

Contents

1 More Linear Regression ................................. 1
  1.1 Introduction ........................................ 1
  1.2 Simple Linear Regression ............................ 1
    1.2.1 Results ...................................... 2
    1.2.2 Prediction of a Single (New) Observation, $Y_{h(new)}$ .... 3
  1.3 Coefficient of Determination ....................... 4
  1.4 One-way Analysis of Variance ....................... 5
  1.5 Lack of Fit Test ..................................... 5
  1.6 Testing Parallelism of Straight Lines ............. 8
    1.6.1 Two Straight Lines ............................ 8
    1.6.2 $k$ Lines ($k > 2$) — An Example of the General Linear Test .... 9
    1.6.3 Residual Sums of Squares ..................... 10
    1.6.4 Degrees of Freedom ........................... 12
    1.6.5 Hypotheses and Tests ......................... 12

2 Matrix Approach to Regression ........................ 18
  2.1 Multiple Linear Regression ........................ 18
    2.1.1 Notation .................................... 19
    2.1.2 Matrix Approach .............................. 20
    2.1.3 Results Using Matrices ...................... 21
    2.1.4 Partitioning the Total SS .................... 24
    2.1.5 Cochran’s Theorem ........................... 24
  2.2 Distribution of the Regression and Residual SS ... 25
    2.2.1 Distribution assuming $\beta=0$ ................ 25
    2.2.2 Distribution assuming $\beta \neq 0$ ............ 25
  2.3 Testing $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ ....... 28
  2.4 General Linear Test-Testing Subsets of Variables .... 30
    2.4.1 Testing $\beta_i = 0$ .......................... 32
    2.4.2 Coefficient of Multiple Determination ......... 33
  2.5 Weighted Least Squares ............................ 34
3 Model Building

3.1 Introduction .......................................................... 36
3.2 Stepwise Regression ............................................... 36
  3.2.1 Search Procedure ............................................. 37
  3.2.2 Forward Selection ........................................... 42
3.3 Linear Models and Experimental Design ......................... 46
3.4 One-Way Analysis of Variance ................................... 47
  3.4.1 Using Cochran’s Theorem .................................... 50
  3.4.2 Testing the Hypothesis ...................................... 50
3.5 Two-Way Analysis of Variance .................................... 50
  3.5.1 Testing the Hypotheses ....................................... 53
  3.5.2 Hypotheses and Tests ......................................... 54

4 Model Assessment ...................................................... 55

4.1 Introduction .......................................................... 55
4.2 Residual Analysis ................................................... 55
  4.2.1 Standardized Residuals ....................................... 56
  4.2.2 Residuals versus an Independent Variable .................. 56
  4.2.3 Residuals versus Fitted Values ............................... 58
  4.2.4 Testing for Normality: Q-Q Plots ............................. 59
  4.2.5 Time Sequence Plots ........................................... 59
  4.2.6 Durbin-Watson Test for Autocorrelation .................... 59
4.3 Transformations ....................................................... 61
  4.3.1 Linearity Transformations ..................................... 61
  4.3.2 Non-Constant Variance ....................................... 62
4.4 Multicollinearity ..................................................... 62
  4.4.1 Problems Associated with Multicollinearity ................ 63
  4.4.2 Polynomial Regression ........................................ 67
  4.4.3 A Note on Extrapolation ...................................... 68
4.5 Detecting Multicollinearity ........................................ 68
  4.5.1 Variance Inflation Factor .................................... 69
  4.5.2 Outliers ......................................................... 72
  4.5.3 Deleted Residuals ................................................ 73
  4.5.4 Influential Observations ...................................... 76
  4.5.5 Cook’s Distance ................................................ 76
4.6 Comment .............................................................. 77

# List of Figures

1.1 Multiple Y values for each X ........................................... 7  
1.2 Parallelism ................................................................. 13  
1.3 Scatter Plot of Cholesterol against Age ................................. 14  

4.1 Typical Patterns in Residual Plots .................................... 57  
4.2 Residuals vs Fitted Values .............................................. 58  
4.3 Quantile-quantile (Q-Q) plot. ......................................... 59  
4.4 Time Sequence Plot of Residuals- Negative Autocorrelation ...... 60  
4.5 Time Sequence Plot of Residuals - Positive Autocorrelation ...... 61  
4.6 Region of Observations .................................................. 68  
4.7 Graphs for Model 1, Naval Hospital Data .............................. 74  
4.8 Graphs for Model 2, Naval Hospital Data .............................. 75
Chapter 1

More Linear Regression

1.1 Introduction

In Statistics 261 the question of how to fit a line of best fit to a set of data was considered. The method of least squares was used to fit a straight line of the form \( \hat{Y} = \beta_0 + \beta_1 X \).

The following section is a summary of Simple Linear Regression from Statistics 261 and should not be new to you.

1.2 Simple Linear Regression

Given \( n \) data points

\[
(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)
\]

consider the model:

\[
Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ where } i = 1, 2, \ldots, n. \tag{1.1}
\]

\( Y \) is the response(dependent) variable,

\( X \) is the independent variable and is assumed known,

\( \beta_0 \) and \( \beta_1 \) are (unknown) parameters to be estimated,

\( \epsilon_i \) is a random error term.

The error terms, \( \epsilon_i \), are assumed to be independently normally distributed with mean zero and variance \( \sigma^2 \).
1.2.1 Results

The following results should be known. Make sure you can derive them and that you understand their meaning.

1. $\hat{\beta}_1 = \frac{\sum xy}{\sum x^2}$, where $\sum xy = \sum_{i=1}^{n}(X_i - \bar{X})(y_i - \bar{y})$, $\sum x^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$.

2. $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\overline{X}$.

3. $\text{Cov}(\bar{y}, \hat{\beta}_1) = 0$.

4. $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$. (That is, the sum of the residuals equals zero.)

5. Regression SS = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum x^2 = \frac{(\sum xy)^2}{\sum x^2}$

(Note: NWK denotes the regression SS by SSR and the residual, or error SS by SSE)

6. Residual SS = Total SS - Regression SS = $\sum y^2 - \frac{(\sum xy)^2}{\sum x^2}$.

7. $\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum x^2}$.

8. $\text{var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\overline{X}^2}{\sum x^2} \right)$.

9. Since $\sigma^2$ is not usually known it is replaced by an (unbiased) estimate,

$$\hat{\sigma}^2 = \frac{\text{Residual SS}}{n - 2}.$$ 

10. To test the hypothesis that the slope of the fitted line is zero ($\beta_1 = 0$) it is useful to use an Analysis of Variance Table of the form:

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression (due to fitting $\hat{\beta}_1$)</td>
<td>1</td>
<td>$\frac{(\sum xy)^2}{\sum x^2}$</td>
<td>$\frac{(\sum xy)^2}{\sum x^2}$</td>
</tr>
<tr>
<td>Residual (dev’ns about fitted line)</td>
<td>$(n - 2)$</td>
<td>$\sum y^2 - \frac{(\sum xy)^2}{\sum x^2}$</td>
<td>$\frac{\sum y^2 - \frac{(\sum xy)^2}{\sum x^2}}{n - 2}$</td>
</tr>
<tr>
<td>Total</td>
<td>$(n - 1)$</td>
<td>$\sum y^2$</td>
<td></td>
</tr>
</tbody>
</table>
This leads to the usual F test of $\beta_1 = 0$, with 1 and $(n - 2)$ degrees of freedom. Notice that the residual or error mean square (MSE) in the AOV table is the estimate of $\sigma^2$ given in result 9 above.

11. $(1 - \alpha)\%$ Confidence interval for $\beta_0$

$$\hat{\beta}_0 \pm t_{(1-\frac{\alpha}{2})}\, \text{se}(\hat{\beta}_0).$$

12. $(1 - \alpha)\%$ Confidence interval for $\beta_1$

$$\hat{\beta}_1 \pm t_{(1-\frac{\alpha}{2})}\, \text{se}(\hat{\beta}_1).$$

13. The mean response at $X = X_h$ is estimated by $\hat{Y}_h = \hat{\alpha} + \hat{\beta} X_h$ and $E(\hat{Y}_h|X_h) = \alpha + \beta X_h$.

14. The variance of the mean response, $Y_h$, at $X_h$ is

$$\text{Var}(\hat{Y}_h) = \sigma^2 \left( \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_j (X_j - \bar{X})^2} \right). \tag{1.2}$$

As $\sigma^2$ is usually unknown it is replaced by $\hat{\sigma}^2 = \text{MSE}$.

15. The interval estimate of $E(\hat{Y}_h|X_h)$ is given by

$$\hat{Y}_h \pm t_{(1-\frac{\alpha}{2})}\, \text{se}(\hat{Y}_h).$$

1.2.2 Prediction of a Single (New) Observation, $Y_{h\text{(new)}}$.

The value of $Y_{h\text{(new)}}$ at a particular $X_h$ is estimated by $\hat{Y}_h$. The prediction interval is estimated by

$$\hat{Y}_h \pm t_{(1-\frac{\alpha}{2})}\, \sqrt{\text{MSE} \left( 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right)}. \tag{1.3}$$

Proof:

If $Y_{h\text{(new)}}$ is a new observation, $\hat{Y}_h$ and $Y_{h\text{(new)}}$ are independent since $Y_{h\text{(new)}}$ is not used to calculate $\hat{Y}_h$. But since $\hat{Y}$ and $Y_{h\text{(new)}}$ are independent

$$\text{var}(\hat{Y}_h - Y_{h\text{(new)}}) = \text{Var}(\hat{Y}_h) + \text{Var}(Y)$$

$$= \text{Var}(\hat{Y}) + \sigma^2$$

since $\text{Var}(Y_{h\text{(new)}}) = \text{Var}(Y) = \sigma^2$. 

3
Noting that $\sigma^2$ is usually unknown and relacing it by the MSE in result 14 the standard error $\hat{Y}_h - Y_{h(new)}$ is estimated by

$$se(\hat{Y}_h - Y_{h(new)}) = \sqrt{\text{MSE} \left( 1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_j(X_j - \bar{X})^2} \right)}.$$  

Then,

$$\frac{\hat{Y}_h - Y_{h(new)}}{se(\hat{Y}_h - Y_{h(new)})}$$

is distributed as a $t$ on $(n-2)$ degrees of freedom.

Thus a $(1-\alpha)$ prediction interval can be based on the above statistic by writing

$$P \left( \frac{|\hat{Y}_h - Y_{h(new)}|}{se(\hat{Y}_h - Y_{h(new)})} \leq t \right) = 1 - \alpha$$

which can be rewritten as

$$\hat{Y}_l = \hat{Y}_h - t_{(1-\alpha)} \cdot se[\hat{Y}_h - Y_{h(new)}]$$

$$\hat{Y}_u = \hat{Y}_h + t_{(1-\alpha)} \cdot se[\hat{Y}_h - Y_{h(new)}]$$

$$P[\hat{Y}_l \leq Y_{h(new)} \leq \hat{Y}_u] = (1 - \alpha)$$

and the result follows. Note this is not a confidence interval as $\hat{Y}_{h(new)}$ is a random variable, not a parameter of the distribution.

Take particular note of the difference between result 15, the confidence interval for the mean response (a function of the unknown parameters) and the prediction interval for one observation (a random variable).

### 1.3 Coefficient of Determination

The total sum of squares, $\sum y^2$, measures the variation in $Y$ but does not take into account the independent variable $X$. On the other hand the residual (or error) sum of squares (SSE) measures the variation in the $Y_i$ when the contribution of the independent variable, $X$, in reducing the overall variation has been taken into account through the regression model.

A measure of the effect of $X$ in reducing the variation in $Y$ is given by

$$r^2 = \frac{\text{Regression SS}}{\sum y^2} = 1 - \frac{\text{SSE}}{\sum y^2} \quad (1.4)$$

which is called the coefficient of determination.

$$0 \leq r^2 \leq 1.$$
The measure $r^2$ represents the proportion of the total variation that can be explained by the use of the independent variable, $X$.

The square root of $r^2$

$$r = \pm \sqrt{r^2}$$

is the \textit{correlation coefficient} introduced in chapter 8 of Stat 261. The sign of $r$ is the same as the sign associated with the slope of the regression equation.

\subsection{1.4 One-way Analysis of Variance}

Single Factor Analysis of Variance was introduced in Stat 261 where the Analysis of Variance Table was of the form

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
<th>$F$ Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>$(k - 1)$</td>
<td>$\sum_i n_i (X_i - \bar{X}_.)^2$</td>
<td>$\frac{SS_B}{(k - 1)}$</td>
<td>$MS_B$</td>
</tr>
<tr>
<td>Within Groups</td>
<td>$(n - k)$</td>
<td>$\sum_{i,j} (X_{ij} - \bar{X}_i)^2$</td>
<td>$\frac{SS_W}{(n - k)}$</td>
<td>$MS_W$</td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>$\sum_{i,j} (X_{ij} - \bar{X}_.)^2$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This should be revised.

\subsection{1.5 Lack of Fit Test}

If there are (independent) repeat observations at one or more $X$ levels then it is possible to test if the regression line appears to fit the data. One of the assumptions made in regression analysis is that the residuals have a common variance, $\sigma^2$. When there are replicate $Y$ values for one or more of the $X$’s two independent estimates of $\sigma^2$ can be obtained. Together these two components make up the SSE for the regression analysis. Suppose there are $c$ $X$ levels with $n_i$ repeat observations at each level. Then:

(i) \textbf{Pure Error}: Denote the $i$th observation at the $j$th level of $X$ by $y_{ij}$. Consider only the $n_j$ $y$ observations at this level of $X$ and calculate the SS of deviations from their mean, $\bar{y}_j$, that is

$$\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2.$$  

These are then added over all levels of $X$ to give the SS for Pure Error.

$$\text{Pure Error SS} = SS_{pe} = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2.$$  \hspace{1cm} (1.5)

Note that this SS is not affected by the actual regression line fitted as it is obtained from within single levels of $X$. The pure error mean square provides an unbiased
estimate of \( \sigma^2 \) (df = \( N - c \), \( N = \sum n_i \)).

(Notice that the Pure Error SS is the **Within Treatments SS** if we regard each level of \( X \) as a treatment and carry out a One-Way analysis of variance.)

**Lack of Fit:** The Lack of Fit SS is given by

\[
\text{Lack of Fit SS} = SS_{\text{lof}} = \sum_{j=1}^{c} n_j (\bar{y}_j - \hat{\gamma})^2.
\]

Note that the estimate of \( \sigma^2 \) obtained from this SS does depend on the fitted line as it represents deviations of the means from the fitted line. The closer the \( \bar{y} \) are to the fitted values \( \hat{\gamma} \) the better the indication that the line appears to fit the data. Large values would tend to indicate the line was not fitting.

The hypothesis that the line appears to provide an adequate fit is tested by

\[
\frac{(\text{Lack Fit SS})/(c - 2)}{(\text{Pure Error SS})/(N - c)}
\]

which is distributed as an F with \((c - 2)\) and \((N - c)\) degrees of freedom.

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
<th>F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>Regression SS</td>
<td>Regression MS</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>((N - 2))</td>
<td>Error SS</td>
<td>Error MS</td>
<td></td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>((c - 2))</td>
<td>lof SS</td>
<td>lof MS</td>
<td>lof MS/ pe MS</td>
</tr>
<tr>
<td>Pure Error</td>
<td>((N - c))</td>
<td>pe SS</td>
<td>pe MS</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>((N - 1))</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Example 1.1**

The following set of (fictional) data is used to illustrate the test for lack of fit. Notice that at each \( X \) value there are three \( Y \) values. The regression line has been added to the scatter plot.

<table>
<thead>
<tr>
<th>( X )</th>
<th>2.00</th>
<th>2.00</th>
<th>2.00</th>
<th>2.50</th>
<th>2.50</th>
<th>2.50</th>
<th>3.00</th>
<th>3.00</th>
<th>3.00</th>
<th>3.50</th>
<th>3.50</th>
<th>3.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td>3.47</td>
<td>2.57</td>
<td>3.18</td>
<td>4.93</td>
<td>4.60</td>
<td>4.38</td>
<td>5.68</td>
<td>4.38</td>
<td>5.21</td>
<td>5.37</td>
<td>5.98</td>
<td>5.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( X )</th>
<th>4.00</th>
<th>4.00</th>
<th>4.00</th>
<th>4.60</th>
<th>4.60</th>
<th>4.60</th>
<th>5.00</th>
<th>5.00</th>
<th>5.00</th>
<th>5.50</th>
<th>5.50</th>
<th>5.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td>6.06</td>
<td>7.53</td>
<td>6.20</td>
<td>7.68</td>
<td>6.35</td>
<td>6.87</td>
<td>6.70</td>
<td>6.87</td>
<td>7.68</td>
<td>7.20</td>
<td>6.97</td>
<td>7.53</td>
</tr>
</tbody>
</table>

A linear regression of \( Y \) on \( X \) gives the regression equation

\[
Y = 1.48 + 1.14X
\]

which leads to the usual analysis of variance table,
Figure 1.1: Multiple Y values for each X

AOV for Regression Analysis

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
<th>F Value</th>
<th>Pr(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>41.56</td>
<td>41.56</td>
<td>106.50</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>22</td>
<td>8.59</td>
<td>0.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>50.15</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The calculations are not shown but are the usual calculations for simple linear regression.

The error term for this AOV (Within Treatments) is the PURE error term. It does not depend on the regression line fitted.

One way AOV

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Treatments</td>
<td>7</td>
<td>45.302</td>
<td>6.472</td>
</tr>
<tr>
<td>Error(Within Treatments)</td>
<td>16</td>
<td>4.845</td>
<td>0.303</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>50.147</td>
<td></td>
</tr>
</tbody>
</table>

Then we obtain

AOV – Lack of Fit

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
<th>F Value</th>
<th>Pr(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>41.56</td>
<td>41.56</td>
<td>106.50</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>22</td>
<td>8.59</td>
<td>0.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>6</td>
<td>3.74</td>
<td>0.622</td>
<td>2.06</td>
<td>0.12</td>
</tr>
<tr>
<td>Pure Error</td>
<td>16</td>
<td>4.85</td>
<td>0.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>50.15</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The lack of fit test is not significant. Hence it appears a straight line does provide a reasonable fit.

**Exercise:** Show that the Residual SS can be partitioned into a Lack fo Fit SS and a Pure Error SS. That is

\[
\text{Residual SS} = (\text{Lack of Fit SS}) + (\text{Pure Error SS}).
\]

### 1.6 Testing Parallelism of Straight Lines

If instead of one data set there are \( k \) data sets, to which \( k \) lines may be fitted, questions of interest may be:

1. Are the lines parallel?
2. If the lines are parallel are the intercepts the same?

Before considering the general situation of \( k \), \( (k > 2) \) data sets we first examine the case where there are only two data sets.

#### 1.6.1 Two Straight Lines

Given two independent sets of data (with sample sizes \( n_1 \) and \( n_2 \)) it is a simple matter to fit different regressions to each set. It is often of interest to know if the two lines are parallel, that is if they have a common slope. A \( t \)-test can be used to test the hypothesis \( \beta_1 = \beta_2 \).

To do this we first calculate the Pooled Residual SS by summing the residual SS’s from the individual regressions. Then

\[
Pooled \text{ Residual SS} = \text{Residual SS(line 1) + Residual SS(line 2)}. \\
s^2 = (\text{Pooled Residual SS})/(n_1 + n_2 - 4).
\]

\[
t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\text{se}(\hat{\beta}_1 - \hat{\beta}_2)}
\]

\[
= \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{s^2 \left( \frac{1}{\sum x_1^2} + \frac{1}{\sum x_2^2} \right)}}.
\]

\[
df = n_1 + n_2 - 4
\]

where \( \sum x_1^2 \) is the sum of squares for the \( X \)'s from data set 1 and \( \sum x_2^2 \) is the sum of squares of the \( X \)'s from data set 2.

(Revise pooling of sums of squares and the \( t \)-test for independent samples from second year.)
1.6.2 \( k \) Lines \((k > 2)\) – An Example of the General Linear Test

Now consider the situation where there are \( k \) sets of data to which (possibly) \( k \) lines may be fitted. While a \( t \)-test is useful for testing the equality of two slopes, using a large number of \( t \)-tests to carry out pairwise tests of equality of slopes can not be recommended, especially if \( k \) is large. A more general analysis is required to enable testing the hypothesis:

\[
\beta_1 = \beta_2 = \cdots = \beta_k.
\]

(Compare: the \( t \)-test and its extension to a one-way AOV from second year.)

The method used is an example of a very general testing procedure called in NWT the General Linear Test. To begin with the full or unrestricted model is first fitted. In the context of this section this is the model where \( k \)-lines are fitted each having a different intercept and a different slope.

**Case 1:** A different intercept and slope for each line, \((k \) separate lines). This is the full or unrestricted model and will have the largest regression sum of squares. Consequently, it will also have the smallest residual (error) sum of squares. The error SS for this model will be denoted by \( \text{SSE}(kI, kS) \).

**Case 2:** A common slope but a separate intercept for each line, \((k \) parallel lines). This model is more restrictive than the model in case 1 since only \( k + 1 \) parameters are fitted compared to the \( 2k \) fitted for case 1. The model is called a reduced or restricted model.

The regression sum of squares will never be greater than the regression sum of squares for case 1 and will usually be smaller. This means the error sum of squares, \( \text{SSE}(kI, 1S) \) will be larger than for case 1. If this increase in the error sum of squares is small then it is an indication that the \( k \) lines were close to parallel to begin with and that allowing each line to have a separate slope was probably unnecessary. Likewise if the increase is large then it is an indication that at least some of the \( k \) lines are not parallel and that a common slope does not provide an adequate representation of the \( k \) data sets.

**Case 3:** A common slope and common intercept, \((one \) common line). This is the most restrictive model having only 2 parameters. It has the smallest regression sum of squares of the three models and the largest error sum of squares, \( \text{SSE}(1I, 1S) \). If \( \text{SSE}(1I, 1S) \) is not much larger than \( \text{SSE}(kI, kS) \) then it is an indication that only one intercept is really required so that a common line would provide an adequate model for all \( k \) data sets.

While it is true that

\[
\text{SSE}(1I, 1S) \geq \text{SSE}(kI, 1S) \geq \text{SSE}(kI, kS)
\]

the question is whether the reduction in the error SS due to using the less restrictive model is worthwhile. Is there a significant reduction in the error SS? To do this the error SS for each of the three models must be found. This is done in the next section.
1.6.3 Residual Sums of Squares

The residual SS in each of the 3 cases listed above is now found.

**Case 1** \((k \text{ intercepts, } k \text{ slopes})\)

For this case we have Model 1:

Line 1: \(y_{1j} = \beta_0^{(1)} + \beta_1^{(1)}X_{1j} + \epsilon_{1j}, \quad j = 1, 2, \ldots, n_1\)

Line 2: \(y_{2j} = \beta_0^{(2)} + \beta_1^{(2)}X_{2j} + \epsilon_{2j}, \quad j = 1, 2, \ldots, n_2\)

\[\vdots\]

Line 1: \(y_{kj} = \beta_0^{(k)} + \beta_1^{(k)}X_{kj} + \epsilon_{kj}, \quad j = 1, 2, \ldots, n_k\)

The parameters \(\hat{\beta}_0^{(i)}\) and \(\hat{\beta}_1^{(i)}\) \((\text{for } i = 1, 2, \ldots, k)\) are estimated using the method of least squares. That is we minimize the error SS,

\[
\sum_{i=1}^{k} \sum_{j=1}^{n_i} \epsilon_{ij}^2 = \sum_{j=1}^{n_1} (y_{1j} - \beta_0^{(1)} - \beta_1^{(1)}X_{1j})^2
+ \sum_{j=1}^{n_2} (y_{2j} - \beta_0^{(2)} - \beta_1^{(2)}X_{2j})^2
+ \cdots
+ \sum_{j=1}^{n_k} (y_{kj} - \beta_0^{(k)} - \beta_1^{(k)}X_{kj})^2
\]

where there are \(n_i, \quad i = 1, 2, \ldots, k\) points in each of the \(k\) lines respectively. The total number of points will be denoted by \(N = \sum_{i=1}^{k} n_i\).

Taking the partial derivatives with respect to \(\beta_0^{(1)}, \beta_1^{(1)}, \beta_0^{(2)}, \beta_1^{(2)}, \ldots, \beta_0^{(k)}, \beta_1^{(k)}\) and equating the resulting expressions to zero leads to \(k\) separate sets of normal equations, each having two unknowns. These give the usual solutions:

\[
\hat{\beta}_1^{(j)} = \frac{\sum x_j y_j}{\sum x_j^2},
\]

\[
\hat{\beta}_0^{(j)} = \bar{y}_j - \hat{\beta}_1^{(1)} \bar{x}_j.
\]

**Note on Notation:** \(\sum x_i y_i\) and \(\sum x_i^2\) are the sum of products and sum of squares calculated from the \(i\) th set of data.

**Exercise:** Derive the estimates for the parameters in model 1 using the method of least squares.

The regression sum of squares \((k \text{ intercepts, } k \text{ slopes}), \text{SSR}(kI, kS)\) is then given by
\[
\frac{\sum x_1 y_1^2}{\sum x_1^2} + \frac{\sum x_2 y_2^2}{\sum x_2^2} + \cdots + \frac{\sum x_k y_k^2}{\sum x_k^2}
\]

and the residual sum of squares, \(\text{SSE}(kI, kS)\), is given by

\[
\text{SSE}(kI, kS) = \sum y_1^2 - \frac{\sum x_1 y_1^2}{\sum x_1^2} - \cdots - \frac{\sum x_k y_k^2}{\sum x_k^2}.
\]

**Case 2** (\(k\) intercepts and 1 (common) slope)

If \(k\) separate lines are fitted but the slopes of all \(k\) lines are restricted to have a common value of \(\beta\) then the model, Model 2, is:

Line 1: \(y_{1j} = \beta_{0}^{(1)} + \beta_1 X_{1j} + \epsilon_{1j}, \quad j = 1, 2, \ldots, n_1,\)

Line 2: \(y_{2j} = \beta_{0}^{(2)} + \beta_1 X_{2j} + \epsilon_{2j}, \quad j = 1, 2, \ldots, n_2,\)

\[\vdots\]

Line \(k\): \(y_{kj} = \beta_{0}^{(k)} + \beta_1 X_{kj} + \epsilon_{kj}, \quad j = 1, 2, \ldots, n_k,\)

The method of least squares is now used to minimize the error SS,

\[
\sum_i \sum_j \epsilon_{ij}^2 = \sum_{j=1}^{n_1} (y_{1j} - \beta_{0}^{(1)} - \beta_1 X_{1j})^2
\]

\[+ \sum_{j=1}^{n_2} (y_{2j} - \beta_{0}^{(2)} - \beta_1 X_{2j})^2
\]

\[+ \cdots
\]

\[+ \sum_{j=1}^{n_k} (y_{kj} - \beta_{0}^{(k)} - \beta_1 X_{kj})^2
\]

with respect to \(\beta_{0}^{(1)}, \beta_{0}^{(2)}, \ldots, \beta_{0}^{(k)}\) and \(\beta_1\). Taking partial derivatives, equating the resulting expressions to zero and solving the \(k + 1\) normal equations so obtained gives the following estimates for the \(k + 1\) parameters

\[
\hat{\beta}_{0}^{(1)} = \bar{y}_1 - \hat{\beta}_1 \bar{X}_1
\]

\[\vdots
\]

\[
\hat{\beta}_{0}^{(k)} = \bar{y}_k - \hat{\beta}_1 \bar{X}_k
\]

\[
\hat{\beta}_1 = \frac{\left(\sum x_1 y_1 + \sum x_2 y_2 + \cdots + \sum x_k y_k\right)}{\sum x_1^2 + \sum x_2^2 + \cdots + \sum x_k^2}.
\]

**Exercise:** Verify the above results for Model 2.
The residual sum of squares, \( \text{SSE}(kI, 1S) \), is then given by

\[
\text{SSE}(kI, 1S) = \sum y_1^2 + \sum y_2^2 + \cdots + \sum y_k^2 - \frac{(\sum x_1 y_1 + \sum x_2 y_2 + \cdots + \sum x_k y_k)^2}{\sum x_1^2 + \sum x_2^2 + \cdots + \sum x_k^2}.
\]

**Case 3** (Common intercept, common slope (1 intercept, 1 slope))

In this case one overall line is being fitted to the data. This involves fitting one \( \beta_0 \) and one \( \beta_1 \), essentially treating the data as one combined set. As in simple linear regression this leads to the residual sum of squares, \( \text{SSE}(1I, 1S) \) given by

\[
\text{SSE}(1I, 1S) = \sum y^2 - \frac{\left(\sum xy\right)^2}{\sum x^2}
\]

where all \( N \) data points have been used in the calculation.

### 1.6.4 Degrees of Freedom

For case 1, \( 2k \) parameters (\( k \) intercepts and \( k \) means) are estimated. Thus, the degrees of freedom for the residual SS, \( \text{SSE}(kI, kS) \), will be \( N - 2k \).

In a similar fashion, since in case 2, \( (k + 1) \) parameters (\( k \) intercepts and 1 slope) are estimated, the degrees of freedom for \( \text{SSE}(kI, 1S) \) will be \( N - (k + 1) \).

For the final case where a common slope and intercept are estimated the degrees of freedom for \( \text{SSE}(1I, 1S) \) will be \( N - 2 \).

### 1.6.5 Hypotheses and Tests

Two hypotheses are of interest.

1. **Test of Parallelism**

   The hypothesis is that all the slopes are equal, that is:

   \[
   \beta_1^{(1)} = \beta_1^{(2)} = \cdots = \beta_1^{(k)}.
   \]

   To test this hypothesis the reduction in residual SS technique discussed in 1.6.2 is utilised. The test statistic is

   \[
   \frac{[\text{SSE}(kI, 1S) - \text{SSE}(kI, kS)]/(k - 1)}{\text{SSE}(kI, kS)/(N - 2k)}
   \]

   which is distributed as an \( F \) with \( (k - 1) \) and \( (N - 2k) \) degrees of freedom. Note that the numerator is a measure of the reduction in residual SS due to allowing each line to have a separate slope rather than forcing them to be parallel. The degrees of freedom for the numerator is the difference in the degrees of freedom of \( \text{SSE}(kI, 1S) \) and \( \text{SSE}(kI, kS) \), that is \( [N - (k + 1)] - [N - 2k] = k - 1 \). A significant result provides evidence that more than one slope is required and that the hypothesis appears incorrect.
2. Test of Equal Intercepts

The hypothesis is that the $k$ intercepts are equal, that is

$$\beta_0^{(1)} = \beta_0^{(2)} = \cdots = \beta_0^{(k)}.$$

The test statistic appropriate for testing this hypothesis is

$$\frac{[\text{SSE}(1I, 1S) - \text{SSE}(kI, 1S)]/(k - 1)}{\text{SSE}(kI, 1S)/[N - (k + 1)]}$$

which is distributed as an $F$ with $(k - 1)$ and $[N - (k + 1)]$ degrees of freedom where the numerator degrees of freedom are obtained from the difference in the degrees of freedom of $\text{SSE}(1I, 1S)$ and $\text{SSE}(kI, 1S)$. That is $[N - 2] - [N - (k + 1)] = k - 1$. Note the numerator is simply a measure of the reduction in the residual SS if rather than forcing a common line through the data, $k$ parallel lines are allowed, each being allowed a separate intercept. A significant result implies that the common line does not provide an adequate fit to the data.

Note: The test of parallelism (equal slopes) is carried out first. If this test is significant it is not usual to proceed further since the intercepts will almost certainly be different due to the unequal slopes (see diagram below), the exception being when there is a theoretical reason to believe that the point where the individual lines intersect should be on the $Y$–axis.

Figure 1.2: Parallelism
Parallel - Test Intercepts          Not Parallel - Don’t Test Intercepts
Example 1.2

The age (X) and concentration of cholesterol (Y) (mg/100ml) in the blood serum of women in Iowa and Nebraska is recorded in the table below. The question to be answered is whether the linear regressions of cholesterol on age is the same in Iowa and Nebraska. (Source: Snedecor GW and Cochran WG, *Statistical Methods* (7th ed) Iowa State Uni. Press, Page 386.)

**Iowa**

<table>
<thead>
<tr>
<th>Age</th>
<th>33</th>
<th>39</th>
<th>41</th>
<th>46</th>
<th>49</th>
<th>52</th>
<th>54</th>
<th>58</th>
<th>65</th>
<th>71</th>
<th>76</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chol</td>
<td>201</td>
<td>182</td>
<td>112</td>
<td>181</td>
<td>121</td>
<td>228</td>
<td>259</td>
<td>189</td>
<td>249</td>
<td>224</td>
<td>339</td>
</tr>
</tbody>
</table>

**Nebraska**

<table>
<thead>
<tr>
<th>Age</th>
<th>18</th>
<th>19</th>
<th>21</th>
<th>30</th>
<th>31</th>
<th>33</th>
<th>42</th>
<th>43</th>
<th>44</th>
<th>44</th>
<th>47</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chol</td>
<td>137</td>
<td>189</td>
<td>191</td>
<td>140</td>
<td>159</td>
<td>177</td>
<td>214</td>
<td>223</td>
<td>190</td>
<td>173</td>
<td>196</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>51</th>
<th>56</th>
<th>58</th>
<th>58</th>
<th>63</th>
<th>67</th>
<th>70</th>
<th>78</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chol</td>
<td>225</td>
<td>197</td>
<td>262</td>
<td>257</td>
<td>337</td>
<td>356</td>
<td>261</td>
<td>241</td>
</tr>
</tbody>
</table>

**Figure 1.3: Scatter Plot of Cholesterol against Age**
# ______ cholesterol.r _____________
options(digits=2)

Age <- c(33,39,41,46,49,52,54,58,65,71,76,
18,19,21,30,31,33,42,43,44,44,47,51,56,58,58,63,67,70,78)
Chol <- c(201,182,112,181,121,228,259,189,249,224,339,
137,189,191,140,159,177,214,223,190,173,196,225,197,262,257,337,356,261,241)
State <- rep(c("Iowa","Nebraska"),c(11,19))
choldf <- data.frame(State, Age, Chol)

library(lattice)
trellis.device(color=F)
xyp <- xyplot(Chol ~ Age|State,data=choldf,
panel=function(x,y){panel.xyplot(x,y)
panel.abline(lsfit(x,y),lty=1)})
print(xyp)
dev.copy2eps(file="chol_regression.ps",horizontal=F,width=6,height=5)

# ______ 2 intercepts and 2 slopes _________
model11 <- lm(Chol ~ State/Age,data=choldf)
# ______ 2 intercepts and 1 slope _________
model12 <- lm(Chol ~ State + Age,data=choldf)
# ______ 1 intercept and 1 slope1 _________
model13 <- lm(Chol ~ Age,data=choldf)
# ______ test slopes _____________
model14 <- lm(Chol ~ State + Age + State:Age,data=choldf)

sink("cholesterol_results.txt")
cat("________ 2 intercepts and 2 slopes ___________ \n\n")
print(anova(model1))
print(summary(model11)$coefficients)
cat("________ 2 intercepts and 1 slope ___________ \n\n")
print(anova(model2))
print(summary(model12)$coefficients)
cat("________ 1 intercept and 1 slope ___________ \n\n")
print(anova(model3))
print(summary(model13)$coefficients)
print(confint(model3,level=0.95))
cat("________ test slopes ___________\n\n")
print(anova(model4))
print(summary(model4)$coefficients)
sink()
2 intercepts and 2 slopes 

Response: Chol

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>1</td>
<td>613</td>
<td>613</td>
<td>0.33</td>
</tr>
<tr>
<td>State:Age</td>
<td>2</td>
<td>54529</td>
<td>27265</td>
<td>14.65</td>
</tr>
<tr>
<td>Residuals</td>
<td>26</td>
<td>48395</td>
<td>1861</td>
<td>Estimate Std. Error t value Pr(&gt;</td>
</tr>
</tbody>
</table>

(Intercept) 35.8 55.12 0.65 0.52156
StateNebraska 65.5 61.98 1.06 0.30045
StateIowa:Age 3.2 1.01 3.21 0.00352
StateNebraska:Age 2.5 0.58 4.36 0.00018

2 intercepts and 1 slope 

Response: Chol

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>1</td>
<td>613</td>
<td>613</td>
<td>0.34</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>53820</td>
<td>53820</td>
<td>29.59</td>
</tr>
<tr>
<td>Residuals</td>
<td>27</td>
<td>49104</td>
<td>1819</td>
<td>Estimate Std. Error t value Pr(&gt;</td>
</tr>
</tbody>
</table>

(Intercept) 64.5 29.3 2.2 3.6e-02
StateNebraska 28.7 16.5 1.7 9.5e-02
Age 2.7 0.5 5.4 9.4e-06

1 intercept and 1 slope 

Response: Chol

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1</td>
<td>48976</td>
<td>48976</td>
<td>25.1</td>
</tr>
<tr>
<td>Residuals</td>
<td>28</td>
<td>54560</td>
<td>1949</td>
<td>Estimate Std. Error t value Pr(&gt;</td>
</tr>
</tbody>
</table>

(Intercept) 91.6 25.7 3.6 1.3e-03
Age 2.5 0.5 5.0 2.7e-05

2.5 % 97.5 %

(Intercept) 39.0 144.1
Age 1.5 3.5
test slopes

Response: Chol

|          | Df | Sum Sq | Mean Sq | F value  | Pr(>|F|) |
|----------|----|--------|---------|----------|----------|
| State    | 1  | 613    | 613     | 0.33     | 0.57     |
| Age      | 1  | 53820  | 53820   | 28.91    | 1.2e-05  *** |
| State:Age| 1  | 709    | 709     | 0.38     | 0.54     |
| Residuals| 26 | 48395  | 1861    |          |          |

| Estimate   | Std. Error | t value | Pr(>|t|) |
|------------|------------|---------|----------|
| (Intercept)| 35.81      | 55.1    | 0.65     | 0.5216   |
| StateNebraska | 65.49 | 62.0    | 1.06     | 0.3005   |
| Age       | 3.24       | 1.0     | 3.21     | 0.0035   |
| StateNebraska:Age | -0.72 | 1.2     | -0.62    | 0.5425   |

2.5 % 97.5 %

(Intercept) | -77.5 | 149.1 |
StateNebraska | -61.9 | 192.9 |
Age       | 1.2    | 5.3   |
StateNebraska:Age | -3.1  | 1.7   |
Chapter 2

Matrix Approach to Regression

2.1 Multiple Linear Regression

In Chapter 1 only simple linear regression was considered. The method of least squares was used to fit a straight line of best fit, which had the form \( \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \). In this chapter simple linear regression is generalised to multiple linear regression, that is \( p \) independent variates, \( X_1, X_2, \ldots, X_p \) will now be considered and the model will be of the form

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i .
\]  

(2.1)

For the special case where only one independent variable, \( X_1 \), is considered then a least squares line is fitted to the data. When two independent variables, \( X_1 \) and \( X_2 \) are considered, a plane is fitted to the data.

As in chapter 1, it will be assumed that the \( \epsilon_i \) are independently distributed \( N(0, \sigma^2) \). The data will consist of \( n \), \((p + 1)\) tuples

\[(Y_i, X_{i1}, X_{i2}, \ldots, X_{ip}), \text{ for } i = 1, 2, \ldots, n.\]

The parameters in the model are estimated by minimizing the error sum of squares, \( \Sigma \epsilon_i^2 \), that is

\[
\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip})\}^2 .
\]

As in Chapter 1 the Normal Equations are found by taking the partial derivatives,

\[
\frac{\partial (\sum_{i=1}^{n} \epsilon_i^2)}{\partial \beta_j} \text{ for } j = 1, 2, \ldots, p,
\]

and putting the resultant expressions equal to zero to get
\[
\begin{align*}
n\hat{\beta}_0 & + \sum X_{i1}\hat{\beta}_1 & + \cdots + \sum X_{ip}\hat{\beta}_p & = \sum Y_i \\
\sum X_{i1}\hat{\beta}_0 & + \sum X_{i1}^2\hat{\beta}_1 & + \cdots + \sum X_{i1}X_{ip}\hat{\beta}_p & = \sum X_{i1}Y_i \\
\vdots & \vdots & \vdots & \vdots \\
\sum X_{ip}\hat{\beta}_0 & + \sum X_{i1}X_{ip}\hat{\beta}_1 & + \cdots + \sum X_{ip}^2\hat{\beta}_p & = \sum X_{ip}Y_i .
\end{align*}
\]

The solution to these equations, \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p \) are then least squares estimates of the \( p + 1 \) parameters \( \beta_0, \beta_1, \ldots, \beta_p \).

**Exercise:** Verify that taking partial derivatives and equating the resultant expressions to zero does give the normal equations (2.2) above.

From the first of the normal equations (2.2) we obtain

\[
\hat{\beta}_0 = \frac{\sum Y - \hat{\beta}_1\sum X_1 - \hat{\beta}_2\sum X_2 - \cdots - \hat{\beta}_p\sum X_p}{n} .
\]

Substituting for \( \hat{\beta}_0 \) in the remaining equations then gives \( p \) equations in \( p \) unknowns of the form

\[
\begin{align*}
\sum x_1^2\hat{\beta}_1 + \sum x_1x_2\hat{\beta}_2 + \cdots + \sum x_1x_p\hat{\beta}_p & = \sum x_1y \\
\sum x_2x_1\hat{\beta}_1 + \sum x_2^2\hat{\beta}_2 + \cdots + \sum x_2x_p\hat{\beta}_p & = \sum x_2y \\
\vdots & \vdots & \vdots & \vdots \\
\sum x_px_1\hat{\beta}_1 + \sum x_px_2\hat{\beta}_2 + \cdots + \sum x_p^2\hat{\beta}_p & = \sum x_py .
\end{align*}
\]

Note: \( \sum x_ix_j = \sum x_jx_i = \sum_{k=1}^n (X_{j,k} - \bar{X}_j)(X_{i,k} - \bar{X}_i) \).

### 2.1.1 Notation

Before considering the matrix approach to linear models it is worthwhile introducing some notation for use when discussing regression and error (residual) sums of squares. The notation will be defined for the regression model with two independent variables \( X_1 \) and \( X_2 \), that is,

\[
Y_k = \beta_0 + \beta_1X_{k1} + \beta_2X_{k2} + \epsilon_i
\]

but the reader should have little difficulty extending this to more general cases.

We will use the following notation for regression and residual sums of squares. For the two variables \( X_1 \) and \( X_2 \), \( SSR(X_1,X_2) \) denotes the regression sum of squares and \( SSE(X_1,X_2) \) the error(or residual) sum of squares for the model where both variables are included. Similarly, \( SSR(X_1) \) and \( SSE(X_1) \) refer to the model where only \( X_1 \) has been included. (This notation can easily be extended to include more than two variables in the obvious way.)

The increase in the regression sum of squares (or equivalently the reduction in the error
SS) due to fitting $X_2$ after $X_1$ has been fitted is then denoted by

$$\text{SSR}(X_2|X_1) = \text{SSE}(X_1) - \text{SSE}(X_1,X_2) = \text{SSR}(X_1,X_2) - \text{SSR}(X_1).$$

It is not difficult to extend this notation to the situation where more than two independent variables are being considered. For example

$$\text{SSR}(X_3|X_1,X_2) = \text{SSE}(X_1,X_2) - \text{SSE}(X_1,X_2,X_3)$$

represents the additional regression sum of squares due to $X_3$ having already fitted $X_1$ and $X_2$.

**Important Note:** The model contains a constant and hence $\text{SSR}(X_1)$ is actually the additional SS due to fitting $X_1$ given that the constant ($\beta_0$) has already been fitted. Similarly $\text{SSR}(X_2|X_1)$ is really the additional SS due to fitting $X_2$ given that both the constant and $X_1$ have already been fitted. We will follow the practice of omitting reference to the constant except where the omission may cause confusion.

Extensive use will be made of this notation throughout these notes without further comment.

### 2.1.2 Matrix Approach

So far no use has been made of matrices. They provide a much neater method of writing the equations above as well as reducing the work involved in deriving many of the results. Before proceeding with this section it would be worthwhile revising the material on the distribution of quadratic forms from Statistics 354. (For a discussion of matrices and matrix methods in regression see NWK.)

We will follow the convention of writing vectors and matrices in bold type with a prime (′) denoting the transpose of a matrix.

Consider the multiple regression model (2.1) when there are $n$ observations, $y_1, y_2, \ldots y_n$. Writing the $n$ equations out in full we see that
\[ Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \epsilon_1 \]
\[ Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} + \epsilon_2 \]
\[ \vdots \]
\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i \]
\[ \vdots \]
\[ Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} + \epsilon_n \]

which can be written (using matrix notation) as

\[ Y = X\beta + \epsilon. \]  

(2.4)

Here,

- \( Y \) is an \((n \times 1)\) vector of observations,
- \( X \) is the \((n \times p)\) design matrix where

\[
X = \begin{bmatrix}
1 & X_{11} & X_{12} & \cdots & X_{1p} \\
1 & X_{21} & X_{22} & \cdots & X_{2p} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & X_{n1} & X_{n2} & \cdots & X_{np}
\end{bmatrix},
\]

- \( \beta \) is the \((p+1) \times 1\) vector of parameters, \((\beta_0, \beta_1, \ldots, \beta_p)'\), with ‘ indicating the transpose of the vector (or matrix) and
- \( \epsilon \) is the \((n \times 1)\) vector of residuals (errors).

### 2.1.3 Results Using Matrices

Results (parallelling those in 1.2.1) will now be derived using matrices.

**Result 2.1** Least squares estimates of the parameters \((\beta_0, \beta_1, \ldots, \beta_p)\) are given by,

\[ \hat{\beta} = (X'X)^{-1}X'Y. \]  

(2.5)

(Assume that \( X \) is of full rank so that the inverse in (2.5) exists. The least squares estimates are obtained by differentiating the error sum of squares with respect to \( \beta_0, \beta_1, \beta_2, \ldots, \beta_p \) and equating the resulting expressions to zero to produce the normal equations (2.2). The normal equations can then be written in matrix form as

\[ 2X'X \hat{\beta} - 2X'Y = 0. \]  

(2.6)
Thus,
\[
\hat{\beta} = ((X'X)^{-1}X'Y) .
\] (2.7)

**Note:** The error sum of squares, \( \epsilon' \epsilon = \sum_{i=1}^{n} \epsilon_i^2 \), can be written as
\[
\epsilon' \epsilon = (Y - X\beta)'(Y - X\beta) = Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta = Y'Y - 2\beta'X'Y + \beta'X'X\beta .
\]

(Note: all the above quantities are scalars (1x1 vectors) so they are unaffected by taking transposes.)

For the next result the following definition is required.

**Definition** A symmetric matrix \( A \) is positive definite if the quadratic form \( y' Ay > 0 \) for every (non-zero) vector, \( y \).

**Result 2.2** \( X'X \) is non-negative.
\[
Y'X'XY = (Y'X')(XY) = (XY)'(XY) .
\]

This last quantity is positive for any (non-zero) vector \( Y \) since it is a sum of squares.

**Result 2.3** \( \hat{\beta} \) is unbiased.
\[
E(\hat{\beta}) = E\{(X'X)^{-1}X'Y\} = E\{(X'X)^{-1}X'(X\beta + \epsilon)\} = E(\beta + (X'X)^{-1}X'\epsilon) = \beta .
\]

since under the assumption that \( E(\epsilon_i) = 0 \), \( E((X'X)^{-1}X'\epsilon) = 0 \).

**Result 2.4** The Variance–Covariance matrix of \( \hat{\beta} \) is given by
\[
\sigma^2(\hat{\beta}) = \sigma^2(X'X)^{-1} .
\] (2.8)

By definition
\[
\sigma^2(\hat{\beta}) = E\{(\hat{\beta} - E(\hat{\beta}))\{\hat{\beta} - E(\hat{\beta})\}' = E\{(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\}.
\] but,
\[
\hat{\beta} - \beta = (X'X)^{-1}X'Y - \beta = (X'X)^{-1}X'(X\beta + \epsilon) - \beta = \beta + (X'X)^{-1}X'\epsilon - \beta = (X'X)^{-1}X'\epsilon .
\]
Thus

\[
\sigma^2(\hat{\beta}) = E\{(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\} \\
= (X'X)^{-1}X'(E[\varepsilon\varepsilon'])X(X'X)^{-1}
\]

(since \(X'X\) is symmetric so that \(((X'X)^{-1})' = (X'X)^{-1}\)).

But it is assumed the \(\varepsilon_i\) are independent and have constant variance \(\sigma^2\) and so

\[
\sigma^2(\hat{\beta}) = (X'X)^{-1}X'(I_n)\sigma^2X(X'X)^{-1} \\
= \sigma^2(X'X)^{-1}.
\]

**Result 2.5** \(\hat{\beta}\) is a best linear unbiased estimate (BLUE estimate).

(Of all unbiased estimators which are linear in the observations, \(Y_i, \hat{\beta}\) has the smallest variance.)

Let \(AY\) be another linear estimate where \(A = (X'X)^{-1}X' + B\). If \(AY\) is an unbiased estimator of \(\beta\) then

\[
E(AY) = E\{(X'X)^{-1}X'Y + BY\} \\
= E\{(X'X)^{-1}X'(X\beta + \varepsilon) + B(X\beta + \varepsilon)\} \\
= \beta + BX\beta - 0 \\
= \beta.
\]

This implies that \(BX\) must be the null matrix, that is

\[
AY = \beta + (X'X)^{-1}X'\varepsilon + BX\beta + B\varepsilon \\
= \beta + \{(X'X)^{-1}X' + B\}\varepsilon.
\]

Then, the var-cov matrix of \(AY\) is given by

\[
\sigma^2(AY) = E\{(AY - \beta)(AY - \beta)'\} \\
= E\{(X'X)^{-1}X' + B\}\varepsilon\varepsilon'\{(X'X)^{-1}X' + B\}' \\
= \sigma^2\{(X'X)^{-1}X'X(X'X)^{-1} + (X'X)^{-1}X'B' + BX(X'X)^{-1} + BB'\} \\
= \sigma^2\{(X'X)^{-1} + BB'\}.
\]

Now, the diagonal elements of \(BB'\) are all positive since they are sums of squares of rows of \(B\). Thus the variance of all linear unbiased estimators is a minimum when the elements of the rows of \(B\) are all zero. This proves the result.

**Result 2.6** The Regression Sum of Squares due to fitting \(\beta\) is

\[
SSR(\beta) = Y'X(X'X)^{-1}X'Y.
\] (2.9)
Regression SS  =  Total SS - Residual SS

=  \( Y'Y - (Y - X \hat{\beta})'(Y - X \hat{\beta}) \)

=  \( Y'Y - Y'Y + Y'X \hat{\beta} + \hat{\beta}' X'Y - \hat{\beta}' X'X \hat{\beta} \).

Using the normal equations, (2.6), it follows that

\[ \hat{\beta}' (X'X) \hat{\beta} = \hat{\beta}' X'Y \]

giving the Regression SS = \( Y'X \hat{\beta} = \hat{\beta}' X'Y \), from which the result follows by substituting for \( \hat{\beta} \) from (2.7).

2.1.4 Partitioning the Total SS

We can partition the Total SS, \( Y'Y \), into two components the regression SS and residual SS. That is

\( Y'Y = \text{Regression SS} + \text{Residual SS} \)

which can be written in matrix notation as,

\[ Y'Y = Y'X(X'X)^{-1}X'Y + Y'(I - X(X'X)^{-1}X')Y. \] (2.10)

If we consider only the matrices involved in the quadratic forms we then have

\[ I = X(X'X)^{-1}X' + (I - X(X'X)^{-1}X'). \]

Both the matrices on the right hand side of the equation are idempotent, (i.e., of the form \( A^2 = A \)), and so we can use Cochran's theorem to find the distribution of the components of the total SS. This will then enable the construction of tests of significance of hypotheses concerning \( \beta_1, \beta_2, \ldots, \beta_p \).

Cochran’s theorem was stated and proved in Statistics 354, revise it before you proceed. It is stated here for convenience.

2.1.5 Cochran’s Theorem

(Theorem 3.6 from Distribution Theory notes.)

Given \( X \sim N_p(0, I) \), suppose that \( X'X \) is decomposed into \( k \) quadratic forms, \( Q_i = X'B_iX \), \( i = 1, 2, \ldots, k \), where the rank of \( B_i \) is \( r_i \) and the \( B_i \) are positive semidefinite, then any one of the following conditions implies the other two:
(a) the ranks of the $Q_i$ add to $p$;

(b) each $Q_i \sim \chi^2_{r_i}$;

(c) all the $Q_i$ are mutually independent.

**Note:** Cochran's Theorem can be extended to apply in cases where $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$. (See for example the statement of Cochran's Theorem in NWK.)

### 2.2 Distribution of the Regression and Residual SS

#### 2.2.1 Distribution assuming $\beta = 0$

When $\beta = 0$, $Y \sim N(0, \sigma^2 I)$ and Cochran's Theorem can be used to find the distribution of the regression sum of squares. Now

\[
Y'Y = Y'X(X'X)^{-1}X'Y + Y'(I - X(X'X)^{-1}X')Y
\]

and both $X(X'X)^{-1}X'$ and $(I - X(X'X)^{-1}X')$ are idempotent. Hence, under the hypothesis, $\beta = 0$, the regression SS is distributed as a $\sigma^2 \chi^2$ variate with degrees of freedom given by

\[
\text{trace}(X(X'X)^{-1}X') = \text{trace}(X'X(X'X)^{-1}) = \text{trace}(I) = p.
\]

When $\beta = 0$, residual SS is also distributed as a $\sigma^2 \chi^2$ and since condition (b) of Cochran's Theorem holds we see the degrees of freedom are $n - p$ and the residual SS is distributed independently of the regression SS. The usual statistic for testing $\beta = 0$ is the ratio of the regression mean square to the residual mean square which is then distributed as an $F_{p,n-p}$, since $F$ is the ratio of independent $\chi^2$'s.

#### 2.2.2 Distribution assuming $\beta \neq 0$

If $\beta$ is not equal to 0, the distribution of the residual remains $\sigma^2 \chi^2_{n-p}$ since

\[
Y'(I - X(X'X)^{-1}X')Y = (\beta'X' + \epsilon')(I - X(X'X)^{-1}X')(X\beta + \epsilon) = \epsilon'(I - X(X'X)^{-1}X')\epsilon
\]

with both $\beta'X'(I - X(X'X)^{-1}X')$ and $(I - X(X'X)^{-1}X')X\beta$ being equal to 0. But,

\[
E(\text{Regression SS}) = E[(\beta'X' + \epsilon'(X(X'X)^{-1}X')(X\beta + \epsilon)]
\]

\[
= E[\beta'X'X\beta + \beta'X'\epsilon + \epsilon'X\beta + \epsilon'(X(X'X)^{-1}X')\epsilon]
\]

\[
= \beta'X'X\beta + E(\epsilon'X(X'X)^{-1}X'\epsilon)
\]

25
\[
\begin{align*}
&= \beta' X' X \beta + E(\text{trace}(\epsilon' X (X' X)^{-1} X' \epsilon)) \quad \text{since scalar} \\
&= \beta' X' X \beta + E(\text{trace}(X (X' X)^{-1} X' \epsilon \epsilon')) \\
&= \beta' X' X \beta + \text{trace}(X (X' X)^{-1} X' \sigma^2) \\
&= \beta' X' X \beta + p \sigma^2.
\end{align*}
\]

Thus if \( \beta \) is not \( 0 \) the distribution of the test statistic has a non-central \( F \) as it is the ratio of a non-central \( \chi^2 \) to a (central) \( \chi^2 \).

**Example 2.2.1 - Simple Linear Regression**

As in Chapter 1 the model is

\[ Y_i = \beta_0 + \beta_1 X_{ii} + \epsilon_i, \quad i = 1, 2, \ldots, n. \tag{2.11} \]

which can be written in matrix notation as

\[ Y = X \beta + \epsilon \]

where

\[ Y' = (Y_1, Y_2, \ldots, Y_n)', \quad \beta' = (\beta_0, \beta_1)' \quad \text{and} \quad \epsilon' = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)' \]

with \('\) indicating the transpose of the vector as usual. Further

\[
X = \begin{bmatrix}
1 & X_{11} \\
1 & X_{21} \\
\vdots & \\
1 & X_{n1}
\end{bmatrix}
\]

Thus we get that

\[
X' X = \begin{bmatrix}
\sum X_i & \sum X_i^2 \\
\sum X_i & \sum X_i^2 \\
\sum X_i & \sum X_i^2 \\
\sum X_i & \sum X_i^2
\end{bmatrix}, \quad (X' X)^{-1} = \begin{bmatrix}
\frac{\sum X_i^2 - \sum X_i}{n} & -\frac{-\sum X_i}{n} \\
-\frac{\sum X_i}{n} & \frac{\sum X_i}{n}
\end{bmatrix}, \quad X'_i Y = \begin{bmatrix}
\sum Y \\
\sum X_i Y
\end{bmatrix}
\]

Thus,

\[
\hat{\beta} = (X' X)^{-1} X'_i Y = \begin{bmatrix}
\frac{\sum Y \sum X_i^2 - \sum X_i \sum X_i Y}{n \sum X_i^2 - (\sum X_i)^2} \\
\frac{-\sum X_i \sum Y + n \sum X_i Y}{n \sum X_i^2 - (\sum X_i)^2}
\end{bmatrix} = \begin{bmatrix}
\hat{\beta}_0 \\
\hat{\beta}_1
\end{bmatrix},
\]

which gives (after some simplification) that \( \hat{\beta}_1 = \frac{\sum xy}{\sum x^2} \) and \( \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 \).

Thus, the regression SS due to fitting both the constant and \( X_1 \), is
\[ Y'X \hat{\beta} = n\bar{Y}^2 + \frac{(\sum xy)^2}{\sum x^2}. \]

Thus we can partition the SS, \( Y'Y \) as
\[
Y'Y = \sum Y^2 = n\bar{Y}^2 + \frac{(\sum xy)^2}{\sum x^2} + \{ \sum (Y - \bar{Y})^2 - \frac{(\sum xy)^2}{\sum x^2} \}. 
\]

**Hypothesis Testing in Simple Linear Regression**

Testing \( \beta_1=0 \) is another example of the general linear test. The linear regression model, (2.11), can be thought of as the **full** model in this example. The **restricted** model is the linear model with \( \beta_1=0 \), that is
\[
Y_i = \beta_0 + \epsilon_i. 
\]

For this model \( X'=1, 1, \ldots, 1 \), \( X'X=n \) and \( X'Y = \sum Y \). Thus for this model \( \hat{\beta}_0 = \bar{Y} \) and the regression sum of squares due to fitting \( \beta_0 \) alone is \( n\bar{Y}^2 \). Thus the additional SS due to fitting \( X_1 \), \( \text{SSR}(X_1) \), is just the difference
\[
\text{SSR}(X_1) = n\bar{Y}^2 + \frac{(\sum xy)^2}{\sum x^2} - n\bar{Y}^2 = \frac{(\sum xy)^2}{\sum x^2}. 
\]

The regression SS for the full model therefore consists of two components

(a) \( n\bar{Y}^2 \) = SS due to fitting \( X_0 \) alone and

(b) \( \frac{(\sum xy)^2}{\sum x^2} = \text{SSR}(X_1) \) = additional SS due to fitting \( \beta_1 \) (after having fitted \( \beta_0 \)).

It is a common practice to subtract \( n\bar{Y}^2 \) from \( Y'Y \) to get the (corrected) SS, \( \sum y^2 \), on \( n-1 \) degrees of freedom so that
\[
\sum Y_i^2 - n\bar{Y}^2 = \frac{(\sum xy)^2}{\sum x^2} + \{ \sum (Y - \bar{Y})^2 - \frac{(\sum xy)^2}{\sum x^2} \}. 
\]

That is,
\[
\sum y^2 = \text{SSR}(X_1) + \text{SSE}(X_1). 
\]

This leads to the AOV table in result 10 in (1.2.1) and the usual test of the hypothesis \( \beta_1=0 \).

**Note:** Considerable simplification can be achieved in Example 2.2.1 by using the alternative form of the regression model given by
\[
Y_i = \beta_0^* + \beta_1(X - \bar{X}) + \epsilon_i. 
\]

That is subtract \( \bar{X} \) from the X’s before forming the matrix \( X \).


2.3 Testing $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

In this section the test of $\beta_1 = 0$ is extended to test the more general hypothesis

$$\beta_1 = \beta_2 = \ldots = \beta_p = 0.$$ 

The regression SS’s and Residual SS have been derived for the model

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

and can be summarized in an AOV table as

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Df</th>
<th>Sum of Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$p$</td>
<td>$Y'X\hat{\beta}$</td>
</tr>
<tr>
<td>Residual</td>
<td>$n - p$</td>
<td>$Y'Y - Y'X\hat{\beta}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n$</td>
<td>$Y'Y$</td>
</tr>
</tbody>
</table>

As was seen in example 2.2.1 the regression SS for fitting the model

$$Y_i = \beta_0 + \epsilon_i$$

is just the correction term $n\bar{Y}^2$ and so the additional SS due to fitting $\beta_1, \beta_2, \ldots, \beta_p$ is given by

$$Y'X\hat{\beta} - n\bar{Y}^2$$
on $p$ degrees of freedom. This leads to an $F$-test of the hypothesis

$$\beta_1 = \beta_2 = \ldots = \beta_p = 0$$

against the alternative that not all $\beta_i$'s are equal to zero. The $F$ test is on $p$ and $n - p$ degrees of freedom. In terms of an AOV table this can be summarized as follows.

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$p$</td>
<td>$Y'X\hat{\beta} - n\bar{Y}^2 = \text{SSR}(X_1, X_2, \ldots, X_p)$</td>
</tr>
<tr>
<td>Residual</td>
<td>$n - p$</td>
<td>$Y'Y - Y'X\hat{\beta} = \text{SSE}(X_1, X_2, \ldots, X_p)$</td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>$Y'Y - n\bar{Y}^2$</td>
</tr>
</tbody>
</table>

This is a test of the utility of the overall model. It essentially tests whether the overall model is contributing information for the prediction of $Y$. It makes no statement about individual $\beta$’s and whether all the $\beta$’s are really required. This question will be studied later.
Example 2.3.1: Regression with Two Variables.

Numerical Calculations

The following table gives the log(diameter), (X1), log(height), (X2), and log(volume), (Y), of trees. The aim is to predict volume from diameter and height.  

<table>
<thead>
<tr>
<th>log(Diam)</th>
<th>2.12</th>
<th>2.15</th>
<th>2.17</th>
<th>2.35</th>
<th>2.37</th>
<th>2.38</th>
<th>2.40</th>
<th>2.40</th>
<th>2.41</th>
<th>2.42</th>
<th>2.42</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(Vol)</td>
<td>2.33</td>
<td>2.33</td>
<td>2.32</td>
<td>2.80</td>
<td>2.93</td>
<td>2.98</td>
<td>2.75</td>
<td>2.90</td>
<td>3.12</td>
<td>2.99</td>
<td>3.19</td>
</tr>
</tbody>
</table>

| log(Diam) | 2.43 | 2.43 | 2.46 | 2.48 | 2.56 | 2.56 | 2.59 | 2.62 | 2.62 | 2.64 | 2.65 |
| log(Height) | 4.33 | 4.33 | 4.23 | 4.32 | 4.30 | 4.44 | 4.45 | 4.26 | 4.16 | 4.36 | 4.38 |
| log(Vol) | 3.04 | 3.06 | 3.06 | 2.95 | 3.10 | 3.52 | 3.31 | 3.25 | 3.21 | 3.54 | 3.46 |

The design matrix $X$ is then given by

$X = \begin{bmatrix} 
1 & 2.12 & 4.25 \\
1 & 2.15 & 4.17 \\
1 & 2.17 & 4.14 \\
\vdots & \vdots & \vdots \\
1 & 2.89 & 4.38 \\
1 & 2.89 & 4.38 \\
1 & 3.03 & 4.47 
\end{bmatrix}$

The matrices $X'X$ and $(X'X)^{-1}$ are then

$X'X = \begin{bmatrix} 
31.000 & 79.260 & 134.120 \\
79.260 & 204.286 & 343.236 \\
134.120 & 343.236 & 580.485 
\end{bmatrix}$

$(X'X)^{-1} = \begin{bmatrix} 
96.5952 & 3.1915 & -24.2052 \\
3.1915 & 0.8554 & -1.2432 \\
-24.2052 & -1.2432 & 6.3294 
\end{bmatrix}$

and
\[
\mathbf{X}'\mathbf{Y} = \begin{bmatrix}
101.440 \\
262.961 \\
439.755
\end{bmatrix}
\]

so that

\[
\hat{\mathbf{\beta}} = \begin{bmatrix}
-6.52362 \\
1.98773 \\
1.08949
\end{bmatrix}
\]

The amount of calculation becomes prohibitive for even a relatively small data set and in practice these calculations are carried out by computer. (The above calculations can be done using the statistical package, SPlus.)

Using the above it is not difficult to show that the regression SS, \( \mathbf{Y}'\mathbf{X}\hat{\mathbf{\beta}}, \) is equal to 340.05 with the total SS, \( \mathbf{Y}'\mathbf{Y}, \) equal to 340.249. The correction term, \( n\overline{y^2}, \) is 331.94

Thus the regression equation is

\[
\log(\text{Vol}) = -6.52 + 1.99 \log(\text{Diam}) + 1.09 \log(\text{Height})
\]

and the Analysis of Variance table is as follows.

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of SqS</th>
<th>Mean Sq</th>
<th>F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>( SSR(X_1, X_2) = 8.12 )</td>
<td>( MSR(X_1, X_2) = 4.06 )</td>
<td>597</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>( SSE(X_1, X_2) = 0.19 )</td>
<td>( MSE(X_1, X_2) = 0.0068 )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>8.31</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The F value calculated here is highly significant (less than .1\% level). That is it is highly unlikely that \( \beta_1 = \beta_2 = 0. \)

### 2.4 General Linear Test-Testing Subsets of Variables

In section 2.3 the hypothesis

\[
\beta_1 = \beta_2 = \ldots = \beta_p = 0
\]

was tested against a general alternative. This test is an example of a much more general form of test which allows testing subsets of the \( k \) parameters. For example how should the hypothesis

\[
\beta_q = \beta_{q+1} = \ldots = \beta_p = 0
\]

be tested against the alternative that not all these \( \beta \)'s are zero given that the parameters \( \beta_1, \beta_2, \ldots, \beta_{q-1} \) have already been included in the model and are non-zero. The method is quite general and will be used repeatedly throughout this course. It involves examining the reduction in the residual sum of squares that occurs when the extra parameters are fitted in the model.
Begin with fitting the full or unrestricted model
\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i . \]

For convenience, denote the error sum of squares for this model by SSE(F).

Next consider the reduced or restricted model under the hypothesis
\[ \beta_q = \beta_{q+1} = \cdots = \beta_p = 0 \]

That is the model
\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{q-1} X_{i,q-1} + \epsilon_i . \]

where \( q < p \). Denote the error sum of squares for this model by SSE(R). Since more parameters are fitted in the full model than the restricted model,
\[ \text{SSE(F)} \leq \text{SSE(R)} \]

and the reduction in the sum of squares, \( \text{SSE(R)} - \text{SSE(F)} \) is a measure of whether it is worth fitting the full model in preference to the restricted model. If \( \text{SSR(F)} \) is not much less than \( \text{SSR(R)} \) then using the full model does not account for much more of the variability of \( Y_i \) than does the restricted model which suggests that the additional coefficients may well be zero. On the other hand a substantial difference would suggest that the full model is better at explaining the variability in \( Y_i \) and that the full model should be used in preference to the restricted model. That is that the alternative hypothesis that not all of the additional parameters are zero is supported.

In order to test the hypothesis the test statistic
\[ F = \frac{\text{SSE(R)} - \text{SSE(F)}}{\frac{\text{df}_R}{\text{df}_F}} \]

is used which follows the F distribution on \( p-q \) and \( n-p \) degrees of freedom provided the hypothesis holds.

Summarizing this in AOV form gives

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced model</td>
<td>q</td>
<td>SSR(R)</td>
</tr>
<tr>
<td>Additional SS due to Full Model</td>
<td>p-q</td>
<td>( \text{SSE(R)} - \text{SSE(F)} )</td>
</tr>
<tr>
<td>Residual</td>
<td>n-p</td>
<td>( \text{SSE(F)} )</td>
</tr>
<tr>
<td>Total</td>
<td>n</td>
<td>( \sum Y_i^2 )</td>
</tr>
</tbody>
</table>
2.4.1 Testing $\beta_i = 0$

The question being asked is whether a particular variable has coefficient zero in the multiple regression model,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

given that all the other variables have been fitted. We are really asking if the contribution of that variable is worthwhile, does it give additional information above that already provided by the other variables in the equation? (The individual coefficients are sometimes called **partial** regression coefficients.) The hypothesis to be tested is

$$H : \beta_i = 0, \text{ after fitting the other } p - 1 \text{ coefficients.}$$

The alternative is that $\beta_i$ is not zero. The test statistic is

$$\frac{\text{SSR}(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)/1}{\text{SSE}(X_1, X_2, \ldots, X_p)/(n - p)}.$$

Under the hypothesis this is distributed as $F(1, n - p)$.

**Example 2.3.1 (continued)**

The hypothesis

$$\beta_1 = \beta_2 = 0$$

was tested above and found to be significant. Two questions may be of interest.

1. Does $\beta_1 = 0$ if $\beta_2$ is not included in the model?

2. Does it give a significant additional amount of information about the volume($Y$) if $\beta_2$ has already been included in the model? That is does $\beta_1 = 0$ given that $\beta_2$ has already been fitted?

To answer the first question, fit log(Vol) against log(Diam) to get:

The regression equation is

$log(Vol) = -2.36 + 2.20 \log(Diam)$

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
<th>F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SSR($X_1$) = 7.93</td>
<td>MSR($X_1$) = 7.93</td>
<td>605</td>
</tr>
<tr>
<td>Error</td>
<td>29</td>
<td>SSE($X_1$) = 0.38</td>
<td>MSE($X_1$) = 0.0131</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>8.31</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Since the F value is again highly significant there is very strong evidence that $\beta_1$ is not equal to zero and the first question is answered.

The second question is answered by looking at the additional SS that can be attributed to fitting $X_1$ after fitting $X_2$. That is
SSR(X₁|X₂) = SSR(X₁, X₂) − SSR(X₂).

To do this, the regression of log(Vol) against log(Height) is fitted first giving the regression equation,

\[
\log(\text{Vol}) = -13.9 + 3.98 \log(\text{Height}).
\]

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
<th>F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SSR(X₂) = 3.50</td>
<td>MSR(X₂) = 3.50</td>
<td>21</td>
</tr>
<tr>
<td>Error</td>
<td>29</td>
<td>SSE(X₂) = 4.81</td>
<td>MSE(X₂) = 0.1659</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>8.3109</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(which also indicates very strongly that \( \beta_2 \) is not zero).

Using these results we then get

\[
SSR(X₁|X₂) = SSR(X₁, X₂) − SSR(X₂) = 8.11 − 3.50 = 4.61
\]

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
<th>F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR(X₂)</td>
<td>1</td>
<td>3.50</td>
<td>3.50</td>
<td></td>
</tr>
<tr>
<td>SSR(X₁</td>
<td>X₂)</td>
<td>1</td>
<td>4.61</td>
<td>4.61</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>0.1922</td>
<td>0.0069</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>8.31</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The additional contribution of \( X₁ \) after fitting \( X₂ \) is highly significant. Hence the test of the partial regression coefficient indicates that \( \beta_1 \) is not zero given that \( \beta_2 \) has already been included in the model.

This test indicates that it is not sufficient to just include \( X₂ \) in the model. However it does not tell us if \( X₂ \) is required if \( X₁ \) is fitted first.

**Exercise** Show that (i) \( \beta_2 \) is not zero and that the additional information supplied by including \( X₂ \) in the model having already included \( X₁ \) is significant. Hence deduce that the best model includes both \( X₁ \) and \( X₂ \).

### 2.4.2 Coefficient of Multiple Determination

The coefficient of multiple determination is defined as

\[
R^2 = \frac{\text{Regression SS}}{SS_y} = 1 - \frac{SSE}{SS_y}.
\]  

(2.12)

When \( p = 1 \) this is just the coefficient of determination, \( r^2 \), defined by (1.4) and as for \( r^2 \)

\[
0 \leq R^2 \leq 1.
\]  

(2.13)
1. $R^2$ measures the proportion of the total variation in $Y$ that can be explained by fitting $X_1, X_2, \ldots, X_p$.

2. $R^2$ should not be used as a test of how well the model fits.

3. Adding independent variables to the model can only increase $R^2$ as adding more variables must increase the regression SS. A modification of $R^2$ is sometimes used that takes into account the number of variables in the model. The adjusted coefficient of multiple determination is defined by

$$R_a^2 = 1 - \left( \frac{n - 1}{n - (p + 1)} \right) \frac{SSE}{SS_y}. \quad (2.14)$$

$R_a^2$ may decrease when an additional independent variable is added to the model.

4. $R = \sqrt{R^2}$ is called the coefficient of multiple correlation and is the correlation between the responses $Y_i$ and the fitted values $\hat{Y}_i$.

### 2.5 Weighted Least Squares

To this point all observations have received equal weight in estimating the parameters of the linear model. This is not always desirable. For example, the variance of the observations may not be independent of $X$ so that instead of $\epsilon$ having a variance-covariance matrix of the form $\sigma^2 I$ it will be of the form $\sigma^2 W$ where $W$ is a diagonal matrix where not all diagonal elements are equal. That is

$$W = \begin{bmatrix} w_1 & & \\ & w_2 & \\ & & \ddots \\ & & & w_n \end{bmatrix} \quad (2.15)$$

In a more general situation the off diagonal elements may also be non-zero in which case the observations would also be correlated.

The linear model can still be written as

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

where now it is assumed that although $\epsilon$ is still normally distributed with mean vector, $0$, it has variance-covariance matrix given by $\sigma^2 W$. The matrix $W$ is symmetric and non-singular.

The ordinary least squares estimate of $\beta$ is no longer appropriate. However, it can be shown that it is possible to find a non-singular symmetric matrix, $P$, such that

$$P'P = PP' = P^2 = W. \quad (2.16)$$
Now define
\[ Z = P^{-1}Y \]  
so that
\[ Z = P^{-1}X\beta + P^{-1}\epsilon. \]  

**Exercise:** Verify that \( P^{-1}\epsilon \sim N(0, \sigma^2 I) \)

The residual sum of squares can now be written as
\[
\epsilon'P^{-1}P^{-1}\epsilon = \epsilon'W^{-1}\epsilon = (Y - X\beta)W^{-1}(Y - X\beta).
\]

The residual sum of squares can now be minimized by differentiating with respect to \( \beta_0, \beta_1, \ldots, \beta_p \) and equating the resulting \( p \) equations to zero to form the normal equations. However since (2.18) satisfies the usual assumptions it is easier to modify the normal equations, (2.6), by replacing \( X \) by \( P^{-1}X \) to get
\[
(X'P^{-1}P^{-1}X)^\hat{\beta} = X'P^{-1}P^{-1}Y
\]
from which
\[
(X'W^{-1}X)\hat{\beta} = X'W^{-1}Y.
\]  

Thus
\[
\hat{\beta} = (X'W^{-1}X)^{-1}X'W^{-1}Y.
\]

It follows that the regression sum of squares is given by
\[
\hat{\beta}'X'W^{-1}Y = Y'W^{-1}X(X'W^{-1}X)^{-1}X'W^{-1}Y
\]  
and the total sum of squares by
\[
Y'W^{-1}Y.
\]  

The difference between equations (2.23) and (2.22) then gives the residual sum of squares.

Weighted least squares are important not only in the linear model sense discussed in this section, but are also important in the *generalised linear model* which will be discussed later in the course.
Chapter 3

Model Building

3.1 Introduction

The problem to be considered in this chapter is how to choose an appropriate model. It will be assumed here that a linear regression model is appropriate, although in a more general context alternate models would also need to be considered. Thus, given a dependent variable \( Y \) and a set of possible predictor (independent) variables, \( X_1, X_2, \ldots, X_{p-1} \), how can the set of independent variables to be used in the actual regression model be selected?

Given a model of the form

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{i,p-1} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \quad (3.1)
\]

the problem of testing subsets of the parameters was examined in chapter 2, section (2.4). In section 2.4.2 the hypothesis, \( \beta_i = 0 \), was tested to see if the independent variable \( X_i \) provided useful additional information given that other variables had already been fitted. This latter test was then used to select a model when there were only two predictor variables. While a similar method could in theory be used when any number of predictor variables are to be considered the process would be very cumbersome and would not be practical. Two methods of subset selection, stepwise regression and all subsets regression will be introduced.

3.2 Stepwise Regression

Stepwise regression is an automatic search procedure. The procedure begins by selecting an initial \( X \) variable to include in the model and then at each step another \( X \) is added to the model, or an \( X \) that has already been included at a preceding step may be deleted according to some well defined criterion. The procedure continues until no further variables can be added or deleted according to the criteria adopted.
3.2.1 Search Procedure

Note: NWT uses the partial $F$ test as the criterion to add or delete variables and this method will be adopted here also.

Given a dependent variable $Y$ and $p - 1$ independent variables $X_1, X_2, \ldots, X_{p-1};$

1. Fit the $p - 1$ simple regressions, $Y$ against $X_k,$ $k = 1, 2, \ldots, p - 1.$ Select the $X_k$ which has the largest value of

\[
F_k^* = \frac{MSR(X_k)}{MSE(X_k)},
\]

provided $F_k^*$ exceeds a predetermined value, called the $F$ to enter. If all $F_k^*$ are less than this value the process terminates as none of the $X$ variables are then considered as being useful for inclusion in the model. This completes the first step.

By this criterion the selected $X_k$ will be the $X$ variable which is most highly correlated with the dependent variable $Y.$ (Why?)

2. For simplicity, assume $X_1$ is the variable entered at step 1. The stepwise procedure now fits all models of the form

\[
Y = \beta_0 + \beta_1 X_1 + \beta_k X_k + \epsilon
\]

$(k = 2, 3, \ldots, p - 1)$ and selects the variable $X_k$ for inclusion with the largest value of the partial $F$ statistic

\[
F_k^* = \frac{MSR(X_k|X_1)}{MSE(X_1, X_k)}
\]

provided $F_k^*$ exceeds the $F$ to enter.

3. Again, for simplicity, assume $X_2$ is added at the second step. The procedure continues by checking that $X_1$ is really required given that $X_2$ is now in the model by calculating

\[
F_1^* = \frac{MSR(X_1|X_2)}{MSE(X_1, X_2)}
\]

If $F^*$ falls below a predetermined value called the $F$ to remove, then $X_1$ is dropped from the model. This step completes the second stage.

4. In general, each stage consists of two steps. The first step is to decide which one of the $X_k$ of those not already in the model should be included. The second step is to consider those $X$’s already in the model and determine which one, if any, should now be deleted given that an additional variable has been included.

Thus, suppose at the completion of any stage the process has not terminated and that the current model includes $r$ independent variables. For simplicity assume these are $X_1, X_2, \ldots, X_r$ (although in general this could be any subset of $r$ variables from
the $p - 1$ independent variables). Then $X_{r+1}, X_{r+2}, \ldots, X_{p-1}$ are assumed to be the variables that have not already been included in the model. Each of these latter variables are examined for possible inclusion by calculating the partial $F$ statistic

$$F_k^* = \frac{MSR(X_k | X_1, X_2, \ldots, X_r)}{MSE(X_1, X_2, \ldots, X_r, X_k)}$$

for $k = r+1, r+2, \ldots, X_{p-1}$. The $X_k$ with the largest value of $F^*$ is included provided it exceeds the $F$ to enter specified. The procedure terminates if no variable is added.

Assuming that $X_{r+1}$ is the new variable added, then the partial $F$ statistics

$$F_k^* = \frac{MSR(X_k | X_1, X_2, \ldots, X_{k-1}, X_{k+1}, \ldots, X_r, X_{r+1})}{MSE(X_1, X_2, \ldots, X_{k-1}, X_{k}, X_{k+1}, \ldots, X_r, X_{r+1})}$$

for $k = 1, 2, \ldots, r$ are calculated and the $X_k$ with the smallest value of $F^*$ is omitted provided it does not exceed the $F$ to remove value specified.

Notice that variables added can be dropped at a later stage when considered in conjunction with the new variables included at later stages. The procedure stops only when the procedure determines no further variables can be added.

**F to Enter**

The predetermined level that $F^*$ must exceed for $X$ to be included in the model, called the $F$ to enter, must be given a value. This is often set at 4 and this will be the value assumed in the examples that follow. This level corresponds roughly to a significance level of 0.05 for a single test based on about 50 degrees of freedom.

**F to Remove**

The predetermined value for $F^*$ used for deciding if an $X$ variable in the model should be excluded is also often set at a value of 4 and this value will also be used in the examples.

Since the number of calculations required for stepwise regression are often large, a computer and appropriate software are usually required to carry out the procedure. In the following example the steps required are followed through in some detail. A statistical package, SPlus was used to do the calculations.

**Example 3.2.1: Stepwise Regression**

In the following table the need for labour (lhours), in 17 U.S Naval Hospitals is related to average daily patient load (pload), monthly x-ray exposures (x-ray), monthly occupied bed days (beddays), eligible population/1000 (pop), and average stay in days (lstay).

<table>
<thead>
<tr>
<th>Hospital</th>
<th>pload</th>
<th>x-rays</th>
<th>bed-days</th>
<th>pop</th>
<th>lstay</th>
<th>l-hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.57</td>
<td>2463</td>
<td>472.9</td>
<td>18.0</td>
<td>4.45</td>
<td>566.5</td>
</tr>
<tr>
<td>2</td>
<td>44.02</td>
<td>2048</td>
<td>1339.7</td>
<td>9.5</td>
<td>6.92</td>
<td>696.8</td>
</tr>
<tr>
<td>3</td>
<td>20.42</td>
<td>3940</td>
<td>620.2</td>
<td>12.8</td>
<td>4.28</td>
<td>1033.2</td>
</tr>
<tr>
<td>4</td>
<td>18.74</td>
<td>6505</td>
<td>568.3</td>
<td>36.7</td>
<td>3.90</td>
<td>1603.6</td>
</tr>
<tr>
<td>5</td>
<td>49.20</td>
<td>5723</td>
<td>1497.6</td>
<td>35.7</td>
<td>5.50</td>
<td>1611.4</td>
</tr>
<tr>
<td>6</td>
<td>44.92</td>
<td>11520</td>
<td>1365.8</td>
<td>24.0</td>
<td>4.60</td>
<td>1613.3</td>
</tr>
<tr>
<td>7</td>
<td>55.48</td>
<td>5779</td>
<td>1687.0</td>
<td>43.3</td>
<td>5.62</td>
<td>1854.2</td>
</tr>
<tr>
<td>8</td>
<td>59.28</td>
<td>5969</td>
<td>1639.9</td>
<td>46.7</td>
<td>5.15</td>
<td>2160.6</td>
</tr>
<tr>
<td>9</td>
<td>94.39</td>
<td>8461</td>
<td>2872.3</td>
<td>78.7</td>
<td>6.18</td>
<td>2305.6</td>
</tr>
<tr>
<td>10</td>
<td>128.02</td>
<td>20106</td>
<td>3655.1</td>
<td>180.5</td>
<td>6.15</td>
<td>3505.9</td>
</tr>
<tr>
<td>11</td>
<td>96.00</td>
<td>13313</td>
<td>2912.0</td>
<td>60.9</td>
<td>5.88</td>
<td>3571.9</td>
</tr>
<tr>
<td>12</td>
<td>131.42</td>
<td>10771</td>
<td>3921.0</td>
<td>103.7</td>
<td>4.88</td>
<td>3741.4</td>
</tr>
<tr>
<td>13</td>
<td>127.21</td>
<td>15543</td>
<td>3865.7</td>
<td>126.8</td>
<td>5.50</td>
<td>4026.5</td>
</tr>
<tr>
<td>14</td>
<td>252.90</td>
<td>36194</td>
<td>7684.1</td>
<td>157.7</td>
<td>7.00</td>
<td>10343.8</td>
</tr>
<tr>
<td>15</td>
<td>409.20</td>
<td>34703</td>
<td>12446.3</td>
<td>169.4</td>
<td>10.78</td>
<td>11732.2</td>
</tr>
<tr>
<td>16</td>
<td>463.70</td>
<td>39204</td>
<td>14098.4</td>
<td>331.4</td>
<td>7.05</td>
<td>15414.9</td>
</tr>
<tr>
<td>17</td>
<td>510.22</td>
<td>86533</td>
<td>15524.0</td>
<td>371.6</td>
<td>6.35</td>
<td>18854.4</td>
</tr>
</tbody>
</table>

(Note that hospital number is only used for identification and is not used as an X variable.)

**Stepwise Procedure**

1. The following table gives the correlation coefficient between the dependent variable labour, (l-hours), and the five independent variables as well as the F-statistic from the simple linear regressions.

<table>
<thead>
<tr>
<th>X Variable</th>
<th>Correlation</th>
<th>F-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. pload</td>
<td>0.98565</td>
<td>511.41</td>
</tr>
<tr>
<td>2. x-rays</td>
<td>0.94518</td>
<td>125.67</td>
</tr>
<tr>
<td>3. bedday</td>
<td>0.98599</td>
<td>524.30</td>
</tr>
<tr>
<td>4. pop</td>
<td>0.94038</td>
<td>114.65</td>
</tr>
<tr>
<td>5. lstay</td>
<td>0.57859</td>
<td>7.55</td>
</tr>
</tbody>
</table>

The largest F to enter is obtained for the monthly occupied bed days, (bedday), and is greater than 4. This is the first variable added to the model.

2. The next table gives the F* statistic (F to Enter) for the variables not in the model.

<table>
<thead>
<tr>
<th>X Variable</th>
<th>F To Enter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. pload</td>
<td>0.17</td>
</tr>
<tr>
<td>2. x-rays</td>
<td>15.33</td>
</tr>
<tr>
<td>3. pop</td>
<td>1.81</td>
</tr>
<tr>
<td>4. lstay</td>
<td>11.55</td>
</tr>
</tbody>
</table>
Xrays is the next variable to enter the equation as it has the largest value of $F^*$ and $F^* > 4$.

3. The partial $F^*$-statistic ($F$ to Remove) for the $X$ variable (bedday) already in the model gives a value of 98.4 so that bedday remains in the model.

4. Now,

<table>
<thead>
<tr>
<th>X VARIABLE</th>
<th>$F$ to Enter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. pload</td>
<td>0.62</td>
</tr>
<tr>
<td>4. pop</td>
<td>0.04</td>
</tr>
<tr>
<td>5. lstay</td>
<td>4.39</td>
</tr>
</tbody>
</table>

and length of stay (lstay) is included in the model.

5. Using

<table>
<thead>
<tr>
<th>X Variable</th>
<th>$F$ to Remove</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. xrays</td>
<td>6.96</td>
</tr>
<tr>
<td>4. bedday</td>
<td>86.62</td>
</tr>
</tbody>
</table>

neither bedday or xrays is removed.

6. Continuing,

<table>
<thead>
<tr>
<th>X Variable</th>
<th>$F$ to Enter</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. pload</td>
<td>0.60</td>
</tr>
<tr>
<td>5. pop</td>
<td>0.97</td>
</tr>
</tbody>
</table>

no more variables are now included since the $F$ to enter for all remaining $X$’s is less than 4. The process is now terminated and selects the model containing the three variables, bedday, xrays and lstay.
Example 3.2.2 - R for Stepwise Regression
The R functions add1() and drop1() calculate the F statistic for adding and dropping.

The iterative use of them requires deft programming.

```r
#_____ hospitals.R __________
options(digits=2)
hospitals <- read.table("hospitals.txt",header=T)
pairs(hospitals)

Xnames <- names(hospitals)[1:5]  # names of the x variables
form1 <- formula(1hours ~ 1)  # start with a model which only fits Ybar
model1 <- lm(form1,data=hospitals)

stop <- 0  # a dummy variable to regulate the while loop
m <- 0  # a counter for the while loop
while(stop==0){
m <- m+1
cat("iteration ",m,"\n")
k <- length(Xnames)  # k is the number of X variables available
#
# the formula for adding
form.add <- formula(paste("~ ",paste(Xnames,collapse="+")," + . ") )
#
# calculate and extract the F statistics for adding
add.model <- add1(object=model1,scope=form.add,test="F")
Fincl <- array(unlist(add.model["F value"])[-1],c(1,k,dimnames=list(NULL,Xnames))
cat("F incl \n"); print(Fincl)

if(max(Fincl) >=4){  # only include if F > 4
max.position <- order(Fincl,decreasing=T)[1]
include.X <- Xnames[max.position]  # choose the X variable with Fmax
Xnames <- Xnames[-max.position]  # remaining X variables
#
# update the model with include.X
upd.form <- formula(paste("- ",include.X,include="+"))
model1 <- update(model1,upd.form)

current.X <- attributes(model1$terms)$term.labels
form.drop <- formula(paste(" - ",paste(current.X,collapse="+"))
drop.model <- drop1(model1,form.drop,test="F")
Fdrop <- array(unlist(drop.model["F value"])[-1],c(1,length(current.X)),
               dimnames=list(NULL,current.X))
cat("F remove \n"); print(Fdrop)
}  # end of if()
else stop <- 1  # Fmax < 4
if (k==1) stop <- 1  # end of candidate X variables
}
# end of while()

print(anova(model1))
```

41
The output from this job shows the F to enter and F to remove at each iteration and the final model:-

iteration 1
F incl
pload xrays beddays pop lstay
[1,] 511 126 524 115 7.5
F remove
beddays
[1,] 524
iteration 2
F incl
pload xrays pop lstay
[1,] 0.17 15 1.8 12
F remove
beddays xrays
[1,] 98 15
iteration 3
F incl
pload pop lstay
[1,] 0.62 0.035 4.4
F remove
beddays xrays lstay
[1,] 87 7 4.4
iteration 4
F incl
pload pop
[1,] 0.6 0.97

Response: lhours

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>beddays</td>
<td>1</td>
<td>4.81e+08</td>
<td>4.81e+08</td>
<td>1273.19</td>
<td>2.3e-14 ***</td>
</tr>
<tr>
<td>xrays</td>
<td>1</td>
<td>7.19e+06</td>
<td>7.19e+06</td>
<td>19.04</td>
<td>0.00077 ***</td>
</tr>
<tr>
<td>lstay</td>
<td>1</td>
<td>1.66e+06</td>
<td>1.66e+06</td>
<td>4.39</td>
<td>0.05633 .</td>
</tr>
<tr>
<td>Residuals</td>
<td>13</td>
<td>4.91e+06</td>
<td>3.78e+05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that at the beginning the variable pload has a large F to enter but once beddays enters the model, pload is not required. These variables are highly correlated. The test for removal does not indicate that any variables should be deleted once they have been included.

3.2.2 Forward Selection

The procedure selects variables to enter the model in a similar fashion to stepwise regression. The major difference is that once a variable enters the model there is no mechanism for removing it, so once a variable is selected it must remain in the model with the result
that the final subset selected may be different than in the case of stepwise regression. The procedure terminates when no more variables can be entered into the model.

**Example 3.2.3-Forward Selection** This is the same as the previous except that there is no test for removal.

Stepwise regression examines all terms in the current model as potential candidates to be dropped and all terms not in the model as potential candidates to add. Only one term is added or dropped at each step since for correlated terms this can greatly alter the status of all other terms. Our aim is to drop terms from the model whose removal has a negligible effect on the systematic SS given the change in the degrees of freedom and to add terms that significantly increase the systematic SS. The process is automated in R with the function `step()`.

The statistic used by R for dropping or adding terms is either

(i) Akaike’s Information Criteria (AIC) or

(ii) Bayesian Information Criteria (BIC)

both having the general form:-

$$ [A \text{ or } B]IC = -2 \times (\text{maximised } \log \text{ Likelihood}) + k \times p \quad (3.7) $$

where $p$ is the number of parameters in the fitted model.

When $k = 2$, equation (3.7) is the AIC and if $k = \log(n)$ it is the BIC.

It is desirable to make the $-2\ell_{\text{max}}$ as small as possible but with the smallest set of parameters that is necessary. Overfitting is to be avoided. The AIC (or BIC) penalises the error SS by the number parameter. Too few parameters and the ESS component is too large and conversely overparameterisation reduces the ESS but at the expense of inflating the penalty ($kp$).

For large samples sizes, ($n > 100$ say), the BIC is preferred in order to avert Type I errors.

Usually $\sigma^2$ has to be estimated and for the linear model

$$ AIC = n \log(ESS/n) + kp $$

**Mallows $C_p$**

An alternate (though related) statistic is Mallows’ $C_p$,

$$ C_p = X^2 + k \times p \sigma^2 \quad (3.8) $$

Since $C_p$ is moment based it is more easily computed as a quadratic function and is preferred for speed computing.

The algorithm is implemented in the following manner

1. Fit an initial model.
2. Fit all possible models that can be obtained by dropping (or adding) a term to the current model and compute the AIC for each.

3. Add (or remove) the term which reduces the AIC the most to form a new model. **Small AIC is good.**

4. Repeat the steps above until the AIC cannot be reduced any further at which time the process terminates.

**Backward Elimination**

This is similar to stepwise regression except that the saturated model is chosen as the initial model. Terms are dropped as for the stepwise procedure using the AIC criteria. The big difference is that once a term is dropped it is not allowed to re-enter the model. (Terms cannot be added to the model).

**Forward Selection**

Again this has similarities with stepwise regression. However, we begin with the minimal model we want to consider (the model containing the terms that for whatever reason we deem MUST be included in the model). Terms are added as in the stepwise procedure but once included in the model cannot be dropped.

It should be noted that these three procedures will often produce different results. The final result in each case can also depend on the choice of initial model.

There is no guarantee that these procedures will identify the ‘best’ model or even a sensible model. It is essential to examine how well the final model fits the data using the usual diagnostic procedures.

The model selection for previous example of the hospitals data using the AIC is done in R by

```r
formL <- formula(~ 1)
formU <- formula(~ pload + xrays + beddays + pop + lstay)
step.model <- step(model0, direction="forward", scope=list(lower=formL, upper=formU))
```

**Start:**

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1.38e+07</td>
<td>4.81e+07</td>
<td>2.35e+02</td>
</tr>
<tr>
<td>+ beddays</td>
<td>1.41e+07</td>
<td>4.81e+07</td>
<td>2.36e+02</td>
</tr>
<tr>
<td>+ pload</td>
<td>5.28e+07</td>
<td>4.42e+07</td>
<td>2.58e+02</td>
</tr>
<tr>
<td>+ xrays</td>
<td>5.72e+07</td>
<td>4.37e+07</td>
<td>2.60e+02</td>
</tr>
<tr>
<td>+ pop</td>
<td>3.29e+08</td>
<td>1.66e+08</td>
<td>2.89e+02</td>
</tr>
<tr>
<td>+ lstay</td>
<td>3.95e+08</td>
<td>1.66e+08</td>
<td>3.00e+02</td>
</tr>
</tbody>
</table>

**Step:**

AIC = 235

44
lhours ~ beddays

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ xrays</td>
<td>1</td>
<td>7191211</td>
<td>6568569</td>
</tr>
<tr>
<td>+ lstay</td>
<td>1</td>
<td>6218991</td>
<td>7540789</td>
</tr>
<tr>
<td>+ pop</td>
<td>1</td>
<td>1574286</td>
<td>12185494</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td>1</td>
<td>13759780</td>
<td>13598065</td>
</tr>
<tr>
<td>+ pload</td>
<td>1</td>
<td>161715</td>
<td>13598065</td>
</tr>
</tbody>
</table>

Step: AIC= 225
lhours ~ beddays + xrays

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ lstay</td>
<td>1</td>
<td>1657870</td>
<td>4910699</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td>1</td>
<td>6568569</td>
<td>225</td>
</tr>
<tr>
<td>+ pload</td>
<td>1</td>
<td>301141</td>
<td>6267428</td>
</tr>
<tr>
<td>+ pop</td>
<td>1</td>
<td>17828</td>
<td>6550741</td>
</tr>
</tbody>
</table>

Step: AIC= 222
lhours ~ beddays + xrays + lstay

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>1</td>
<td>4910699</td>
<td>222</td>
</tr>
<tr>
<td>+ pload</td>
<td>1</td>
<td>365703</td>
<td>4544996</td>
</tr>
<tr>
<td>+ pop</td>
<td>1</td>
<td>234493</td>
<td>4676205</td>
</tr>
</tbody>
</table>

print(anova(step.model))

Analysis of Variance Table

Response: lhours

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>beddays</td>
<td>1</td>
<td>4.81e+08</td>
<td>4.81e+08</td>
<td>1273.19</td>
</tr>
<tr>
<td>xrays</td>
<td>1</td>
<td>7.19e+06</td>
<td>7.19e+06</td>
<td>19.04</td>
</tr>
<tr>
<td>lstay</td>
<td>1</td>
<td>1.66e+06</td>
<td>1.66e+06</td>
<td>4.39</td>
</tr>
<tr>
<td>Residuals</td>
<td>13</td>
<td>4.91e+06</td>
<td>3.78e+05</td>
<td></td>
</tr>
</tbody>
</table>

At the start, the NULL model has AIC = 294. Of all the candidate variables, beddays improves (i.e. reduces) the AIC most so it is the first chosen predictor. At the next step xrays is best and pload actually increases the AIC indicating that including that variable in the model would not improve the fit by more than the 2 degrees of freedom penalty. At the last step lstay is selected.

Obviously, this is more convenient than the previous.

An argument for step() is to put trace=0 and this suppresses the output. For a large
number of candidate predictors, this option should be used because the output will be copious and with that comes a corresponding slowing of the program.

### 3.3 Linear Models and Experimental Design

The examples of linear models considered so far have dealt with *quantitative* variables, that is variables that take numerical values such as temperature, weight, age, income, height, number of seeds or proportion of males to females.

If a variable is not quantitative it is *qualitative*. For example, sex is a qualitative variable having two categories (male, female) rather than being measured on a numerical scale. Other examples include fertilizer status (fertilized, unfertilized), contamination level (none, low, medium, high) or colour (red, blue, green, black, white). Not only quantitative but also qualitative variables can be used in regression analysis. Qualitative variables are used by first defining (quantitative) *indicator* (or *dummy*) variables that identify the levels (or categories) of the qualitative variable. It is these dummy variables (sometimes in conjunction with quantitative variables) that are used to build linear models for experimental designs.

The connection between experimental design and the linear model will now be illustrated by means of examples.

**Example 3.4.1**

Consider an experiment where there are three treatment groups. Treatment can be considered as an independent qualitative variable having three levels (treatment 1, treatment 2, treatment 3). Three indicator variables $X_1$, $X_2$ and $X_3$ could then be defined as

\[
X_1 = \begin{cases} 
1 & \text{if observation from treatment 1} \\
0 & \text{otherwise.}
\end{cases}
\]

\[
X_2 = \begin{cases} 
1 & \text{if from treatment 2} \\
0 & \text{otherwise}
\end{cases}
\]

\[
X_3 = \begin{cases} 
1 & \text{if from treatment 3} \\
0 & \text{otherwise}
\end{cases}
\]

The regression model can then be written as

\[
y_{ij} = \mu_1 X_1 + \mu_2 X_2 + \mu_3 X_3 + \epsilon_{ij}
\]

If the first two observations come from treatment 1, the next three from treatment 2 and the last two from treatment 3 then the $X$ matrix would be

\[
\begin{pmatrix}
X_1 & X_2 & X_3
\end{pmatrix}
\]
\[
\begin{bmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
\end{bmatrix}
\]

This example is just a special case of one-way analysis of variance (see Stats 261 notes).

### 3.4 One-Way Analysis of Variance

Suppose there are \( k \) independent samples,

\[
y_{11}, y_{12}, \ldots, y_{1n_1} \\
y_{21}, y_{22}, \ldots, y_{2n_2} \\
\vdots \\
y_{k1}, y_{k2}, \ldots, y_{kn_k}
\]
drawn from \( k \) normal populations \( N(\mu_1, \sigma^2) \), \( N(\mu_2, \sigma^2) \), \( \ldots \), \( N(\mu_k, \sigma^2) \) respectively. The model to be considered is

\[
y_{ij} = \mu_i + \epsilon_{ij}
\]

for \( i = 1, 2, \ldots, k \) and \( j = 1, 2, \ldots, n_i \). Equation (3.16) can then be rewritten as

\[
y_{ij} = \mu_1 X_1 + \mu_2 X_2 + \ldots + \mu_k X_k + \epsilon_{ij}
\]

where the indicator variable \( X_r \) is defined as

\[
X_r = \begin{cases} 
1 & \text{all } y_{ij}, \ i = r, \ j = 1, 2, \ldots, n_r \\
0 & \text{all other } y_{ij}
\end{cases}
\]

That is \( X_r = 1 \) if \( y_{ij} \) is an observation from the \( r \)th group. Thus the \( X \) matrix and vector of parameters \( \beta \) are given by

\[
X = \begin{bmatrix}
1_{(n_1 \times 1)} & 0 & 0 \\
0 & 1_{(n_2 \times 1)} & 0 \\
\vdots & \ddots & \ddots \\
0 & 0 & 1_{(n_k \times 1)}
\end{bmatrix} \quad (N \times k) \\
\beta = \begin{bmatrix}
\mu_1 \\
\mu_2 \\
\vdots \\
\mu_k
\end{bmatrix}
\]

noting that \( 1_{(n_i \times 1)} \) is a \( (n_i \times 1) \) column vector of 1’s, making \( X \) a \( (N \times k) \) matrix. Then,

\[
X'X = \begin{bmatrix}
n_1 & 0 & 0 \\
0 & n_2 & 0 \\
\vdots & \ddots & \ddots \\
0 & 0 & n_k
\end{bmatrix} \quad X'y = \begin{bmatrix}
\sum y_{1i} \\
\sum y_{2i} \\
\vdots \\
\sum y_{ki}
\end{bmatrix}
\]

47
so that

\[
(X'X)^{-1} = \begin{bmatrix}
    n_1^{-1} & 0 & 0 \\
    0 & n_2^{-1} & 0 \\
    0 & 0 & \ldots
\end{bmatrix}
\]

Using results 2.1 and 2.6 of (2.1.2),

\[
\hat{\beta} = (X'X)^{-1}X'y = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)' = (\overline{y}_1, \overline{y}_2, \ldots \overline{y}_k)'
\]

and the regression SS due to fitting \(\mu_1, \mu_2, \ldots, \mu_k\) is given by

\[
SSR(\mu_1, \mu_2, \ldots, \mu_k) = \hat{\beta}' X' y = n_1\overline{y}_1^2 + n_2\overline{y}_2^2 + \ldots + n_k\overline{y}_k^2.
\]

Thus

\[
\text{Residual SS} = \text{Total SS} - \text{Regression SS} = y'y - (n_1\overline{y}_1^2 + n_2\overline{y}_2^2 + \ldots + n_k\overline{y}_k^2). \tag{3.15}
\]

The hypothesis of interest is that all the means are equal or:

\[
H: \mu_1 = \mu_2 = \ldots = \mu_k. \tag{3.16}
\]

If the hypothesis is assumed correct then model (3.16) becomes

\[
y_{ij} = \mu X + \epsilon_{ij} \tag{3.17}
\]

where \(X = 1\) for all \(y_{ij}, i = 1, 2, \ldots, k, j = 1, 2, \ldots, n_i\). Thus

\[
X'X = N (= \sum n_{ij}),
\]

so that

\[
\hat{\mu} = (X'X)^{-1}X'y = (\sum y_{ij})/N = \overline{y} \tag{3.18}
\]

and the regression SS due to fitting \(\mu\) is given by

\[
SSR(\mu) = \hat{\mu}' X'y = \overline{y} \sum y_{ij} = N\overline{y}^2. \tag{3.19}
\]

The residual SS is given by \(y'y - N\overline{y}^2\).

The reduction in the residual SS due to fitting \(k\) different means instead of the one common mean is then

\[
SSR(\mu_1, \mu_2, \ldots, \mu_k) - SSR(\mu) = (n_1\overline{y}_1^2 + n_2\overline{y}_2^2 + \ldots + n_k\overline{y}_k^2) - N\overline{y}^2. \tag{3.20}
\]
and is the SS that was called the **Between Treatments SS** in Stats 261. The total SS can now be written as the sum of three quadratic forms. That is

\[
y' I y = \text{Regression SS + Residual SS} \\
= N\bar{y}^2 + \{ (n_1\bar{y}_1^2 + n_2\bar{y}_2^2 + \ldots + n_k\bar{y}_k^2) - N\bar{y}^2 \} + (y' y - n_1\bar{y}_1^2 + n_2\bar{y}_2^2 + \ldots + n_k\bar{y}_k^2) \\
= \text{Mean SS + Between Treatments SS + Residual SS} \\
= y' A_1 y + y' A_2 y + y' A_3 y
\]

where the \((N \times N)\) matrices associated with the quadratic forms are given by:

\[
A_1 = \begin{bmatrix}
\frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\
\frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N}
\end{bmatrix}
\]

(3.21)

\[
A_2 = \begin{bmatrix}
B_1 & M_{12} & \cdots & M_{1k} \\
M_{21} & B_2 & \cdots & M_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
M_{k1} & M_{k2} & \cdots & B_k
\end{bmatrix}
\]

(3.22)

where \(B_i\) is a \((n_i \times n_i)\) matrix with all elements equal to \(\frac{1}{n_i} - \frac{1}{N}\). That is:

\[
B_i = \begin{bmatrix}
\frac{1}{n_i} - \frac{1}{N} & \frac{1}{n_i} - \frac{1}{N} \\
\frac{1}{n_i} - \frac{1}{N} & \frac{1}{n_i} - \frac{1}{N} \\
\frac{1}{n_i} - \frac{1}{N} & \frac{1}{n_i} - \frac{1}{N} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{1}{n_i} - \frac{1}{N} & \frac{1}{n_i} - \frac{1}{N}
\end{bmatrix}
\]

(3.23)

and \(M_{ij}\) is a \((n_i \times n_j)\) matrix with all elements equal to \(-N^{-1}\).

Finally

\[
A_3 = \begin{bmatrix}
C_1 & 0 & \cdots & 0 \\
0 & C_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & C_k
\end{bmatrix}
\]

(3.24)

where

\(C_k\) is a \((n_k \times n_k)\) matrix given by
\[
C_i = \begin{bmatrix}
(1 - \frac{1}{n_i}) & -\frac{1}{n_i} & \cdots & -\frac{1}{n_i} \\
-\frac{1}{n_i} & (1 - \frac{1}{n_i}) & \cdots & -\frac{1}{n_i} \\
\vdots & \vdots & \ddots & \vdots \\
-\frac{1}{n_i} & \frac{1}{n_i} & \cdots & (1 - \frac{1}{n_i})
\end{bmatrix}
\]  

(3.25)

The results can be summarized in tabular form in an Analysis of Variance Table:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Quadratic Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due to Grand Mean</td>
<td>1</td>
<td>(N\overline{y}^2)</td>
<td>(y'y)</td>
</tr>
<tr>
<td>Between Treatments</td>
<td>(k - 1)</td>
<td>(n_1\overline{y}_1^2 + n_2\overline{y}_2^2 + \cdots + n_k\overline{y}_k^2 - N\overline{y}^2)</td>
<td>(y'A_1y)</td>
</tr>
<tr>
<td>Residual</td>
<td>(N-k)</td>
<td>(y'y - (n_1\overline{y}_1^2 + n_2\overline{y}_2^2 + \cdots + n_k\overline{y}_k^2))</td>
<td>(y'A_2y)</td>
</tr>
<tr>
<td>Total</td>
<td>(N)</td>
<td>(y'y)</td>
<td>(y'A_3y)</td>
</tr>
</tbody>
</table>

3.4.1 Using Cochran’s Theorem

Now \(A_1 + A_2 + A_3 = I\) and \(A_iA_j = 0\) for \(i \neq j\). Hence Cochran’s Theorem holds and the three components of the analysis of variance table above are distributed as independent \(\chi^2\) variates. The degrees of freedom are then given by the ranks of the matrices \(A_i\). But, by Cochran’s Theorem, the matrices, \(A_i\), are all idempotent, and \(\text{rank}(A_i) = \text{Trace}(A_i) = \sum_j a_{jj}\).

3.4.2 Testing the Hypothesis

The hypothesis \(H: \mu_1 = \mu_2 = \cdots = \mu_k\) is tested using the ratio

\[
\frac{(\text{Between Treatments SS})/(k-1)}{(\text{Residual SS})/(N-k)}
\]

which is distributed as an \(F(k-1, N-k)\).

**Exercise:** (i) Verify the degrees of freedom associated with each term in the AOV for a oneway design.

(ii) Explain why the test above has an \(F\) distribution.

3.5 Two-Way Analysis of Variance

Suppose in an experiment there are two factors to be examined, factor A having \(a\) levels and factor B having \(b\) levels. Consider the case where there is only one observation per cell. The model can be written as

\[
y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}
\]

(3.27)
for \( i = 1,2, \ldots, a, \) and \( j = 1,2, \ldots, b. \) It is assumed that \( \epsilon_{ij} \sim N(0,\sigma^2). \) The model (3.33) can then be rewritten in the form

\[
y_{ij} = X_0 \mu + X_1 \alpha_1 + \ldots + X_a \alpha_a + Z_1 \beta_1 + \ldots + Z_b \beta_b + \epsilon_{ij} \tag{3.28}
\]

where the indicator variables \( X_i \) and \( Z_j \) are defined by

\[
X_0 = \begin{cases} 
1 & \text{for all } i \text{ and } j, \\
0 & \text{all other } y_{ij} 
\end{cases}
\]

\[
X_k = \begin{cases} 
1 & \text{all } y_{ij}, j = 1,2, \ldots, b; \ i = k \\
0 & \text{all other } y_{ij} 
\end{cases}
\]

\[
Z_i = \begin{cases} 
1 & \text{for all } y_{ij}, i = 1,2, \ldots, a; \ j = l \\
0 & \text{for all other } y_{ij} 
\end{cases}
\]

Equation (3.33) can now be written in the usual form of

\[
y = X \beta + \epsilon
\]

That is

\[
\begin{bmatrix}
y_{11} \\
y_{12} \\
\vdots \\
y_{1b} \\
y_{21} \\
\vdots \\
y_{2b} \\
y_{a1} \\
y_{ab}
\end{bmatrix} = \begin{bmatrix}
1 & 1 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\
1 & 1 & 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \\
1 & 0 & 1 & \ldots & 0 & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 1 & \ldots & 0 & 0 & 0 & \ldots & 1
\end{bmatrix} \begin{bmatrix}
\mu \\
\alpha_1 \\
\vdots \\
\alpha_a \\
\beta_1 \\
\vdots \\
\beta_b 
\end{bmatrix} + \begin{bmatrix}
\epsilon_{11} \\
\epsilon_{12} \\
\vdots \\
\epsilon_{1b} \\
\epsilon_{21} \\
\epsilon_{2b} \\
\vdots \\
\epsilon_{a1} \\
\epsilon_{ab}
\end{bmatrix}
\]

Notice that there are \( a + b + 1 \) parameters to estimate but the rank of \( X \) is not greater than \( a + b - 1 \) since adding columns 2 to \( a+1 \) gives column 1 as does adding columns \( a + 2 \) to \( a + b + 1 \). (The rank is \( a + b - 1 \).) To overcome the problem of \( X \) not having full rank two restrictions must be applied. One possibility is the sum to zero restrictions

\[
\sum \alpha_i = 0 \\
\sum \beta_j = 0
\]

That is,

\[
\alpha_a = -\alpha_1 - \alpha_2 - \cdots - \alpha_{a-1}
\]
and

$$\beta_b = -\beta_1 - \beta_2 - \cdots - \beta_{b-1}$$

Model (3.33) can then be rewritten as

$$y_{ij} = X_0\mu + \alpha_1(X_1 - X_a) + \alpha_2(X_2 - X_a) + \cdots + \alpha_{a-1}(X_{a-1} - X_a) + \beta_1(Z_1 - Z_b) + \cdots + \beta_{b-1}(Z_{b-1} - Z_b) + \epsilon_{ij}$$

so that

$$y_{ij} = X_0\mu + \alpha_1U_1 + \alpha_2U_2 + \cdots + \alpha_{a-1}U_{a-1} + \beta_1V_1 + \beta_2V_2 + \cdots + \beta_{b-1}V_{b-1} + \epsilon_{ij}$$

(3.31)

where $U_i = X_i - X_a$ and $V_j = Z_j - Z_b$. In matrix notation this becomes

$$
\begin{bmatrix}
  y_{11} \\
  y_{12} \\
  \vdots \\
  y_{ib} \\
  y_{21} \\
  \vdots \\
  y_{2b} \\
  \vdots \\
  y_{a1} \\
  y_{ab}
\end{bmatrix} = 
\begin{bmatrix}
  1 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
  1 & 1 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\
  \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  1 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \\
  1 & 0 & 1 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
  \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
  1 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 1 \\
  1 & 0 & 1 & \cdots & 0 & -1 & -1 & \cdots & -1
\end{bmatrix}
\begin{bmatrix}
  \mu \\
  \alpha_1 \\
  \vdots \\
  \alpha_{a-1} \\
  \beta_1 \\
  \vdots \\
  \beta_{b-1}
\end{bmatrix} + 
\begin{bmatrix}
  \epsilon_{11} \\
  \epsilon_{ij} \\
  \vdots \\
  \epsilon_{ab}
\end{bmatrix}
$$

and the matrix $X$ now has full rank. Two hypotheses are of interest.

$$H : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

(3.32)

and

$$H : \beta_1 = \beta_2 = \cdots = \beta_b = 0$$

(3.33)

52
Comment: Equation (3.33) implies that the observation $y_{ij}$ is a combination of a mean effect $\mu$ plus a contribution due to adding level $i$ of factor $A$, a contribution due to adding level $j$ of factor $B$ and a random (unexplainable) contribution $\epsilon_{ij}$. If, for example, there is no contribution from level $i$ of factor $A$ then $\alpha_i = 0$. If factor $A$ has no effect then all $\alpha$'s are zero.

Exercise: What does the design matrix $X$ become if the restrictions $\alpha_1 = 0$ and $\beta_1 = 0$ are used.

Comment: Since $X$ is not of full rank there is no unique solution to the normal equations. The solution will depend on the actual restriction applied. However, it can be shown that estimates of quantities such as $\alpha_i - \alpha_j$ and $\beta_i - \beta_j$ will be uniquely estimated.

### 3.5.1 Testing the Hypotheses

The reduction in the residual SS technique is again utilised to test the hypotheses (3.38) and (3.39) above. First fit the restricted model assuming all $\alpha$'s and $\beta$'s are zero. That is fit

$$y_{ij} = \mu X_0 + \epsilon_{ij}$$

(3.34)

to obtain the SS of regression, SSR(\(\mu\)), due to fitting $\mu$ alone and the corresponding Residual SS (or Error SS), SSE(\(\mu\)).

Now fit the model

$$y_{ij} = \mu + \alpha_1 U_1 + \alpha_2 U_2 + \ldots + \alpha_{a-1} U_{a-1} + \epsilon_{ij}$$

(3.35)

and obtain the regression SS due to fitting the $\alpha$'s and $\mu$, SSR(\(\mu, \alpha_1, \alpha_2, \ldots, \alpha_{a-1}\)) and the corresponding Residual SS, SSE(\(\mu, \alpha_1, \alpha_2, \ldots, \alpha_{a-1}\)). Notice that because of dependence only $a - 1$ parameters need (can) be fitted.

The SS due to differences in mean levels of factor $A$ is then given by the difference

$$SSE(\mu) - SSE(\mu, \alpha_1, \alpha_2, \ldots, \alpha_{a-1})$$

which is just the regression SS due to fitting $U_1, U_2, \ldots, U_{a-1}, SSR(U_1, U_2, \ldots, U_{a-1})$, which has $a - 1$ degrees of freedom. Thus for a two way AOV we can write

$$\text{Factor A SS} = SSR(U_1, U_2, \ldots, U_{a-1})$$

The next step is to fit

$$y_{ij} = \mu + \alpha_1 U_1 + \alpha_2 U_2 + \ldots + \alpha_{a-1} U_{a-1} + \beta_1 V_1 + \beta_2 V_2 + \ldots + \beta_{b-1} V_{b-1} + \epsilon_{ij}$$

(3.36)

to obtain $SSE(\mu, \alpha_1, \alpha_2, \ldots, \alpha_{a-1}, \beta_1, \ldots, \beta_{b-1})$ from which

$$\text{Factor B SS} = SSE(\mu, \alpha_1, \alpha_2, \ldots, \alpha_{a-1}) - SSE(\mu, \alpha_1, \alpha_2, \ldots, \alpha_{a-1}, \beta_1, \ldots, \beta_{b-1})$$

$$= SSR(V_1, V_2, \ldots, V_{b-1}|U_1, U_2, \ldots, U_{a-1})$$

$$= SSR(V_1, V_2, \ldots, V_{b-1})$$

(3.37)
Notice that the order of fitting does not matter since the columns of the matrix \( \mathbf{X} \) are uncorrelated. It is not difficult to show that the sums of squares above are just the usual sums of squares for a two way AOV as given below.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grand Mean</td>
<td>1</td>
<td>( \text{SSR}(\mu) = ab \overline{y}^2 )</td>
</tr>
<tr>
<td>A Factor Means</td>
<td>( a - 1 )</td>
<td>( \text{SSR}(U_1, U_2, \ldots, U_{a-1}) = b \sum_i (\overline{y}_i - \overline{y})^2 )</td>
</tr>
<tr>
<td>B Factor Means</td>
<td>( b - 1 )</td>
<td>( \text{SSR}(V_1, V_2, \ldots, V_{b-1}) = a \sum_j (\overline{y}_j - \overline{y})^2 )</td>
</tr>
<tr>
<td>Residual</td>
<td>((a - 1)(b - 1))</td>
<td>Difference = ( \sum_i \sum_j (y_{ij} - \overline{y}_i - \overline{y}_j + \overline{y})^2 )</td>
</tr>
<tr>
<td>Total</td>
<td>( ab )</td>
<td>( \mathbf{y}'\mathbf{y} = \sum_i \sum_j y_{ij}^2 )</td>
</tr>
</tbody>
</table>

### 3.5.2 Hypotheses and Tests

The \( A \) Factor Means SS (\( A \) SS) is used to test the hypothesis

\[ \alpha_1 = \alpha_2 = \ldots = \alpha_a \]

using

\[
\frac{(A \text{ SS})/(a - 1)}{(\text{Residual SS})/(a - 1)(b - 1)}
\]  \hspace{1cm} (3.38)

which is distributed as an \( F[a - 1, (a - 1)(b - 1)] \).

The \( B \) SS is used to test the hypothesis that

\[ \beta_1 = \beta_2 = \ldots = \beta_b \]

using

\[
\frac{(B \text{ SS})/(b - 1)}{(\text{Residual SS})/(a - 1)(b - 1)}
\]  \hspace{1cm} (3.39)

which is distributed as an \( F[(b - 1), (a - 1)(b - 1)] \).

These results can be proved using Cochran’s Theorem in a similar manner as was used for the oneway AOV.
Chapter 4

Model Assessment

4.1 Introduction

In chapters 2 and 3 the problem of fitting a linear model to a set of data was discussed. A natural question to ask is how well this model fits the data, that is, does the model provide an adequate representation of the data? The lack of fit test (section 1.5) provides a partial answer to the question when there are multiple y observations for each x but doesn’t help in other cases. One objective of this chapter is to try and assess the adequacy of a model as well as examining some problems that can occur.

A second objective is to examine any assumptions made either in the fitting process or in the inference associated with that model.

Finally, some methods of overcoming problems in model fitting will be examined.

4.2 Residual Analysis

In the definition of the linear model (2.4) the vector of residuals, $\epsilon$, was assumed to have a $n$-variate normal distribution, $N(0, \sigma^2 I)$. The residual vector is estimated by

$$
e = Y - \hat{Y}$$

$$= Y - X\hat{\beta}$$

$$= Y - X(X'X)^{-1}X'Y$$

$$= [I - X(X'X)^{-1}X']Y$$

Thus $\epsilon$ can be written as

$$\epsilon = (I - H)Y$$  \hspace{1cm} (4.2)

where

$$H = X(X'X)^{-1}X'$$  \hspace{1cm} (4.3)

is called the hat matrix.
Now $Y \sim N_n(X\beta, \sigma^2I)$ so that $(I - H)Y \sim N_n(0, \sigma^2(I - H))$.

**Exercise:** (a) Show $H$ is idempotent.

(b) Prove $\hat{Y} = HY$

(c) Prove $e \sim N_n(0, \sigma^2(I - H))$.

The vector of estimated residuals has a $n$-variate normal distribution if the assumption of normality is true but the components, $e_i$, $(i = 1, 2, \ldots, n)$ are not independent. However, when the sample size is large compared to the number of parameters in the model this dependency is relatively unimportant and can generally be ignored.

It is reasonable to assume that the estimated residuals, $e_i$, should reflect the properties assumed for the (unknown) true residuals $\epsilon_i$. It is this assumption about the residuals that is utilised in examining the aptness of a model. (For convenience the estimated residuals, $e_i$, will be referred to simply as residuals.) Several departures from the model or model assumptions can be detected using residuals.

1. The $\epsilon_i$ do not have constant variance.
2. The regression function is not linear.
3. The model fits well except at one (or a few) points (called outliers).
4. One or more important independent variables are missing from the model.
5. The $\epsilon_i$ are not normally distributed.
6. The $\epsilon_i$ are not independent.

These will be studied graphically. In these graphical methods it can be an advantage to use standardized residuals instead of raw residuals (although the shape of the plot does not change).

### 4.2.1 Standardized Residuals

We will define standardized residuals as

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

where the MSE (which estimates $\text{var}(\epsilon_i) = \sigma^2$) is obtained from the regression analysis, and $h_{ii}$ is the $i$th diagonal element of the hat matrix, $H$.

**Warning:** We use different definitions of standardized and (later) studentized residuals to NWK. Our definitions are consistent with the definitions in many texts and statistical packages including SPlus.

### 4.2.2 Residuals versus an Independent Variable

There are several possibilities when plotting residuals. One possibility is to plot the residuals against an independent variable, $x_i$. As the residuals should be normally distributed with a zero mean and common variance (and this should in no way depend on the size of the
independent variable) the impression gained from this scatter plot should be of a horizontal band as in figure 4.1(a). Departures from this horizontal band indicate abnormalities in the model assumptions or fit.

![Graphs showing different patterns](image)

(a) Uniform  (b) Increasing Variance  
(c) Linear Term Omitted  (d) Curvilinear Term Omitted

Figure 4.1: Typical Patterns in Residual Plots

**Homogeneity of Variance**

A plot of the form of fig 4.1(b) indicates that the variance of the residuals is increasing as the x value increases. The assumption of constant variance appears to be violated.

**Linear Term Omitted**

A plot of the form 4.1(c) would indicate a linear term may have been omitted.

**Nonlinearity of the Model**

A plot of the form of fig 4.1(d) indicates that the model is not fitting and suggests that a curvilinear function is probably required.
4.2.3 Residuals versus Fitted Values

When the residuals are plotted against the corresponding fitted values $\hat{y}_i$, $i = 1, 2, \ldots, n$ these plots are interpreted in a similar manner to those where residuals are plotted against an independent variate.

The residual plots in Fig 4.2 above were obtained by fitting a straight line to various data sets and then plotting the residuals against $\hat{y}_i$.

There are no obvious patterns or trends in Fig 4.2 (a). No obvious departures from the assumptions are apparent from this graph.

There is an obvious pattern in Fig 4. (b) which indicates that the variance of the residuals increases as X increases. That is the assumption of constant variance appears to be violated.

There is a curvilinear trend apparent in Fig 4.2 (c). This is an indication that the linear model is not appropriate. Try fitting a quadratic.
4.2.4 Testing for Normality: Q-Q Plots

To check the normality assumption graphically a normal probability plot (Q-Q plot), can be used. The ordered residuals are plotted against the expected values of the ordered residuals from a normal distribution. If the residuals are normally distributed they should lie on a straight line.

In Fig 4.3 the points do appear to fall reasonably close to a straight line suggesting the residuals are approximately normally distributed. Any obvious departure from a straight line would indicate a problem with the normality assumption.

4.2.5 Time Sequence Plots

In figure 4.4 the residuals are plotted in the order they are generated. There appears to be little wrong with the impression gained from fig 4.4(a), the plot appears uniform. However figure 4.4(b) is the same plot but with adjacent points joined.

Notice that every positive residual is followed by a negative residual and every negative by a positive. This illustrates a problem with the independence of the residuals. In particular this is typical of a plot when there is negative autocorrelation present.

Figure 4.5 illustrates the situation where there is positive autocorrelation.

4.2.6 Durbin-Watson Test for Autocorrelation

If instead of assuming the errors, \( \epsilon_t \), are independent it is assumed they are of the form

\[
\epsilon_t = \rho \epsilon_{t-1} + u_t
\]

(4.5)
then the random errors are said to follow a first order autoregressive process. The parameter \( \rho \) is the *autocorrelation* parameter and satisfies \(|\rho| < 1\). The \( u_t \) are independent \( \text{N}(0, \sigma^2) \). The error term \( \epsilon_t \) consists of a fraction (\( \rho \)) of the error term in the preceding period \( \epsilon_{t-1} \) plus \( u_t \) a new random component.

To test the hypothesis

\[
H : \rho = 0 \quad \text{against the alternative} \quad A : \rho > 0
\]

the Durbin-watson statistic is defined by

\[
D = \frac{\sum_{i=2}^{n}(\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n} \epsilon_i^2}
\]

(4.6)

When positive autocorrelation is present \( \epsilon_t \) and \( \epsilon_{t-1} \) are related so that the difference \( \epsilon_t - \epsilon_{t-1} \) tends to be small. Tables for \( D \) consist of upper and lower limits for \( D \). Values of \( D \) greater than the upper limit support the hypothesis with values below the lower limit supporting the alternative. Values between the limits are inconclusive.

To test for negative autocorrelation, calculate \( 4 - D \) and proceed in the same manner.
4.3 Transformations

When the assumptions for regression analysis are violated then the situation must be remedied. One common method is to try transforming the variables in the equation. Transformations can be used to

1. achieve linearity,
2. stabilize variance,
3. make the distribution of the data closer to a normal distribution.

4.3.1 Linearity Transformations

Instead of the usual simple linear regression model assume the data fits a model of the form

\[ y = \alpha e^{-\beta x} \epsilon. \]

This equation is not linear in the parameters and has a multiplicative error term. Taking logs to the base \( e \) of both sides,

\[ \log_e(y) = \log_e(\alpha) + \beta x + \log_e(\epsilon) \]

which is linear in the parameters and has a linear error term so that the usual methods can be employed to regress \( \log(y) \) against \( x \). (Logs to any base can be used although it is very uncommon to use other than base \( e \) or base 10.)

In the same way if the model were

\[ y = \alpha x^\beta \epsilon \]
then linearity would be obtained by taking logarithms of both $y$ and $x$ and regressing $\log(y)$ against $\log(x)$.

Likewise a reciprocal transformation of $x$ would achieve linearity if

$$y = \alpha + \frac{\beta}{x} + \epsilon .$$

### 4.3.2 Non-Constant Variance

If the plot of residuals against $X_i$ fans out then it indicates the variance increases as $X_i$ increases. If the variance is related to the mean response, $E(Y)$, then statistical theory indicates the following transformations as being appropriate:

1. $E(Y)$ proportional to $\sigma^2$, use $\sqrt{Y}$.
2. $E(Y)$ proportional to $\sigma$, use $\log(Y)$ (base $e$ or 10).
3. $E(Y)$ proportional to $\sqrt{\sigma}$, use $\frac{1}{\sqrt{Y}}$.

After a transformation has been selected, residual plots and other remedial methods should be used with the analysis on the transformed data. Also be aware that transformations on the $X$ variable do not affect the shape or variability of the error distribution. Transformations on the $Y$ variable do affect the error distribution.

**Comment: Non-Normality.**

Pronounced departures from normality can often be remedied by transformation of the dependent variable. It is important to note that several problems already discussed may make the error term *look* non-normal. The incorrect functional form, omitted variables, and non-constant variance can cause the error term to appear non-normal. Fixing these problems often corrects the non-normality problem as well.

### 4.4 Multicollinearity

In chapter 2, section 2.4 the problem of testing $\beta_i=0$ was considered. The question asked was if having fitted the other $p-2$ predictor variables was it really necessary to include $X_i$ as well. In particular, this is a question that must be considered when the independent variables, $X_1, X_2, \ldots, X_{p-1}$, are interrelated (or dependent on each other). When this occurs *multicollinearity* is said to exist among the independent variables. In example 2.3.1, it is to be expected that the diameter of a tree and its height may be related to each other, height increasing with diameter.

Many of the problems associated with multicollinearity stem from the need to invert the matrix $X'X$. If a linear relationship exists between the columns of $X$ then $X$ is not of full rank which implies $X'X$ is also not of full rank and can’t be inverted. While in regression it is unusual to find a perfect relationship between the columns of $X$, in practice
a near relationship may exist so that $X'X$ is close to singular (ill conditioned). Estimates that involve inverting $X'X$ then tend to be sensitive to small changes in the data, leading to the type of problems now discussed.

4.4.1 Problems Associated with Multicollinearity

The results obtained in example 2.3.1 will be used to illustrate some of the problems associated with multicollinearity.

1. Addition or deletion of an independent variable changes the regression coefficients. In example 2.3.1 where log(Vol) was the dependent variable and log(Diam) and log(Height) the independent variables, the following three models were obtained

   (a) log(Vol) = -6.52 + 1.99 log(Diam) + 1.09 log(Height)
   (b) log(Vol) = -2.36 + 2.20 log(Diam)
   (c) log(Vol) = -13.9 + 3.98 log(Height)

   The coefficients change as the independent variables in the model change. It is not unusual for the signs of the coefficients to change as well as the magnitude. This can cause interpretation of the equation to be very difficult.

2. The additional sum of squares attributable to an independent variable changes depending on what variables have already been fitted. Notice in example 1.2 that $SSR(X_1) = 7.93$ but $SSR(X_1|X_2) = 4.61$.

3. The test of the partial regression coefficients may not be statistically significant even though a strong relationship exists between the dependent variable and the independent variables. That is the test of the hypothesis

   $$\beta_1 = \beta_2 = \ldots \beta_p = 0$$

   may be highly significant but all of the tests of the hypotheses

   $$\beta_i = 0$$

   may be non-significant. This is not a contradiction but rather would reflect a situation where the addition of the last variable to the equation added little or no information due to the fact that the independent variables were related. Hence there is an overlap of information between the variables.

   Note that the problems above can arise without substantial multicollinearity between variables although it is unusual. It is not unusual however that multicollinearity may occur due to linear dependence between more than two variables in which case it may not be particularly evident if only the correlations between pairs of variables are examined.
Multicollinearity is often found in observational data where the values of the independent variables are not controlled (such as observing diameter and height in the tree data).

As a contrast to multicollinearity and example follows where the predictor variables are independent. This is experimental data in which it was possible to design the experiment so that the predictor variables were orthogonal to each other and so uncorrelated.

**Example 4.4.1: Uncorrelated Predictor Variables**

The independent variables are the dose of antibiotic, $X_1$ and the dose of vitamin $B_{12}$, $X_2$. The average daily weight gain of pigs, $Y$, was the measured variable. The results are given in the following table.


<table>
<thead>
<tr>
<th>antibiotic (g)</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>40</th>
<th>40</th>
<th>40</th>
<th>40</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_{12}$ (g)</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>wt-gain</td>
<td>1.30</td>
<td>1.19</td>
<td>1.08</td>
<td>1.26</td>
<td>1.21</td>
<td>1.19</td>
<td>1.05</td>
<td>1.00</td>
<td>1.05</td>
<td>1.52</td>
<td>1.56</td>
</tr>
</tbody>
</table>

1. The regression equation is

$$\text{wt-gain} = 1.07 + 0.00208 \text{ antibiotic} + 0.0540 \text{ B}_{12}$$

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>0.23953</td>
<td>0.11977</td>
</tr>
<tr>
<td>Error</td>
<td>9</td>
<td>0.20213</td>
<td>0.02246</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>0.44167</td>
<td></td>
</tr>
</tbody>
</table>

2. The regression equation is

$$\text{wt-gain} = 1.21 + 0.00208 \text{ antibiotic}$$

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>0.02084</td>
<td>0.02084</td>
</tr>
<tr>
<td>Error</td>
<td>10</td>
<td>0.42083</td>
<td>0.04208</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>0.44167</td>
<td></td>
</tr>
</tbody>
</table>

3. The regression equation is

$$\text{wt-gain} = 1.11 + 0.0540 \text{ B}_{12}$$

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>0.21870</td>
<td>0.21870</td>
</tr>
<tr>
<td>Error</td>
<td>10</td>
<td>0.22297</td>
<td>0.02230</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>0.44167</td>
<td></td>
</tr>
</tbody>
</table>

64
From above notice that

\[
SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2) = 0.22297 - 0.20213 = 0.02084
\]

\[
SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2) = 0.42083 - 0.20213 = 0.21870
\]

Thus in the case of uncorrelated independent variables

- SSR(X_1) = SSR(X_1|X_2), i.e., the SS due to fitting X_1 alone is not affected by fitting X_2 first. Thus the reduction in the residual SS due to fitting X_1 is the same whether or not X_2 has already been included in the model. (Similarly for X_2.)

- The regression coefficients of the X’s remain the same whether one or both independent variables are fitted.

**Example 4.4.2: Correlated Predictor Variables**

Return again to the relationship between labour required in U.S Naval Hospitals and the 5 predictor variables in example 3.2.1 of chapter 2.

Minitab was used to carry out all calculations for the results that follow.

**Correlation Matrix**

<table>
<thead>
<tr>
<th></th>
<th>lhours</th>
<th>pload</th>
<th>xrays</th>
<th>beddays</th>
<th>pop</th>
</tr>
</thead>
<tbody>
<tr>
<td>pload</td>
<td>0.986</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xrays</td>
<td>0.945</td>
<td>0.907</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beddays</td>
<td>0.986</td>
<td>1.000</td>
<td>0.907</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pop</td>
<td>0.940</td>
<td>0.936</td>
<td>0.910</td>
<td>0.933</td>
<td></td>
</tr>
<tr>
<td>lstay</td>
<td>0.579</td>
<td>0.671</td>
<td>0.447</td>
<td>0.671</td>
<td>0.463</td>
</tr>
</tbody>
</table>

The regression equation for the full model is

\[ l\text{hours} = 1962 - 15.8 \text{pload} + 0.05599 \text{xrays} + 1.59 \text{beddays} - 4.21 \text{pop} - 394 \text{lstay} \]

| Predictor | Value | Std. Error | t value | Pr(> |t|) |
|-----------|-------|------------|---------|------|
| (Intercept) | 1962  | 1071       | 1.83    | 0.094|
| pload      | -15.76| 97.65      | -0.16   | 0.875|
| xrays      | 0.05594| 0.02126    | 2.63    | 0.023|
| beddays    | 1.586 | 3.092      | 0.51    | 0.618|
| pop        | -4.210| 7.176      | -0.59   | 0.569|
| lstay      | -394.1| 209.6      | -1.88   | 0.087|

For the full model \( R^2 = 99.1\% \) and \( R^2(\text{adj}) = 98.7\% \)

The Analysis of Variance Table for the full model is then
<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
<th>F Value</th>
<th>Pr(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>5</td>
<td>490172288</td>
<td>98034456</td>
<td>237.82</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>11</td>
<td>4534340</td>
<td>412213</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>494706624</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thus the test of $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ is very highly significant, where $\beta_1$ is the coefficient of pload, $\beta_2$ is the coefficient of x-rays, $\beta_3$ is the coefficient of beddays, $\beta_4$ is the coefficient of pop and $\beta_5$ the coefficient of lstay. This implies not all the $\beta$'s are zero.

The test that $\beta_2=0$ given that the other four $\beta$'s are already in the model is also significant. Tests of all other partial regression coefficients are non-significant.

The five simple regressions for lhours against each of the other variables is now fitted. The slope of the line is tested in each case using a t-test.

1. lhours = - 68 + 34.0 pload  
   Predictor      Value  Std. Error  t value  Pr(>|t|)  
   (Intercept)   -67.7    324.2     -0.21   0.837  
   pload          34.033   1.505     22.61   0.000  

2. lhours = 492 + 0.247 xrays  
   Predictor       Value  Std. Error  t value  Pr(>|t|)  
   (Intercept)    492.3     605.8      0.81   0.429  
   xrays          0.24700   0.02203    11.21   0.000  

3. lhours = - 28 + 1.12 beddays  
   Predictor       Value  Std. Error  t value  Pr(>|t|)  
   (Intercept)   -28.0     319.0     -0.09   0.931  
   beddays      1.11739     0.04880    22.90   0.000  

4. lhours = - 171 + 48.4 pop  
   Predictor       Value  Std. Error  t value  Pr(>|t|)  
   (Intercept)  -171.1     675.1     -0.25   0.803  
   pop           48.437     4.524     10.71   0.000  

5. lhours = - 6991 + 2031 lstay  
   Predictor       Value  Std. Error  t value  Pr(>|t|)  
   (Intercept)  -6991.0     4502     -1.55   0.141  
   lstay        2031.0      739.2      2.75   0.015  

Observations:

1. The correlation matrix indicates the predictor variables are correlated.

2. The tests of $\beta = 0$ in the individual equations are all highly significant whereas only the coefficient of ‘x-rays’ is significant in the full model.
3. The coefficients of the variables are different in the full model to when they are fitted alone in the five simple regression models. Three of the coefficients change sign.

This indicates that multicollinearity is a problem in this data set.

**Comment on Interpretation:**

Notice that in the individual equations pload, pop, and lstay all have positive signs but that they have negative signs in the full equation. If we try to interpret the coefficient of say pload by holding all other variables except pload constant in the full model, then increasing pload would seem to imply that the labour hours required decrease which is contrary to what is actually expected. This method of interpretation cannot be used when the variables are correlated. When variables are correlated the value of one variable cannot be assumed to change without affecting the values of the other variables.

### 4.4.2 Polynomial Regression

Polynomial regression is a special case of multiple regression. Consider the polynomial model with one independent variable

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_{p-1} X_i^{p-1} + \epsilon$$

In terms of the multiple regression model equate $X_1$ with $X$, $X_2$ with $X^2$, ... and $X_{p-1}$ with $X^{p-1}$. Hence the methods used for multiple regression can be used for polynomial regression. It should also be realised that $X$, $X^2$, $X^3$, ... are usually highly correlated which must be taken into account when fitting and assessing polynomial models.

A question often asked is what order polynomial should be fitted to a given set of data. To do this the individual $\beta$’s are tested in order (i.e., $\beta_1, \beta_2, \ldots$) Thus the procedure is

1. Fit $X$ alone and test $H : \beta_1 = 0$.

2. Add $X^2$ to the model and test $H : \beta_2 = 0$. That is test if adding the quadratic term is worthwhile after having fitted the linear term. If this is not significant it is an indication the linear may be sufficient. If it is significant

3. Add $X^3$ to the model and test $H : \beta_3 = 0$.

This process is continued until non-significance is obtained.

Note that in polynomial regression once we decide the order of the polynomial it is usual to fit all terms of lower order as well. That is if we decide to fit a cubic it would be unusual to leave out the quadratic (or linear) terms.
4.4.3 A Note on Extrapolation

Multicollinearity does not effect the ability to find a model that ‘fits’ the data or inferences about mean responses or predictions of new observations, provided there is no extrapolation.

Example 4.4.3

Consider the independent variables $X_1$, and $X_2$ each which has the range $[0, 1]$ with the predicted value of $Y$ satisfying $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$. Two different situations are represented in Figure 4.4 where $x$ represents the point $(0.6, 0.1)$.

![Figure 4.4: Region of Observations](image)

As the range of both independent variables is $[0, 1]$ it would seem safe to predict the value of $Y$ at the point $(0.6, 0.1)$. However the shaded area represents the actual region covered by the combinations of $(X_1, X_2)$ in the data, then $(0.6, 0.1)$ is outside the region in Figure 4.6(a) and it is not safe to use the equation for prediction purposes. The point $(0.6, 0.1)$ is inside the region in Figure 4.6(b) and the equation could be used for prediction purposes.

Be very careful! When there are two (or more) independent variables it is very easy to fall into the trap of extrapolating outside the range covered by the data. As is illustrated in the above example it is not sufficient to look at the ranges of the individual independent variables in isolation.

4.5 Detecting Multicollinearity

The following are indicators that multicollinearity may be a problem.
1. Adding or deleting variables cause large changes in estimated regression coefficients.

2. Estimated regression coefficients for important independent variables may have large standard deviations leading to non-significant tests of individual regression coefficients or wide confidence intervals.

3. Estimated regression coefficients may have a sign that it opposite to that expected from theoretical considerations.

4. Correlations between pairs of independent variables may be large.

One formal method of detection is provided by the Variance Inflation Factor, (VIF).

### 4.5.1 Variance Inflation Factor

The variance inflation factor is defined as

\[ (VIF)_k = \frac{1}{1 - R^2_k} \] (4.7)

where \( R^2_k \) is the coefficient of multiple determination obtained when \( X_k \) is regressed on the \( p - 2 \) other \( X \) variables.

If \( X_k \) is not related to the other variables then \( R^2_k = 0 \) and \( (VIF)_k = 1 \). When \( R^2_k \neq 0 \) then \( (VIF)_k > 1 \) indicating an inflated variance for the regression coefficient. As \( R^2_k \) increases so does the value of the variance inflation factor. Since the value of \( R^2_k \) is an indicator of the strength of the linear association between \( X_k \) and the other \( X \) variables the largest \( (VIF)_k \) is often used as as indicator of the severity of of multicollinearity. A value larger than 10 is often regarded as indicating multicollinearity may be a problem in the estimation of the coefficients.

**Calculation:** The VIF’s of the \( k \) predictors can be calculated using the the fact that if each of the predictors is first standardized, that is

\[ x^*_i = \frac{x_i - \text{mean}(x_i)}{\text{stdev}(x_i)}, i = 1, 2, \ldots, k \]

then the VIF’s are given by the diagonal elements of \( r^{-1}_{xx} \) where \( r_{xx} \) is the correlation matrix of the \( k \) standardized predictors, \( x^*_i \).

**Example 4.4.2(cont)**

If the U.S Naval Hospital data is in the data frame, hosp.df, then the Variance Inflation factors for this example can be calculated using SPlus by:

```r
hosp1.df_hosp.df  #Declares a new data frame to hold standardized variates.
#Calculate standardized variates.
for (i in 1:7) hosp1.df[,i]_ (hosp.df[,i] - mean(hosp.df[,i]))/stdev(hosp.df[,i])
```

69
hosp1.df_as.data.frame(hosp1.df)

#Calculate VIF's
vif.diag(solve(cor(hosp1.df[,2:6])))

These have been added to the usual output from the SPlus summary() command to give:

| Predictor  | Value | Std. Error | t value | Pr(>|t|) | VIF  |
|------------|-------|------------|---------|---------|------|
| (Intercept)| 1962  | 1071       | 1.83    | 0.094   |      |
| pload      | -15.76| 97.65      | -0.16   | 0.875   | 9597.6 |
| xrays      | 0.05594 | 0.02126   | 2.63    | 0.023   | 7.9  |
| beddays    | 1.586 | 3.092      | 0.51    | 0.618   | 8933.1 |
| pop        | -4.210| 7.176      | -0.59   | 0.569   | 23.3 |
| lstay      | -394.1| 209.6      | -1.88   | 0.087   | 4.3  |

The extremely high values for the variance inflation factor of two of the variables would indicate that multicollinearity is a severe problem in this data set. This result should probably have been expected given the the correlation matrix and results of example 4.4.2.

**Example 4.4.3**

Now consider the model selected for this problem by stepwise (and all subsets regression. The results are summarised in the following table.

| Predictor  | Value  | Std. Error | t value | Pr(>|t|) | VIF  |
|------------|--------|------------|---------|---------|------|
| (Intercept)| 1523.0 | 786.7      | 1.94    | 0.075   |      |
| xrays      | 0.05300 | 0.02009    | 2.64    | 0.020   | 7.7  |
| beddays    | 0.9784  | 0.1051     | 9.31    | 0.000   | 11.3 |
| lstay      | -320.8  | 153.1      | -2.09   | 0.056   | 2.5  |

Multicollinearity does not seem to be such a problem with only one variable having a VIF greater than 10.

This table would also indicate that it may be useful to consider omitting the non-significant variable, lstay, from the model. For this model

| Predictor  | Value  | Std. Error | t value | Pr(>|t|) | VIF  |
|------------|--------|------------|---------|---------|------|
| (Intercept)| -68.2  | 228.4      | -0.30   | 0.770   |      |
| xrays      | 0.07487| 0.01912    | 3.91    | 0.002   | 5.6  |
| beddays    | 0.82284| 0.08294    | 9.92    | 0.000   | 5.6  |

Both x-rays and beddays have significant coefficients and VIF’s less than 10. Also the adjusted $R^2$ is 98.5 and the $C_p$ value is 4.9 which would also indicate the model may be reasonable.

These two models will be considered further in the examples that follow. Tables for various quantities calculated using the two models are on page 80. These will be used in the examples.
Model 1: hours = 1523 + 0.0530 xrays + 0.978 beddays - 321 lstay

<table>
<thead>
<tr>
<th>hosp</th>
<th>lhours</th>
<th>predicted</th>
<th>stdres</th>
<th>stdres</th>
<th>leverage</th>
<th>cook</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>566.5</td>
<td>688.5</td>
<td>-0.21164</td>
<td>-0.20369</td>
<td>0.120749</td>
<td>0.00154</td>
</tr>
<tr>
<td>2</td>
<td>696.8</td>
<td>722.1</td>
<td>-0.04674</td>
<td>-0.04491</td>
<td>0.226128</td>
<td>0.00016</td>
</tr>
<tr>
<td>3</td>
<td>1033.2</td>
<td>965.5</td>
<td>0.11805</td>
<td>0.11348</td>
<td>0.129664</td>
<td>0.00052</td>
</tr>
<tr>
<td>4</td>
<td>1603.6</td>
<td>1172.5</td>
<td>0.76471</td>
<td>0.75182</td>
<td>0.158762</td>
<td>0.02759</td>
</tr>
<tr>
<td>5</td>
<td>1611.4</td>
<td>1526.9</td>
<td>0.14363</td>
<td>0.13810</td>
<td>0.084914</td>
<td>0.00048</td>
</tr>
<tr>
<td>6</td>
<td>1613.3</td>
<td>1994.0</td>
<td>-0.65742</td>
<td>-0.64239</td>
<td>0.112011</td>
<td>0.01363</td>
</tr>
<tr>
<td>7</td>
<td>1854.2</td>
<td>1676.7</td>
<td>0.30172</td>
<td>0.29090</td>
<td>0.084078</td>
<td>0.00209</td>
</tr>
<tr>
<td>8</td>
<td>2160.6</td>
<td>1791.5</td>
<td>0.62704</td>
<td>0.61176</td>
<td>0.083005</td>
<td>0.00890</td>
</tr>
<tr>
<td>9</td>
<td>2305.6</td>
<td>2798.9</td>
<td>-0.83892</td>
<td>-0.82876</td>
<td>0.084596</td>
<td>0.01626</td>
</tr>
<tr>
<td>10</td>
<td>3505.9</td>
<td>4191.6</td>
<td>-1.18939</td>
<td>-1.21049</td>
<td>0.120262</td>
<td>0.04835</td>
</tr>
<tr>
<td>11</td>
<td>3571.9</td>
<td>3191.1</td>
<td>0.64495</td>
<td>0.62981</td>
<td>0.077335</td>
<td>0.00872</td>
</tr>
<tr>
<td>12</td>
<td>3741.4</td>
<td>4364.5</td>
<td>-1.11750</td>
<td>-1.12926</td>
<td>0.177058</td>
<td>0.06717</td>
</tr>
<tr>
<td>13</td>
<td>4026.5</td>
<td>4364.3</td>
<td>-0.56825</td>
<td>-0.55287</td>
<td>0.064498</td>
<td>0.00557</td>
</tr>
<tr>
<td>14</td>
<td>10343.8</td>
<td>8713.5</td>
<td>2.87111</td>
<td>4.56022</td>
<td>0.146451</td>
<td>0.35359</td>
</tr>
<tr>
<td>15</td>
<td>11732.2</td>
<td>12081.0</td>
<td>-1.00621</td>
<td>-1.00673</td>
<td>0.681763</td>
<td>0.54225</td>
</tr>
<tr>
<td>16</td>
<td>15414.9</td>
<td>15132.8</td>
<td>0.99126</td>
<td>0.99055</td>
<td>0.785480</td>
<td>0.89947</td>
</tr>
<tr>
<td>17</td>
<td>18854.4</td>
<td>19260.7</td>
<td>-1.78743</td>
<td>-1.97740</td>
<td>0.863247</td>
<td>5.04194</td>
</tr>
</tbody>
</table>

Model 2: hours = - 68 + 0.0749 xrays + 0.823 beddays

<table>
<thead>
<tr>
<th>hosp</th>
<th>lhours</th>
<th>predicted</th>
<th>stdres</th>
<th>stdres</th>
<th>leverage</th>
<th>cook</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>566.5</td>
<td>505.4</td>
<td>0.09413</td>
<td>0.09073</td>
<td>0.100524</td>
<td>0.00033</td>
</tr>
<tr>
<td>2</td>
<td>696.8</td>
<td>1187.6</td>
<td>-0.75329</td>
<td>-0.74106</td>
<td>0.095450</td>
<td>0.01996</td>
</tr>
<tr>
<td>3</td>
<td>1033.2</td>
<td>737.2</td>
<td>0.45501</td>
<td>0.44173</td>
<td>0.098234</td>
<td>0.00752</td>
</tr>
<tr>
<td>4</td>
<td>1603.6</td>
<td>886.5</td>
<td>1.10936</td>
<td>1.11933</td>
<td>0.109419</td>
<td>0.05040</td>
</tr>
<tr>
<td>5</td>
<td>1611.4</td>
<td>1592.6</td>
<td>0.02800</td>
<td>0.02756</td>
<td>0.082312</td>
<td>0.00002</td>
</tr>
<tr>
<td>6</td>
<td>1613.3</td>
<td>1918.2</td>
<td>-0.47153</td>
<td>-0.45803</td>
<td>0.108545</td>
<td>0.00902</td>
</tr>
<tr>
<td>7</td>
<td>1854.2</td>
<td>1752.6</td>
<td>0.15458</td>
<td>0.14908</td>
<td>0.080599</td>
<td>0.00070</td>
</tr>
<tr>
<td>8</td>
<td>2160.6</td>
<td>1728.1</td>
<td>0.65838</td>
<td>0.64448</td>
<td>0.080582</td>
<td>0.01266</td>
</tr>
<tr>
<td>9</td>
<td>2305.6</td>
<td>2928.8</td>
<td>-0.94570</td>
<td>-0.94188</td>
<td>0.074420</td>
<td>0.02397</td>
</tr>
<tr>
<td>10</td>
<td>3505.9</td>
<td>4444.8</td>
<td>-1.43022</td>
<td>-1.49145</td>
<td>0.081594</td>
<td>0.06058</td>
</tr>
<tr>
<td>11</td>
<td>3571.9</td>
<td>3324.7</td>
<td>0.37349</td>
<td>0.36171</td>
<td>0.066571</td>
<td>0.00332</td>
</tr>
<tr>
<td>12</td>
<td>3741.4</td>
<td>3964.6</td>
<td>-0.33990</td>
<td>-0.32889</td>
<td>0.080640</td>
<td>0.00338</td>
</tr>
<tr>
<td>13</td>
<td>4026.5</td>
<td>4276.4</td>
<td>-0.37624</td>
<td>-0.36440</td>
<td>0.059837</td>
<td>0.00300</td>
</tr>
<tr>
<td>14</td>
<td>10343.8</td>
<td>8964.6</td>
<td>2.13249</td>
<td>2.50084</td>
<td>0.108437</td>
<td>0.18436</td>
</tr>
<tr>
<td>15</td>
<td>11732.2</td>
<td>12771.5</td>
<td>-1.94949</td>
<td>-2.20092</td>
<td>0.394229</td>
<td>0.82445</td>
</tr>
<tr>
<td>16</td>
<td>15414.9</td>
<td>14467.9</td>
<td>1.99325</td>
<td>2.26960</td>
<td>0.51850</td>
<td>1.42812</td>
</tr>
<tr>
<td>17</td>
<td>18854.4</td>
<td>19184.6</td>
<td>-1.28720</td>
<td>-1.32100</td>
<td>0.859757</td>
<td>3.38582</td>
</tr>
</tbody>
</table>
4.5.2 Outliers

An observation may be an outlier with respect to its $Y$ value, its $X$ value or both. The hat matrix defined by (4.3) can be used to assist in identifying outliers.

The variance-covariance matrix of $\mathbf{e}$ has been shown to be

$$\sigma^2(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

so that

$$\sigma^2(\mathbf{e}_i) = \sigma^2(1 - h_{ii})$$

where $h_{ii}$ is the $i$th element on the main diagonal of the hat matrix. Thus

$$h_{ii} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i$$

where

$$\mathbf{X}_i' = (1, X_{i1}, X_{i2}, \ldots, X_{ip-1})$$

the vector associated with the $i$th observation, $y_i$. Since the hat matrix is a $(p \times p)$ idempotent matrix,

$$\text{tr}(\mathbf{H}) = \sum_i h_{ii} = p$$

Also,

$$0 \leq h_{ii} \leq 1$$

since

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{i \neq j} h_{ij}^2$$

$$h_{ii}(1 - h_{ii}) \geq 0$$

and (4.13) follows.

It can be shown that $h_{ii}$ can be interpreted as the distance of $X_i$ from the centre of the $X$ observations and so indicates whether or not it is an outlier. The element $h_{ii}$ is called the leverage of the observation. A large value indicates it is distant from the centre of the $X$ observations and is likely to have a substantial effect in determining the fitted value $\hat{y}_i$. Further, $\hat{\mathbf{Y}} = \mathbf{HY}$, so that

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij}y_j$$

Here, $h_{ii}$ is the coefficient of $y_i$. The larger the value of $h_{ii}$ is the more important $y_i$ is in determining the value of $\hat{y}_i$. Since $h_{ii}$ depends only on $X$ values it measures the influence of the $X$ values in determining the importance of $y_i$ in estimating $\hat{y}_i$.

Also notice from (4.9) that the larger the value of $h_{ii}$ the smaller the value of $\sigma^2(e_i)$. Observations with high leverage tend to have small residuals and so may not be detected by residual plots alone.
The mean leverage is defined by

\[
\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n}. 
\]

(4.14)

Values of \( h \) greater than twice \( \bar{h} \), that is values greater than \( \frac{2p}{n} \), are usually regarded as large.

**Example 4.5.1**

Consider again the US Naval Hospital example with the selected model

\[
\text{lhours } = 1523 + 0.0530 \text{ xrays} + 0.978 \text{ beddays} - 321 \text{ lstay}
\]

Using the table on page 77 hospitals 15, 16 and 17 have a leverage greater than \( \frac{2 \times 4}{17} = 0.471 \) and so might be seen as having a substantial effect in determining the fitted value.

The same hospitals have leverage greater than \( \frac{2 \times 3}{17} = 0.35 \) in the second model.

**Example 4.5.2**

The standardized residuals for both model 1 (three predictors) and model 2 (two predictors) are given in the tables on page 77. They are graphed in figures 4.5(a) and 4.6(a) on pages 80 and 81. Both residual plots look reasonable with no obvious trends although there is an outlier evident for model 1.

### 4.5.3 Deleted Residuals

As seen above if the \( i \)th observation exhibits a very strong influence on \( \hat{y}_i \) then the residuals must also be severely affected. If the \( i \)th observation is omitted and the regression fitted using only the remaining \( n - 1 \) observations, the \( i \)th observation cannot affect the predicted value. The predicted value is then denoted by \( \hat{Y}_{(i)} \) to indicate the \( i \)th observation has been omitted in the fitting of the regression line.

The *deleted residual* is then defined by

\[
d_i = Y_i - \hat{Y}_{(i)}.
\]

(4.15)

The estimated variance of \( d_i \) is then given by

\[
s^2(d_i) = MSE_{(i)}(1 + X_i'(X_i'(X_{(i)}X_{(i)})^{-1}X_i))
\]

(4.16)

where \( MSE_{(i)} \) is the mean square error when the \( i \)th observation is omitted in fitting the regression.

**Exercise:** Verify (4.16).
Figure 4.7: Graphs for Model 1, Naval Hospital Data
Figure 4.8: Graphs for Model 2, Naval Hospital Data
The *studentized deleted residual*, (which we will refer to simply as the *studentized residual*), is now defined by

\[ d_i^* = \frac{d_i}{s(d_i)} \quad (4.17) \]

It follows that

\[ d_i^* \sim t(n - p - 1) \quad (4.18) \]

although the \( d_i^* \) are not independent.

The studentized residual can be calculated from the residual, \( e_i \), the residual SS, \( SSE \), and the leverage, \( h_{ii} \) using

\[ d_i^* = e_i \left[ \frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2} \right]^{1/2} \quad (4.19) \]

Outlying \( Y \) observations are found by comparing the appropriate \( d_i^* \) values to tabulated values for the \( t \)-distribution.

### 4.5.4 Influential Observations

Having identified the \( i \)th observation as an outlier in either \( X \) or \( Y \) the aim is now to determine if it is exerting any undue influence in the fitting of the regression function. Two measures that suggest themselves are

\[ b - b_{(i)} \quad (4.20) \]

the difference between the estimated regression coefficient based on all \( n \) observations and that based on the \( n - 1 \) observations after deleting the \( i \)th, and

\[ \hat{Y}_i - \hat{Y}_{(i)} \quad (4.21) \]

#### Example 4.5.3

The studentized residuals are tabulated on page 77 and graphed in figures 4.5(b) and 4.6(b). The outlier is emphasised in Fig4.5(b) for model 1 but essentially disappears for the simpler model 2. This would provide further evidence that model 2 with only 2 predictors is probably adequate. This conclusion is also supported when the probability plots are examined (Fig4.5(c) and 4.6(c)). Fig 4.5(c) again appears to indicate the presence of an outlier that is not evident in Fig 4.6(c).

### 4.5.5 Cook’s Distance

Cook’s distance is defined as

\[ D_i = \frac{(b - b_{(i)})X'X(b - b_{(i)})}{p(MSE)} \quad (4.22) \]

76
and measures the impact on the differences in the regression coefficients when the \( i \)th observation is deleted.

It is known that Cook’s distance does not follow the \( F \) distribution but it has still found to be useful to compare the calculated values of \( D_i \) to the tabulated values of the \( F(p, n - p) \) distribution. If the percentile value is small (say less than 20\%) then the \( i \)th observation has little influence. If on the other hand the percentile exceeds 50\% the \( i \)th observation would appear to exert substantial influence on the regression function.

**Calculation of Cook’s Distance**

Cook’s distance, \( D_i \) can be calculated without fitting new regression functions where the \( i \)th observation is deleted. It can be shown that

\[
D_i = \frac{e_i^2}{pMSE \left(1 - h_{ii}\right)^2}
\]

where \( e_i \) is the \( i \)th residual and \( h_{ii} \) the leverage.

**4.6 Comment**

In this chapter some of the methods commonly used to examine the regression model have been presented. The treatment is by no means exhaustive and you are referred to books on regression diagnostics for further reading. However, it should be remembered that an important part of any regression analysis should be to check the assumptions and to examine the fit of the model.