}_i = x_i^T b_M$ where $b_M$ is the least-squares estimator of the parameters of the model fit on the modelling set, $M$. We might then choose the model which has the smallest discrepancy measure as our preferred model.

Of course, there are still some difficulties with which to be dealt. Specifically, we must decide which and how many data points to use in the modelling and validation sets. If the data is time-dependent, then it is often suggested that the most recent data points should be used in the validation set, as these data points will most closely correspond to the desired future predictions of the model. This example is a specific instance of the situation in which the model builder has a specific region within the predictor space where future predictions are desired. In such cases, we clearly should select those data points nearest to this region as our validation set. As to the relative sizes of the two sets, there is no generally accepted rule of thumb. One very common liberal suggestion is to use roughly equal sizes for the two sets. However, it is clear that there should be enough data points in the fitting sample to ensure adequate information for a regression analysis to make a reliable fit to the data. Moreover, whatever the sizes, and whichever candidate model or models are chosen, the final regression estimates and predicted values should be based on the *entire* data set. In other words, data splitting and cross-validation should be thought of as a mechanism for exploring the performance of various candidate models, but the model or models which are finally adopted should be fit using all the available information from the dataset.

In smaller datasets, data splitting may not be practical. To overcome this problem, or simply to alleviate the problem of deciding how to split the dataset in cases where splitting is feasible, we can construct a measure of performance for our candidate models which uses the conceptual notion behind cross-validation, but does not actually require splitting of the data. We do so through the use of the *PRESS* residuals,

$$e_{i,-i} = Y_i - \hat{Y}_{i,-i} = Y_i - x_i^T b_{-i} = \frac{e_i}{1 - h_{ii}}.$$

Note that the $i^{\text{th}}$ *PRESS* residual is the measure of discrepancy we would arrive at if we performed a data splitting cross-validation where the validation set consisted of solely the $i^{\text{th}}$ data point. The *PRESS* statistic, sometimes denoted as $PRESS_p$, is defined as:

$$PRESS_p = \sum_{i=1}^{n} e_{i,-i}^2 = \sum_{i=1}^{n} \left( \frac{e_i}{1 - h_{ii}} \right)^2.$$

We might then favor the model or models with the smallest values of $PRESS_p$. In addition, note that the *PRESS* statistic is just a sum of squares for the *PRESS* residuals. As such, we might expect that its value would be close to the $SSE$ if the candidate model under investigation was appropriate for the data. Thus, a comparison between the $PRESS_p$ and $SSE$ values can yield useful information regarding the appropriateness of a particular model.

iii. Underfitting, Overfitting and Mallows' $C_p$ Statistic

We have already discussed the fact that if we fit an "underspecified" model to our data, we will incur a bias in our estimate of the error variance, $\sigma^2$. Specifically, if we fit the model $Y = X_1\beta_{(1)} + \epsilon$ when the true mean of the response variable is $E(Y) = X_1\beta_{(1)} + X_2\beta_{(2)}$, then it can be shown that the expected value of the $MSE$ is:

$$E(s_p^2) = \sigma^2 + \frac{1}{n-p}\sum_{i=1}^{n}\{E(\hat{Y}_i) - E(Y_i)\}^2 = \sigma^2 + \frac{1}{n-p}\sum_{i=1}^{n}\{Bias(\hat{Y}_i)\}^2,$$

where $p$ is the number of parameters in the underspecified model and $\hat{Y}_i$ is the fitted value for the $i^{\text{th}}$ data point from the underspecified model, while the expectations are taken with respect to the true underlying model. In fact, we will incur a bias in other estimates as well. For instance, the least-squares estimates from the underspecified regression have expectation:

$$\begin{aligned} E(b_{(1)}) &= E\{(X_1^T X_1)^{-1} X_1^T Y\} \\ &= (X_1^T X_1)^{-1} X_1^T E(Y) \\ &= (X_1^T X_1)^{-1} X_1^T (X_1\beta_{(1)} + X_2\beta_{(2)}) \\ &= \beta_{(1)} + (X_1^T X_1)^{-1} X_1^T X_2 \beta_{(2)} \\ &= \beta_{(1)} + A\beta_{(2)}, \end{aligned}$$

where $A = (X_1^T X_1)^{-1} X_1^T X_2$ is sometimes referred to as the *alias matrix*. Thus, the least-squares estimates from an underspecified model will be biased, which implies that any predictions will also be biased. Therefore, we obviously want to avoid underspecifying our model. On the other hand, if we "overspecify" our model, then typically:

- The variances of the least-squares estimates are inflated. In other words, suppose that the true model for our response is $Y = X\beta + \epsilon$ and that $b$ represents the least-squares estimates from a regression of $Y$ using the design matrix $X$. If we were to fit the model $Y = X\beta + \beta_{k+1}x_{k+1} + \epsilon$, then the new least-squares estimator, $b^\star$, would have the property:

$$Var(b_j^\star) \geq Var(b_j) \qquad \text{for } (j = 1, \ldots, k).$$

- The variance of predicted values will also be inflated. In other words, at a particular set of predictor values $x_0 = (x_{1,0}, \ldots, x_{k,0}, x_{k+1,0})$, we would define the predicted values from the two regressions noted above by

$$\hat{Y}(x_0) = \sum_{i=1}^{k} b_i x_{i,0} \qquad \text{and} \qquad \hat{Y}^\star(x_0) \sum_{i=1}^{k+1} b_i^\star x_{i,0},$$

respectively, and the variances of these predictions would satisfy:

$$Var\{\hat{Y}(x_0)\} \leq Var\{\hat{Y}^\star(x_0)\}.$$

- The $MSE$ from the overspecified model will still be an unbiased estimate of $\sigma^2$, so that

$$E(s_{k+1}^2) = \sigma^2.$$

However, it will be a less precise estimate than $s_k^2$, the $MSE$ from the correct model. This is because it is now an estimator based on one fewer degree of freedom.

Generally speaking then, we see that model selection amounts to finding an appropriate compromise between the bias of an underspecified (or *underfit*) and the inflated variances of an overspecified (or *overfit*) model. To help decide on a reasonable compromise between these two poles, we need a criterion which is sensitive to the discrepancies inherent in both under- and overfitting. As a start, we might consider a measure based on how well the particular model under investigation predicts, such as the *mean squared error of prediction* at a point $x_0$:

$$\begin{aligned} MSE\{\hat{Y}(x_0)\} &= Var\{\hat{Y}(x_0)\} + \left[E\{\hat{Y}(x_0)\} - E\{Y(x_0)\}\right]^2 \\ &= Var\{\hat{Y}(x_0)\} + \left[Bias\{\hat{Y}(x_0)\}\right]^2, \end{aligned}$$

which clearly incorporates both of the issues at hand. Unfortunately, this measure depends upon the particular values, $x_0$, we choose for the predictor variables. To overcome this problem, we could look at the scaled mean squared errors at each of the fitted values:

$$\sum_{i=1}^{n} \frac{MSE\{\hat{Y}_i\}}{\sigma^2} = \sum_{i=1}^{n} \frac{Var\{\hat{Y}_i\} + \left[Bias\{\hat{Y}_i\}\right]^2}{\sigma^2}.$$

Now, the above quantity will not reflect the interpolation or extrapolation capabilities of the candidate model, however, it will give us a quantity that can be used to obtain a workable balance between bias and inflated variances. A little algebra will show that if the candidate model has $p$ parameters, then

$$\sum_{i=1}^{n} \frac{Var\{\hat{Y}_i\}}{\sigma^2} = p,$$

which follows using a nearly identical argument to that used to demonstrate that the sum of the leverages, $\sum_{i=1}^{n} h_{ii}$, is equal to the number of parameters in the model (which is also the number of columns in the design matrix $X_{(1)}$). Similarly, using the previous result regarding the expectation of the $MSE$ of an underspecified model, we can see that

$$\begin{aligned} \sum_{i=1}^{n} \frac{\left[Bias\{\hat{Y}(x_0)\}\right]^2}{\sigma^2} &= \frac{(n-p)(E(s_p^2) - \sigma^2)}{\sigma^2} \\ &\approx \frac{(n-p)(s_p^2 - \sigma^2)}{\sigma^2}, \end{aligned}$$

where $s_p^2$ is the $MSE$ from the candidate model under investigation. So, if we had an "independent" estimate of $\sigma^2$, say $\hat{\sigma}^2$, then we could estimate the sum of the scaled mean squared errors of the fitted values using *Mallows' $C_p$* statistic:

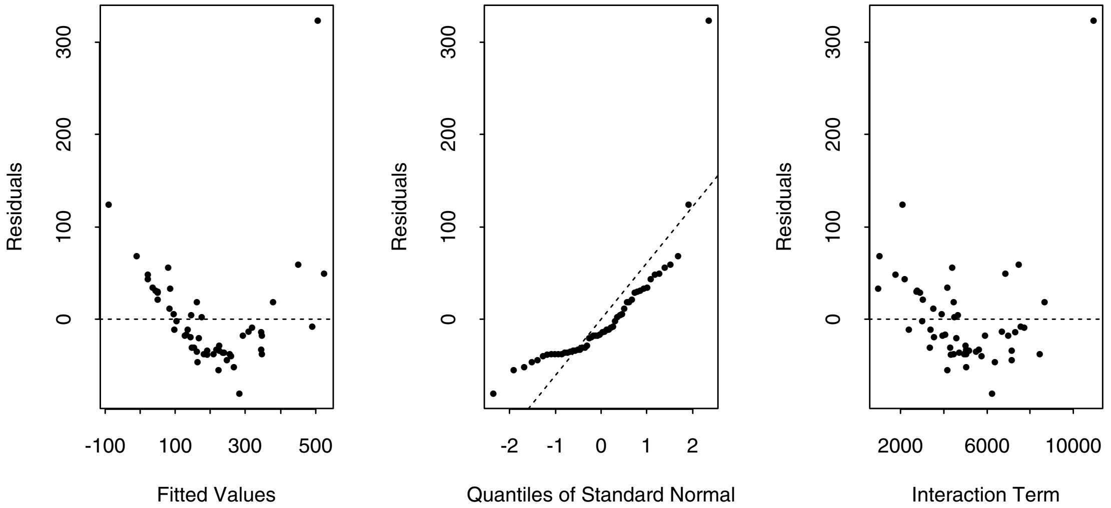
$$C_p = p + \frac{(n-p)(s^2 - \hat{\sigma}^2)}{\hat{\sigma}^2}.$$

We might then favor the model or models with the smallest values of $C_p$. Also, note that a potentially reasonable norm by which to judge the size of the $C_p$ value can be based on the fact that a model with no bias would have $C_p = p$. Of course, we don't generally have an "independent" estimate of the true error variance. However, it is often suggested that a reasonable choice for $\hat{\sigma}^2$ is the $MSE$ from the "full" model, $s_P^2$, derived using all the predictors under consideration (recall that the $MSE$ from an overspecified model is still an unbiased estimate of $\sigma^2$, although it is somewhat less precise than the $MSE$ calculated from the true model). Notice, however, that this choice for $\hat{\sigma}^2$ means that the $C_p$ value for the "full" model will necessarily be exactly equal to its total number of parameters. Frequently, a plot of the $C_p$ values versus the associated number of parameters, $p$, for each of the candidate models is used to visually display the information contained in this selection criterion. For reference, the line $C_p = p$ is also generally superimposed on the plot.

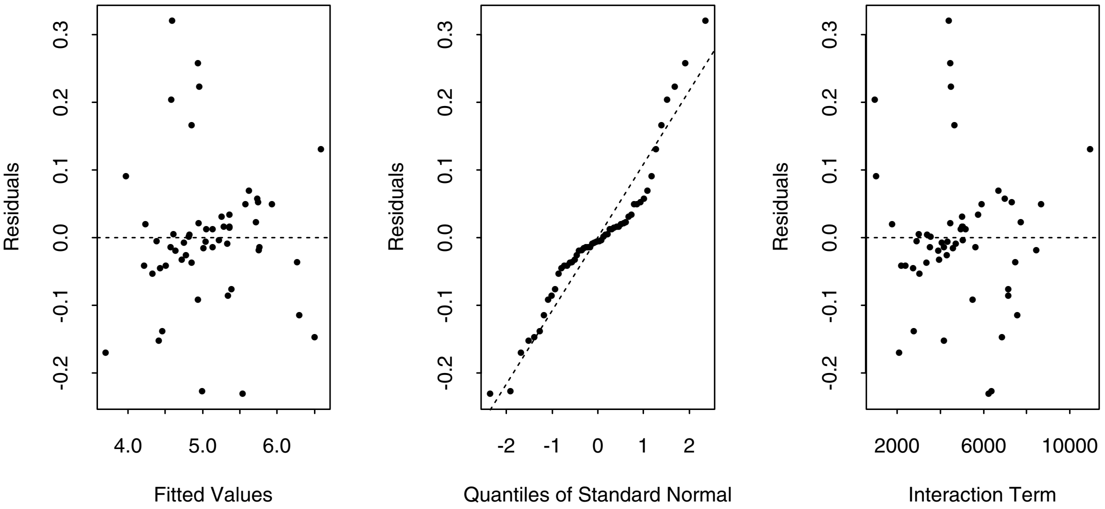
*Example 7 - Surgical Data:* A particular hospital conducted a study to investigate the survival time in patients undergoing a certain type of liver operation. A random sample of 54 patients was selected and from each patient's medical record, the following information was gathered:

$x_1$ - A blood clotting score;
$x_2$ - A prognostic index (calculated incorporating age);
$x_3$ - An enzyme function test score; and,
$x_4$ - A liver function test score.

Initially, a full model of the patients' survival times was fit, including all four predictor variables. The residuals versus fitted value and normal q-q plots are shown below:



In addition, a plot of the residuals versus the product $x_2 \cdot x_3$ is also shown, indicating that there is a potential relationship between the response and this product (generally referred to as an *interaction*) which is not accounted for by the other predictors. Clearly, the residual and normal q-q plots indicate non-normality and non-linearity in the data, and a logarithmic transformation is suggested. The corresponding residual and normal q-q plot are shown below, along with the plot of the new residuals versus the product $x_2 \cdot x_3$:



The residual plot now looks much better, however, it still appears that there may be some question about the normality of the residuals, as the q-q plot displays an inverted-"S" shape characteristic of a heavy-tailed distribution. There is some cause for concern here, but typically, skewness presents a bigger problem than does heavy or light tails, and so we will proceed with our analysis on this scale. Also, it appears that the relationship with the interaction term has disappeared. Nonetheless, its initial appearance suggests that there may still be the potential for some curvature in the relationship between the logarithm of the response and the second and third predictors, so we shall still include the terms $x_2^2$, $x_3^2$ and $x_2 \cdot x_3$ as potential predictors in our model selection procedures (and the inclusion of these terms may help with the heavy-tailed appearance of the residuals as well).

We first calculate the $s_p^2$'s, $R_a^2$'s, $PRESS_p$'s and $C_p$'s for all of the models, using *S-Plus*:

```
> selcrit <- function(preds,resp,sig2) {
+   regout <- lsfit(preds,resp)
+   regdiag <- ls.diag(regout)
+   regsum <- ls.print(regout,print.it=F)
+   mse <- regdiag$std.dev^2
+   R2 <- regsum$summary[2]
+   nn <- regsum$summary[3]
+   np <- regsum$summary[5]+1
+   adjr2 <- (((nn-1)*R2)/(nn-np))-((np-1)/(nn-np))
+   levs <- regdiag$hat
+   delres <- regout$resid/(1-levs)
+   pressp <- sum(delres^2)
+   Cp <- np + (((nn-np)*(mse-sig2))/sig2)
+   tmp <- cbind(mse,adjr2,pressp,Cp)
+   dimnames(tmp) <- list(NULL,c("MSE","R2a","PRESSp","Cp"))
+   tmp
+ }
> modsel <- function(predind,dta) {
+   cols <- ncol(dta)
+   preds <- dta[,-cols]
+   resp <- dta[,cols]
+   sig2 <- ls.diag(lsfit(preds,resp))$std.dev^2
+   selcrit(preds[,predind],resp,sig2)
+ }
> srg.df <- as.data.frame(surgery)
> attach(srg.df)
> names(srg.df)
> [1] "clot" "prog" "enzyme" "liver" "survival"
> modlst <- list(c(1),c(2),c(3),c(4),
+   c(1,2),c(1,3),c(1,4),c(2,3),c(2,4),c(2,5),c(3,4),c(3,7),
+   c(1,2,3),c(1,2,4),c(1,2,5),c(1,3,4),c(1,3,7),c(2,3,4),c(2,3,5),
+     c(2,3,6),c(2,3,7),c(2,4,5),c(3,4,7),
+   c(1,2,3,4),c(1,2,3,5),c(1,2,3,6),c(1,2,3,7),c(1,2,4,5),c(1,3,4,7),
+     c(2,3,4,5),c(2,3,4,6),c(2,3,4,7),c(2,3,5,6),c(2,3,5,7),c(2,3,6,7),
+   c(1,2,3,4,5),c(1,2,3,4,6),c(1,2,3,4,7),c(1,2,3,5,6),c(1,2,3,5,7),
+     c(1,2,3,6,7),c(2,3,4,5,6),c(2,3,4,5,7),c(2,3,4,6,7),c(2,3,5,6,7),
+   c(1,2,3,4,5,6),c(1,2,3,4,5,7),c(1,2,3,4,6,7),c(1,2,3,5,6,7),
+     c(2,3,4,5,6,7),
+   c(1,2,3,4,5,6,7))
> srg.all <- cbind(clot,prog,enzyme,liver,prog^2,prog*enzyme,enzyme^2,
+   log(survival))
> srg.mod.sel <- lapply(modlst,modsel,dta=srg.all)
```

The user-defined *S-Plus* function `modsel()` uses an input data matrix and vector of index values to indicate the subset of the predictors on which we wish to fit a model and calculate selection criteria. It then passes the appropriate design matrix and response variable, along with the overall $MSE$ from the full model for use in calculating $C_p$, to the user-defined *S-Plus* function `selcrit()` which calculates the $MSE$, $R_a^2$, $PRESS_p$ and $C_p$ values for the input model (see tutorial exercises). In order to create the selection criteria values for a collection of different models, we create a list of the desired subsets (`modlst` in the above command lines) and then use the *S-Plus* function `lapply()` to apply the `modsel` function to each element of our model list. The results for the surgery data are presented below:

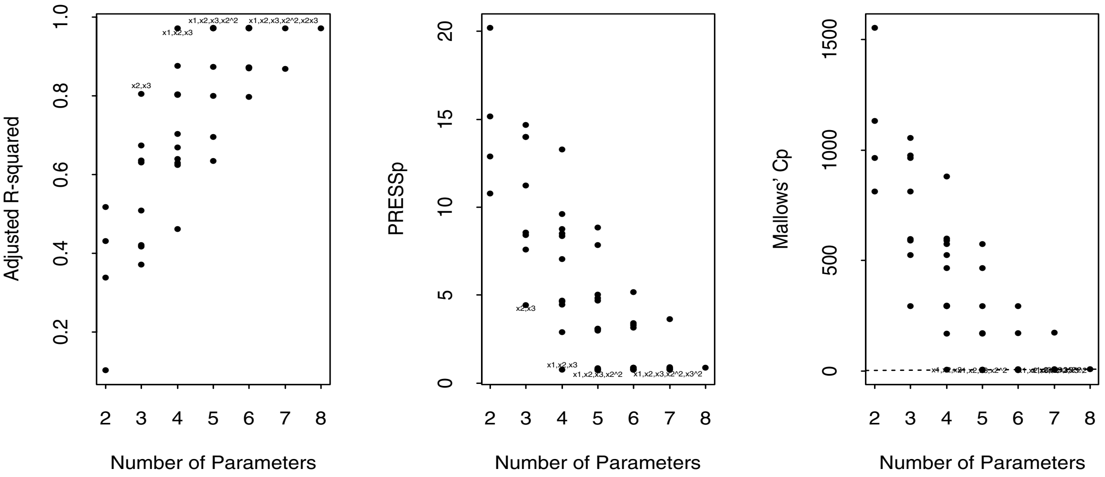
| Model | $p$ | $s_p^2$ | $R_a^2$ | $PRESS_p$ | $C_p$ |
|---|---|---|---|---|---|
| $x_1$ | 2 | 0.3564 | 0.103 | 20.191 | 1552.436 |
| $x_2$ | 2 | 0.2627 | 0.339 | 15.177 | 1130.854 |
| $x_3$ | 2 | 0.2259 | 0.432 | 12.868 | 965.469 |
| $x_4$ | 2 | 0.1914 | 0.518 | 10.759 | 810.660 |
| $x_1 + x_2$ | 3 | 0.2321 | 0.416 | 13.990 | 975.245 |

| Model | $p$ | $s_p^2$ | $R_a^2$ | $PRESS_p$ | $C_p$ |
|---|---|---|---|---|---|
| $x_1 + x_3$ | 3 | 0.1463 | 0.632 | 8.534 | 597.030 |
| $x_1 + x_4$ | 3 | 0.1950 | 0.509 | 11.242 | 811.833 |
| $x_2 + x_3$ | 3 | 0.0773 | 0.806 | 4.428 | 292.616 |
| $x_2 + x_4$ | 3 | 0.1447 | 0.636 | 8.395 | 590.116 |
| $x_2 + x_2^2$ | 3 | 0.2497 | 0.372 | 14.670 | 1053.011 |
| $x_3 + x_4$ | 3 | 0.1295 | 0.674 | 7.575 | 522.872 |
| $x_3 + x_3^2$ | 3 | 0.2297 | 0.422 | 14.002 | 964.789 |
| $x_1 + x_2 + x_3$ | 4 | 0.0117 | 0.971 | 0.745 | 4.368 |
| $x_1 + x_2 + x_4$ | 4 | 0.1474 | 0.629 | 8.755 | 591.367 |
| $x_1 + x_2 + x_2^2$ | 4 | 0.2140 | 0.461 | 13.261 | 879.152 |
| $x_1 + x_3 + x_4$ | 4 | 0.1183 | 0.702 | 7.045 | 465.391 |
| $x_1 + x_3 + x_3^2$ | 4 | 0.1490 | 0.625 | 9.578 | 598.219 |
| $x_2 + x_3 + x_4$ | 4 | 0.0493 | 0.876 | 2.910 | 167.280 |
| $x_2 + x_3 + x_2^2$ | 4 | 0.0784 | 0.803 | 4.459 | 292.736 |
| $x_2 + x_3 + x_2 \cdot x_3$ | 4 | 0.0787 | 0.802 | 4.687 | 294.109 |
| $x_2 + x_3 + x_3^2$ | 4 | 0.0781 | 0.803 | 4.625 | 291.722 |
| $x_2 + x_4 + x_2^2$ | 4 | 0.1433 | 0.639 | 8.511 | 573.333 |
| $x_3 + x_4 + x_3^2$ | 4 | 0.1320 | 0.668 | 8.341 | 524.668 |
| $x_1 + x_2 + x_3 + x_4$ | 5 | 0.0119 | 0.970 | 0.772 | 6.328 |
| $x_1 + x_2 + x_3 + x_2^2$ | 5 | 0.0109 | 0.972 | 0.710 | 2.304 |
| $x_1 + x_2 + x_3 + x_2 \cdot x_3$ | 5 | 0.0115 | 0.971 | 0.807 | 4.549 |
| $x_1 + x_2 + x_3 + x_3^2$ | 5 | 0.0116 | 0.971 | 0.761 | 5.280 |
| $x_1 + x_2 + x_4 + x_2^2$ | 5 | 0.1454 | 0.634 | 8.837 | 572.040 |
| $x_1 + x_3 + x_4 + x_3^2$ | 5 | 0.1207 | 0.696 | 7.863 | 467.345 |
| $x_2 + x_3 + x_4 + x_2^2$ | 5 | 0.0503 | 0.873 | 2.987 | 169.187 |
| $x_2 + x_3 + x_4 + x_2 \cdot x_3$ | 5 | 0.0502 | 0.874 | 3.088 | 168.521 |
| $x_2 + x_3 + x_4 + x_3^2$ | 5 | 0.0503 | 0.873 | 3.055 | 169.066 |
| $x_2 + x_3 + x_2^2 + x_2 \cdot x_3$ | 5 | 0.0795 | 0.800 | 4.686 | 292.702 |
| $x_2 + x_3 + x_2^2 + x_3^2$ | 5 | 0.0796 | 0.800 | 4.842 | 293.394 |
| $x_2 + x_3 + x_2 \cdot x_3 + x_3^2$ | 5 | 0.0795 | 0.800 | 5.013 | 292.734 |
| $x_1 + x_2 + x_3 + x_4 + x_2^2$ | 6 | 0.0112 | 0.972 | 0.738 | 4.290 |
| $x_1 + x_2 + x_3 + x_4 + x_2 \cdot x_3$ | 6 | 0.0117 | 0.971 | 0.845 | 6.453 |
| $x_1 + x_2 + x_3 + x_4 + x_3^2$ | 6 | 0.0118 | 0.970 | 0.787 | 7.273 |
| $x_1 + x_2 + x_3 + x_2^2 + x_2 \cdot x_3$ | 6 | 0.0111 | 0.972 | 0.757 | 4.002 |
| $x_1 + x_2 + x_3 + x_2^2 + x_3^2$ | 6 | 0.0112 | 0.972 | 0.738 | 4.304 |
| $x_1 + x_2 + x_3 + x_2 \cdot x_3 + x_3^2$ | 6 | 0.0115 | 0.971 | 0.855 | 5.855 |
| $x_2 + x_3 + x_4 + x_2^2 + x_2 \cdot x_3$ | 6 | 0.0512 | 0.871 | 3.306 | 170.513 |
| $x_2 + x_3 + x_4 + x_2^2 + x_3^2$ | 6 | 0.0513 | 0.871 | 3.169 | 171.060 |
| $x_2 + x_3 + x_4 + x_2 \cdot x_3 + x_3^2$ | 6 | 0.0512 | 0.871 | 3.417 | 170.422 |
| $x_2 + x_3 + x_2^2 + x_2 \cdot x_3 + x_3^2$ | 6 | 0.0809 | 0.797 | 5.144 | 293.599 |
| $x_1 + x_2 + x_3 + x_4 + x_2^2 + x_2 \cdot x_3$ | 7 | 0.0113 | 0.972 | 0.795 | 6.001 |
| $x_1 + x_2 + x_3 + x_4 + x_2^2 + x_3^2$ | 7 | 0.0114 | 0.971 | 0.767 | 6.290 |
| $x_1 + x_2 + x_3 + x_4 + x_2 \cdot x_3 + x_3^2$ | 7 | 0.0118 | 0.970 | 0.896 | 7.813 |
| $x_1 + x_2 + x_3 + x_2^2 + x_2 \cdot x_3 + x_3^2$ | 7 | 0.0113 | 0.972 | 0.811 | 6.001 |
| $x_2 + x_3 + x_4 + x_2^2 + x_2 \cdot x_3 + x_3^2$ | 7 | 0.0523 | 0.869 | 3.618 | 172.343 |
| Full | 8 | 0.0116 | 0.971 | 0.852 | 8.000 |

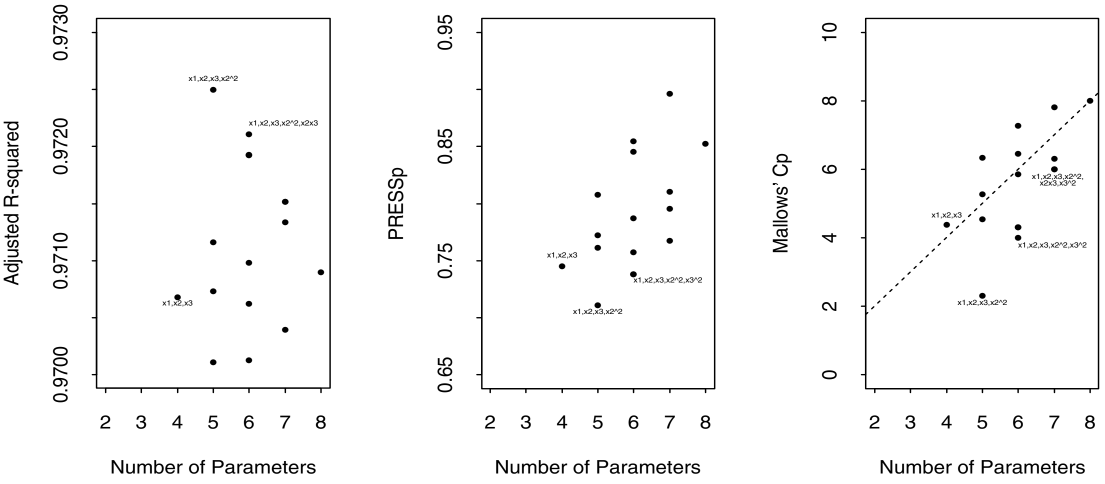
Notice that we have not included the term $x_2^2$ in any model which does not also have the term $x_2$. Similarly, we do not consider models in which $x_3^2$ appears without $x_3$ or in which $x_2 \cdot x_3$ appears without both $x_2$ and $x_3$ also included in the model. This follows a basic model selection principle that to maintain interpretability of the chosen model, we should not include *higher-order* or *interaction* terms in the model if the associated *simple* or *lower-order* terms are not also included in the model. Of course, this principle can be over-ridden if there are compelling external scientific reasons. Among the models described in the above table, we can see several which appear to be performing well. In particular, the models:

- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2^2 + \epsilon$
- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2^2 + \beta_5 (x_2 \cdot x_3) + \epsilon$

appear to be the best models with 4, 5 and 6 parameters, respectively. However, it is difficult to absorb all the numbers in the table, and thus we usually resort to graphical depictions of this information:



Of course, since some of the models are obviously doing very poorly, the plots of all the values are nearly impossible to use, due to the poor scale of the vertical axes. If we restrict the vertical axis, we see a much clearer picture:



It seems that the models noted previously are indeed the best candidate models, as all of the selection criteria recommend them.

iv. Sequential Variable Selection Procedures

The preceding section provided us with some extremely useful model selection criteria. However, in order to use them, we were forced to do a large amount of computational work, since each of the criteria had to be calculated for a large collection of different candidate models. In this section, we will briefly discuss some automated methods to arrive at one (or a few) "best" model(s). The idea behind each of the procedures is a sequential search through all the possible models by either adding or removing a single predictor from the current candidate model at each step in the sequence until a suitable model is found. We must point out, however, that while these methods are enticing in their simplicity and speed, they are not a true substitute for a complete examination and can often arrive at misleading conclusions (and can even commonly suggest models which violate our tenet of not including higher-order and interaction terms without their associated simple and lower-order terms being included in the model), especially in the presence of strong multicollinearity in the predictors.

The first procedure that we shall describe is known as *forward selection*, and it builds a model sequentially by including one predictor at a time to the model until a suitable model is reached. The algorithm proceeds as follows:

1. Start with a model with no predictors and include the single predictor with the highest simple regression $F$-statistic. In other words, we fit a simple regression of the response versus each of the predictors individually, and select for inclusion the one with the "strongest" association with the response. We will denote this predictor as $x_{[1]}$.
2. Next search through the remaining predictors and include the one with the largest *partial* $F$-statistic with respect to the model selected in the initial step. In other words, we include the predictor $x_i$ which has the largest value of

$$F = \frac{SSR(\beta_i|\beta_{[1]}, \beta_0)}{SSE(\beta_i, \beta_{[1]}|\beta_0)},$$

   where $\beta_{[1]}$ is the parameter associated with the predictor $x_{[1]}$, included in the initial step, and

$$SSE(\beta_i, \beta_{[1]}|\beta_0) = SST - SSR(\beta_i, \beta_{[1]}|\beta_0),$$

   is the residual sum of squares from a regression of the response on the two predictors $x_{[1]}$ and $x_i$. The predictor so included will be denoted as $x_{[2]}$.
3. Repeat the previous step, using partial $F$-statistics with respect to all the predictors currently included in the model. In other words, to find the $j^{\text{th}}$ predictor to include, we search through the remaining predictors until we find the one with the largest value of

$$F = \frac{SSR(\beta_j|\beta_{[j-1]}, \ldots, \beta_{[1]}, \beta_0)}{SSE(\beta_j, \beta_{[j-1]}, \ldots, \beta_{[1]}|\beta_0)}.$$

4. Halt the inclusion of predictors the first time that the "largest" partial $F$-ratio is less than a user-specified cut-off value, denoted $F_{IN}$ or *$F$-to-enter*.

The next procedure is known as *backward elimination*, and builds a model sequentially by removing predictors one at a time from the full model until a suitable model is reached. The algorithm proceeds as follows:

1. Start with the full model and find the predictor with the smallest partial $F$-statistic,

$$F = \frac{SSR(\beta_i|\beta_k, \ldots, \beta_{i+1}, \beta_{i-1}, \ldots, \beta_1, beta_0)}{MSE_{full}}.$$

   Remove this predictor, denoted as $x_{[1]}$, from the model.
2. Repeat the preceding step using the model without $x_{[1]}$ as the new "full" model.
3. Halt the removal of predictors the first time that none of the partial $F$-ratios are smaller than a user-specified cut-off value, denoted $F_{OUT}$ or *$F$-to-exit*.

The final procedure is referred to as *forward stepwise regression*, and constitutes a combination of forward selection and backward elimination, choosing a model sequentially by alternating between inclusion and removal of single predictors at each stage of the model building process. Clearly, this is a sensible modification of the "uni-directional" approach of the other two procedures, since multicollinearity can certainly render a predictor of little overall significance even though it may have appeared significant at an earlier stage in the model-building process. The algorithm for forward stepwise regression proceeds as follows:

1. As for forward selection, start with a model with no predictors and include the single predictor with the highest simple regression $F$-statistic. Again, we refer to this predictor as $x_{[1]}$
2. Again as for forward selection, include the predictor with the largest partial $F$-ratio into the model and denote it as $x_{[2]}$. Subsequently, examine the partial $F$-statistic of $x_{[1]}$,
$$F = \frac{SSR(\beta_{[1]}|\beta_{[2]}, \beta_0)}{SSE(\beta_{[1]}, \beta_{[2]}|\beta_0)},$$
to see if it is still significant. If this partial $F$-statistic is smaller that a user-specified value $F_{OUT}$, then remove $x_{[1]}$ from the model, otherwise retain it.
3. Search through the remaining variables and find the predictor which again has the highest partial $F$-statistic with respect to the predictors currently included in the model, denote this predictor as $x_{[3]}$. Include this predictor into the model. Then, search through the predictors in the model to see if their new partial $F$-ratios (i.e., with repect to the model which now has $x_{[3]}$ included) are still significant. If any of the partial $F$-statistics are less than the cut-off $F_{OUT}$, remove the predictor with the smallest partial $F$-value, otherwise retain all the predictors.
4. Repeat this process (i.e., including the "best" remaining predictor based on partial $F$-tests, and then removing the "worst" currently included predictor if any of the partial $F$-statistics are below the $F_{OUT}$ threshhold) until the first time that there is no remaining predictor with a partial $F$-ratio greater than a user-specified cut-off, $F_{IN}$. (NOTE: This algorithm requires that $F_{IN} > F_{OUT}$, since otherwise it is possible for a predictor to be removed and then re-included in an infinite loop, if its associated $F$-ratio were to lie in the range $F_{IN} < F < F_{OUT}$.)

Note that a forward stepwise regression in which $F_{OUT} = 0$ is equivalent to the forward selection procedure. Similarly, there is a procedure closely related to forward stepwise regression called *backward* stepwise regression, in which the order of the inclusion and exclusion steps is reversed and the algorithm starts from the full model rather than the empty model. For backward stepwise regression, a value of $F_{IN} = \infty$ produces the backward elimination procedure. This leads us to the obvious problem of good choices for the $F_{IN}$ and $F_{OUT}$ values. There are no generally accepted values. However, it is commonly suggest that $F_{OUT}$ values should correspond to the traditional testing critical values, that is, to the 90%, 95% or 99% quantile of an appropriate $F$-distribution (which will generally be the $F_{P,n-P}$-distribution, where $P$ is the number of parameters in the full model, giving values on the order of 2 to 4). On the other hand, it is generally agreed that the $F_{IN}$ value should be somewhat less stringent, typically on the order of the 50% or 75% quantile of the $F_{P,n-P}$-distribution (giving values on the order of 1.5 to 3.5). As a basic rationale for this recommendation, recall that in underspecified models, the $MSE$ tends to have a large positive bias. In the early stages of the two forward procedures (i.e., forward selection and forward stepwise regression), the models will tend to be underspecified, so that the $MSE$ will tend to be an inflated estimate of the true error scale. This inflation of the $MSE$ will then tend to artificially deflate the partial $F$-statistics, which may cause the forward sequential procedures to "stop short" and not allow important predictors into the model. Of course, for the forward stepwise regression procedure, both of these suggestions cannot be used, since we require $F_{IN} > F_{OUT}$, and so we will typically use the values $F_{IN} \approx 3.5$ and $F_{OUT} \approx 3.0$ as a reasonable compromise. We should reiterate, moreover, that these procedures should be seen as quick, exploratory attempts at model building, and the models at which the finally arrive are not necessarily the models which we should favor. Indeed, it is generally common practice to follow the *trace* of the algorithms (i.e., the listing of the ordered models through which the sequential procedure went in arriving at its final model) and perhaps taking the last few in the sequence as candidate models deserving further investigation, and thus the exact values of the cut-offs for inclusion and removal are not overly critical.

*Example 7 (cont'd) - Surgical Data:* We now use each of the three methods outlined above to pick a model for the logarithm of patients' survival times from the 7 predictors (including the three higher-order interaction terms) in the full model:

```
> dimnames(srg.all)[[2]][5:7] <- c("prog2","prgenz","enzyme2")
> forsel <- stepwise(srg.all[,1:7],srg.all[,8],method="forward")
> forsel$which
      clot prog enzyme liver prog2 prgenz enzyme2
1(+6)    F    F      F     F     F      T       F
2(+1)    T    F      F     F     F      T       F
3(+7)    T    F      F     F     F      T       T
4(+2)    T    T      F     F     F      T       T
5(+3)    T    T      T     F     F      T       T
6(+5)    T    T      T     F     T      T       T
7(+4)    T    T      T     T     T      T       T
> forsel$f.stat
[1] 1.745790e+02 6.104035e+01 3.032853e+01 3.317351e+01 2.282057e+01
[6] 1.894791e+00 8.864628e-04
```

Notice that in *S-Plus*, if we perform forward selection, we do not need to specify $F_{IN}$ at the outset. The `stepwise()` command will simply produce a complete sequence from the empty model to the full model, showing the order in which the predictors are added and giving the associated $F$-statistics, so that we can see what model would have been reached with any specified $F_{IN}$ value. For instance, for $F_{IN} = 4$, we would stop after the 5th step at the model:

$$\log(\texttt{survival}) = \beta_0 + \beta_1\texttt{clot} + \beta_2\texttt{prog} + \beta_3\texttt{enzyme} + \beta_6\texttt{prgenz} + \beta_7\texttt{enzyme2}.$$

In fact, the sharp drop in the $F$-statistic at this step (from 22.8 down to 1.9) indicates that we would arrive at this model for most reasonable $F_{IN}$ values. Also, note that this is *not* one of the models which our more thorough investigation in the preceding section suggested.

```
> bckel <- stepwise(srg.all[,1:7],srg.all[,8],method="backward")
      clot prog enzyme liver prog2 prgenz enzyme2
6(-4)    T    T      T     F     T      T       T
5(-7)    T    T      T     F     T      T       F
4(-6)    T    T      T     F     T      F       F
3(-5)    T    T      T     F     F      F       F
2(-1)    F    T      T     F     F      F       F
```

```
1(-2)    F   F      T    F    F     F      F
0(-3)    F   F      F    F    F     F      F
> bckel$f.stat
[1] 8.864628e-04 1.110610e-03 3.023078e-01 4.037184e+00 2.650762e+02
[6] 9.113832e+01 3.648910e+01
```

Again, we do not need to specify an $F_{OUT}$ value initially to perform a backward elimination. The *S-Plus* function `stepwise()` will produce the complete sequence from the full model to the empty model, showing the order in which the predictors are removed and providing the associated $F$-statistics. Thus, we can again simply use the list of $F$-ratios to see where the algorithm would have stopped for any specified value of $F_{OUT}$. In this example, the $F$-statistic changes from 0.3 to 4 between the third and fourth step, indicating that for most reasonable $F_{OUT}$ values, the algorithm would stop after the third removal, yielding the model:

$$\log(\texttt{survival}) = \beta_0 + \beta_1\texttt{clot} + \beta_2\texttt{prog} + \beta_3\texttt{enzyme} + \beta_5\texttt{prog2}.$$

Note that this is not the same model produced by the forward selection procedure, but it is one of the models which our earlier more extensive investigation in the previous section suggested.

```
> forstp <- stepwise(srg.all[,1:7],srg.all[,8],f.crit=c(3.5,3))
> forstp$which
      clot prog enzyme liver prog2 prgenz enzyme2
1(+6)    F   F      F    F    F     T      F
2(+1)    T   F      F    F    F     T      F
3(+7)    T   F      F    F    F     T      T
4(+2)    T   T      F    F    F     T      T
3(-6)    T   T      F    F    F     F      T
4(+3)    T   T      T    F    F     F      T
3(-7)    T   T      T    F    F     F      F
4(+5)    T   T      T    F    T     F      F
```

Finally, for the forward stepwise procedure (which is the default method for the *S-Plus* `stepwise()` function) we must specify the $F_{IN}$ and $F_{OUT}$ values initially. Setting $F_{IN} = 3.5$ and $F_{OUT} = 3$, the model at which the procedure arrives is:

$$\log(\texttt{survival}) = \beta_0 + \beta_1\texttt{clot} + \beta_2\texttt{prog} + \beta_3\texttt{enzyme} + \beta_5\texttt{prog2}.$$

It is interesting to note that this model does not contain the interaction `prgenz`, despite the fact that it was the first predictor included into the model. Note also that this is the same model as the backward elimination prodecure produced. This coincidence of the results of the two procedures, along with its prominent place in the more extensive earlier investigation adds credence to the choice of this model as the most appropriate. Nonetheless, this should not be seen as the termination of the modelling process, since we should examine this model (and the other candidate models which performed nearly as well on the different selection criteria) using our residual analysis and influence diagnostic procedures as well.

v. Conclusion

As a closing note, we point out that the model selection techniques we have discussed here are generally most useful when we are interested in either variable screening or future prediction, since in these instances, we simply want to construct a good description of the data. If, however, we are interested in the significance of a specific regressor variable, the situation becomes more nebulous. We could, of course, test whether this variable has a significant $F$-statistic in the model or models that we have arrived at through our model selection procedures (provided of course that the predictor is included in the chosen model). However, this approach has two basic problems. First, we typically select our models based on various $F$-statistics and thus, the $F$-ratios of included terms are by their nature large. Second, if it does happen that we find that an included variable has an insignificant $F$-ratio, we are in somewhat of a quandry, since this indicates that we think this predictor should be removed from the model, even though we are ostensibly dealing with our chosen "best" models. One suggestion then is to ignore $F$-tests and simply say that a predictor is important if it appears in all (or at least most) of the models at which our selection procedures arrive. Alternatively, we could simply ignore model selection altogether and focus on our earlier ideas of sequential and partial $F$-tests. This approach, however, ignores all of the issues which led us to develop model selection procedures in the first place. In the final analysis, what is really at the center of our dilemma is that regression techniques are truly only designed to investigate statistical associations (which can often be interrelated in very confusing and confounding ways) and not true causal relationships. Finally, we might simply force our model selection procedures to include the predictor in which we are interested. Then, from among the selected "best" models, we can examine the $F$-statistic of our predictor of interest (which has not been used in any of our selection procedures) and decide whether it is significant or not (see tutorial exercises).

# SCHOOL OF FINANCE AND APPLIED STATISTICS

## REGRESSION MODELLING<br>(STAT2008/STAT8038)

### ISSUES and EXAMPLE DATASETS

Regression modelling addresses issues such as:

- How do we arrive at a suitable model for our data?
  - · What functional form should we use for the model?
  - · Which predictor variables should we include in the model?
- How do we assess the validity of the underlying assumptions of our chosen model?
- How do we reach conclusions based on our chosen model?
  - · How do we construct confidence intervals and perform hypothesis tests?
- How can we predict future values of the response variable using our chosen model?

Throughout the course, we shall focus on these issues in many different circumstances and we will make reference to various sets of example data, including the following:

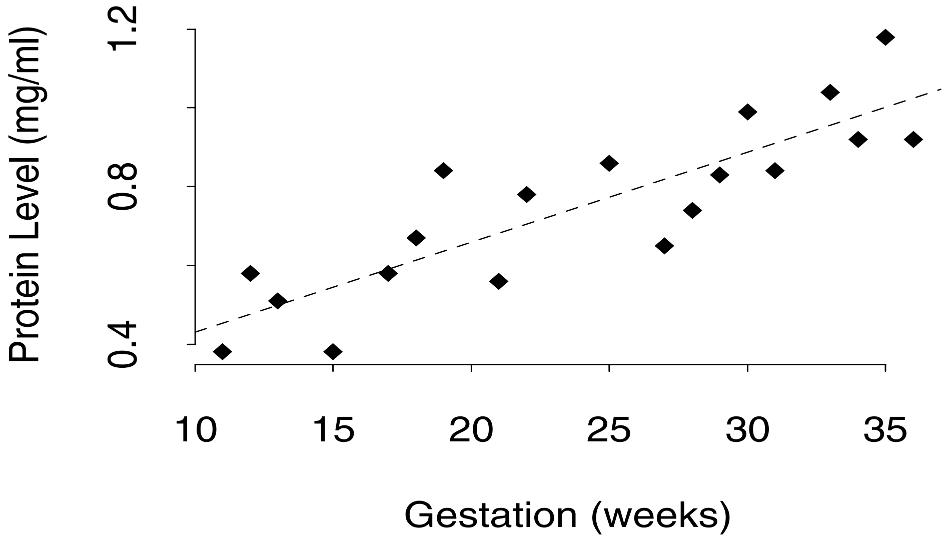
*Example 1: The Protein in Pregnancy data*

Observations were taken on 19 healthy pregnant women. Each woman was at a different stage of her pregnancy (gestation) and her protein level (in milligrams per milliliter) was measured. Below are a listing and a scatterplot of the data (for details see Sutcliffe, *et al* (1982) *Placenta*, **3**, pp71-80.)

| Protein level | Gestation |
|---|---|
| (*mg/ml*) | (weeks) |
| 0.38 | 11 |
| 0.58 | 12 |
| 0.51 | 13 |
| 0.38 | 15 |
| 0.58 | 17 |
| 0.67 | 18 |

| Protein level | Gestation |
|---|---|
| (*mg/ml*) | (weeks) |
| 0.84 | 19 |
| 0.56 | 21 |
| 0.78 | 22 |
| 0.86 | 25 |
| 0.65 | 27 |
| 0.74 | 28 |
| 0.83 | 29 |

| Protein level | Gestation |
|---|---|
| (*mg/ml*) | (weeks) |
| 0.99 | 30 |
| 0.84 | 31 |
| 1.04 | 33 |
| 0.92 | 34 |
| 1.18 | 35 |
| 0.92 | 36 |

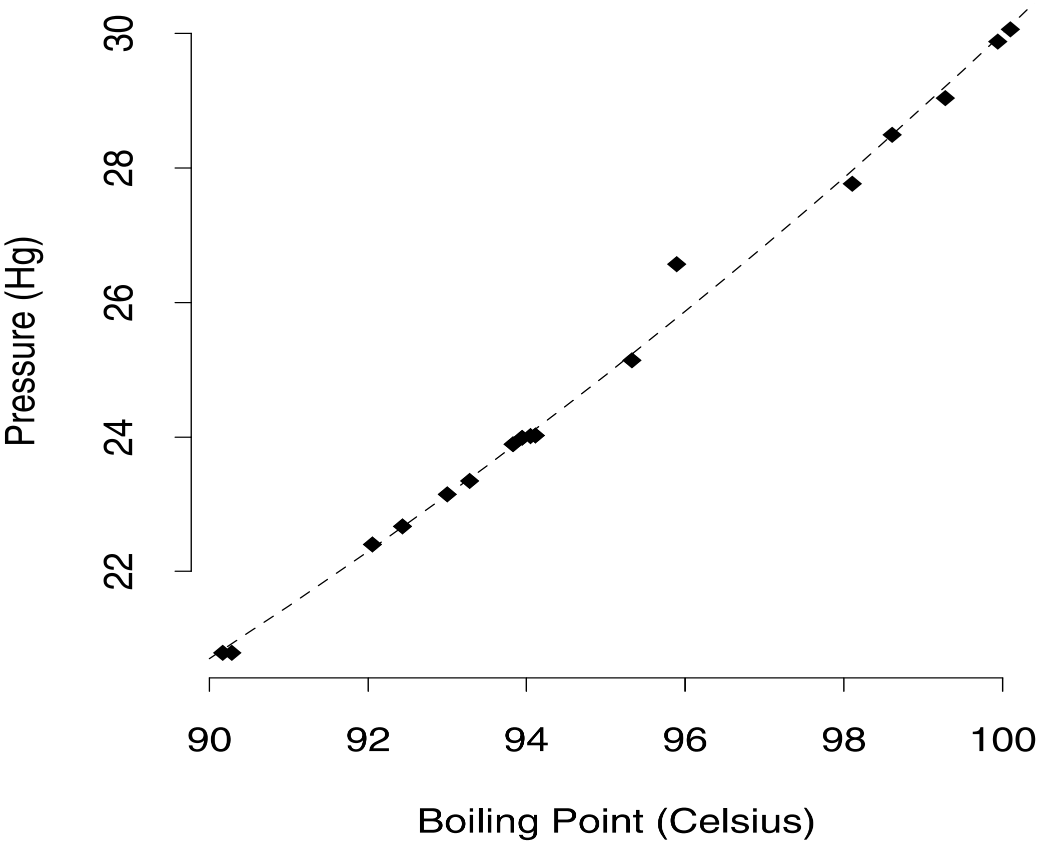


$$Model: \text{ Protein Level} = \beta_0 + \beta_1 \text{Gestation} + \epsilon$$

*Example 2: The Forbes Data*

In the 1840's it was difficult for travellers to estimate altitude. This could be done through measurements of atmospheric pressure, but the barometers of the time were rather fragile. The Scottish physicist James Forbes investigated the use of boiling point (which itself is related to atmospheric pressure) as a means of identifying altitude. The following data relating the boiling point of water (in Celsius degrees, $^\circ C$) to the atmospheric pressure (in inches of mercury, $Hg$) were collected by him in the Alps and in Scotland (for details see Forbes (1857) *Trans. R. Soc. Edinburgh*, **21**, pp135-143.)

| Boiling point | Pressure |
|---|---|
| ($^\circ C$) | ($Hg$) |
| 90.28 | 20.79 |
| 90.17 | 20.79 |
| 92.06 | 22.40 |
| 92.44 | 22.67 |
| 93.00 | 23.15 |
| 93.28 | 23.35 |
| 93.83 | 23.89 |
| 93.94 | 23.99 |
| 94.11 | 24.02 |
| 94.05 | 24.01 |
| 95.33 | 25.14 |
| 95.89 | 26.57 |
| 98.61 | 28.49 |
| 98.11 | 27.76 |
| 99.28 | 29.04 |
| 99.94 | 29.88 |
| 100.1 | 30.06 |



Forbes suggested a model of the form:

$$Model: \ \ln(\text{Pressure}) = \beta_0 + \beta_1 \text{Boiling Point} + \epsilon$$

or

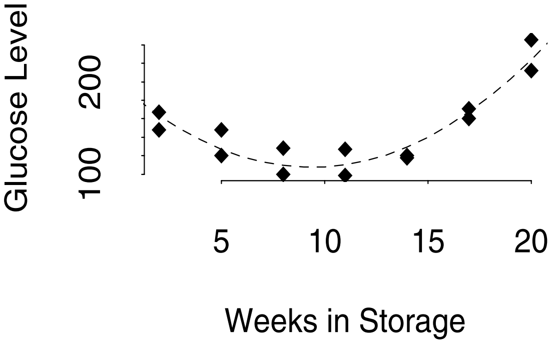
$$Model: \ \text{Pressure} = e^{\beta_0 + \beta_1 \text{Boiling Point}} + \epsilon$$

though a simple linear model may work equally well here.

*Example 3: Sugar in Potatoes*

The following data were collected in an experiment to investigate the glucose content of potatoes during storage:

| Glucose | Weeks |
|---|---|
| 148 | 2 |
| 167 | 2 |
| 120 | 5 |
| 148 | 5 |
| 100 | 8 |
| 128 | 8 |
| 99 | 11 |
| 127 | 11 |
| 118 | 14 |
| 120 | 14 |
| 160 | 17 |
| 171 | 17 |
| 212 | 20 |
| 245 | 20 |



The underlying chemical process would indicate an initial fall in glucose levels followed by an increase, so we might use:

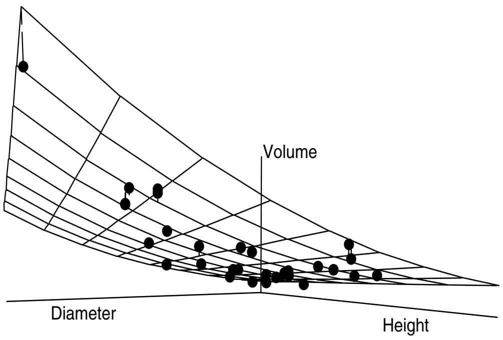
$$Model: \ \text{Glucose} = \beta_0 + \beta_1\text{Weeks} + \beta_2\text{Weeks}^2 + \epsilon$$

*Example 4: Black Cherry Trees*

Data were obtained from a sample of black cherry trees in order to examine the relationship between the volume (cubic feet) of wood in the tree and its height (feet) and diameter (inches). (For details of the data set see Meyer (1953) *Forest Mensuration*, State College, PA., Penns Valley Publishers.)

| Diameter | Height | Volume |
|---|---|---|
| (inches) | (feet) | (feet$^3$) |
| 8.3 | 70 | 10.3 |
| 8.6 | 65 | 10.3 |
| 8.8 | 63 | 10.2 |
| 10.5 | 72 | 16.4 |
| 10.7 | 81 | 18.8 |
| 10.8 | 83 | 19.7 |
| 11.0 | 66 | 15.6 |
| 11.0 | 75 | 18.2 |
| 11.1 | 80 | 22.6 |
| 11.2 | 75 | 19.9 |
| 11.3 | 79 | 24.2 |
| 11.4 | 76 | 21.0 |
| 11.4 | 76 | 21.4 |
| 11.7 | 69 | 21.3 |
| 12.0 | 75 | 19.1 |
| 12.9 | 74 | 22.2 |
| 12.9 | 85 | 33.8 |
| 13.3 | 86 | 27.4 |
| 13.7 | 71 | 25.7 |
| 13.8 | 64 | 24.9 |
| 14.0 | 78 | 34.5 |
| 14.2 | 80 | 31.7 |

| Diameter | Height | Volume |
|---|---|---|
| (inches) | (feet) | (feet$^3$) |
| 14.5 | 74 | 36.3 |
| 16.0 | 72 | 38.3 |
| 16.3 | 77 | 42.6 |
| 17.3 | 81 | 55.4 |
| 17.5 | 82 | 55.7 |
| 17.9 | 80 | 58.3 |
| 18.0 | 80 | 51.5 |
| 18.0 | 80 | 51.0 |
| 20.6 | 87 | 77.0 |



The underlying geometery suggests a model of the form:

$$Model: \ \text{Volume} = \beta_0\text{Diameter}^{\beta_1}\text{Height}^{\beta_2} + \epsilon$$

or

$$Model: \ \ln(\text{Volume}) = \beta_0 + \beta_1\text{Diameter} + \beta_2\text{Height} + \epsilon$$

*Example 5: Giving in the Church of England*

The following data record the amount of annual giving per church member in a sample of 20 dioceses in the Church of England. Three other potentially relevant factors, employment rate, the percentage of the population on the electoral roll of the church and the percentage of the population who usually attend church, are also recorded. (For details see Pickering (1983) An anlysis of giving in the Church of England. *Applied Economics* **17**, 619-32.)

| Annual giving | Employment | Electoral Roll | Attendance |
|---|---|---|---|
| *(£)* | (%) | (%) | (%) |
| 43 | 89.9 | 7.2 | 4.6 |
| 61 | 83.6 | 1.9 | 1.4 |
| 37 | 86.4 | 5.7 | 3.1 |
| 54 | 87.1 | 3.2 | 2.3 |
| 71 | 89.6 | 3.0 | 2.4 |
| 37 | 87.7 | 8.7 | 4.1 |
| 55 | 89.3 | 2.3 | 1.9 |
| 43 | 85.0 | 3.7 | 2.6 |
| 43 | 82.7 | 3.4 | 1.9 |
| 49 | 92.8 | 5.1 | 3.4 |
| 48 | 84.9 | 3.5 | 2.4 |
| 36 | 86.3 | 5.4 | 3.1 |
| 44 | 82.6 | 3.1 | 2.5 |
| 43 | 85.9 | 3.3 | 2.3 |
| 56 | 92.0 | 3.9 | 3.0 |
| 43 | 89.4 | 3.8 | 2.5 |
| 36 | 87.7 | 4.6 | 2.9 |
| 56 | 88.1 | 2.7 | 2.1 |
| 45 | 86.6 | 3.0 | 2.2 |
| 40 | 90.0 | 4.9 | 3.3 |

One possible model for this dataset might be:

$$Model: \ \text{Giving} = \beta_0 + \beta_1\text{Employment} + \beta_2\text{Electoral Roll} + \beta_3\text{Attendance} + \epsilon.$$