Ordinary Least-Squares Regression

In: The Multivariate Social Scientist

By: Graeme D. Hutcheson

Pub. Date: 2011
Access Date: July 29, 2020
Publishing Company: SAGE Publications, Ltd.
Print ISBN: 9780761952015
Online ISBN: 9780857028075
DOI: https://dx.doi.org/10.4135/9780857028075
Print pages: 56-113

Ordinary Least-Squares Regression

Ordinary least-squares (OLS) regression is one of the most popular statistical techniques used in the social sciences. It is used to predict values of a continuous response variable using one or more explanatory variables and can also identify the strength of the relationships between these variables (these two goals of regression are often referred to as prediction and explanation).

OLS regression assumes that all variables entered into the analysis are continuous and the regression procedure attaches importance to actual values. Response variables, i.e., those variables which are being modelled, must be continuous and be recorded on at least an interval scale if they are to be modelled using OLS regression[1]. Response variables which cannot be assumed to be continuous may be more appropriately analysed using other generalized linear modelling techniques discussed in this book, such as logistic regression (for variables which are dichotomous) or loglinear analysis (for categorical variables in the form of frequency counts). Though explanatory variables are also required to be continuous, dichotomous data can legitimately be used in a regression model. This is particularly useful as it makes it possible to include multi-category ordered and unordered categorical explanatory variables in a regression model provided that they are appropriately coded into a number of dichotomous 'dummy' categories.

OLS regression is a generalized linear modelling technique, which, as the name suggests, models linear relationships. The three components of generalized linear models for OLS regression are a random component for the response variable, which is assumed to be Normally distributed, a systematic component representing the fixed values of the explanatory variables in terms of a linear function, and finally, a link function which maps the systematic component onto the random component. In OLS regression, this is simply an identity link which means that the fitted value of the response variable is the same as the linear predictor arising from the systematic component. This might appear to be quite restrictive since a number of the relationships one might wish to model are likely to be non-linear. It is possible, however, to model non-linear relationships using OLS regression if appropriate transformations are applied to one or more of the variables which render the relationships linear.

OLS regression is a powerful technique for modelling continuous data, particularly when it is used in conjunction with dummy variable coding and data transformation. This chapter discusses in some depth its application to different types of data, which are related both linearly and non-linearly, and demonstrates how models can be constructed for explanatory and predictive purposes.

3.1 Simple OLS Regression

A description of simple linear regression, where there is just a single explanatory variable, serves as an introduction to the more complex technique of multiple regression where a number of explanatory variables can be entered into a model simultaneously. Simple regression is used to model the relationship between a

continuous response variable *y* and an explanatory variable x.

3.1.1 The Regression Equation

For this discussion we will deal with a simple example of OLS regression where both the response and explanatory variables are continuous. A direct linear relationship between two such variables can be expressed in the form of an equation which identifies a straight line.

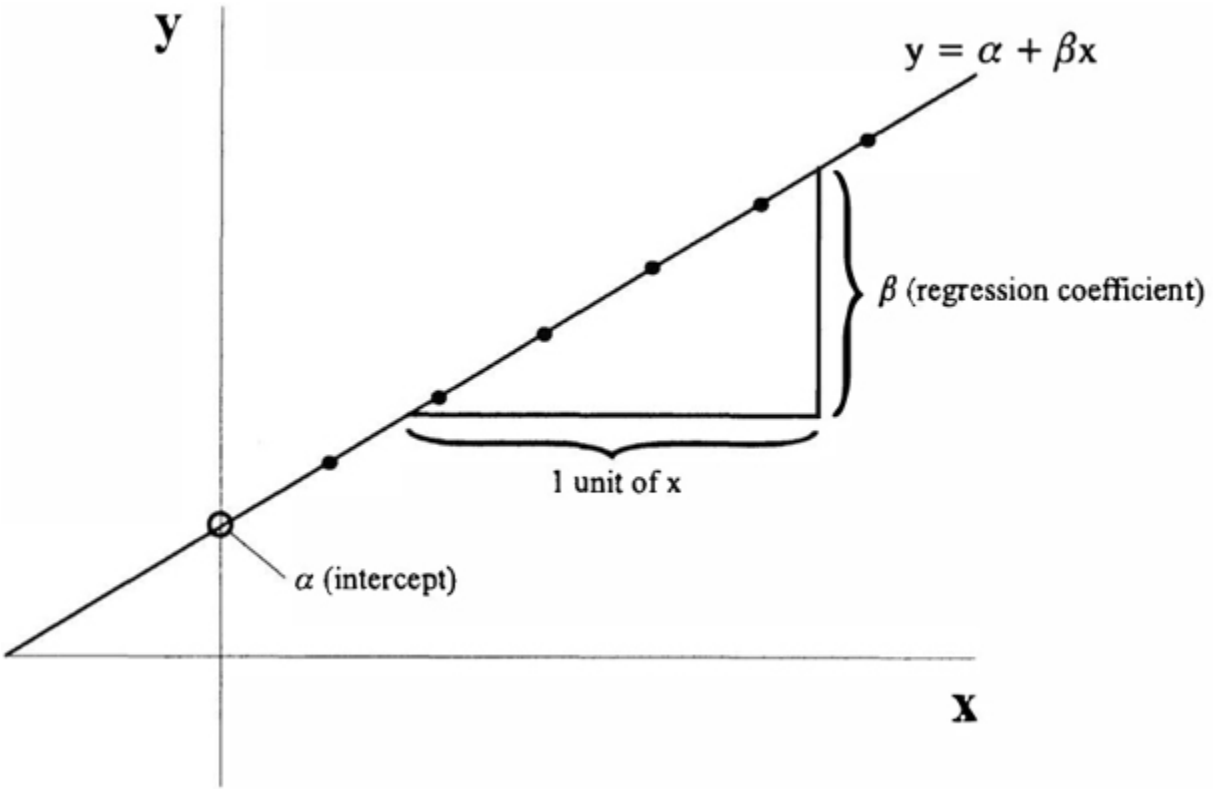
$$y = \alpha + \beta x \tag{3.1}$$

where α is the intercept of the line on the y axis,
and β is the slope of the line.

where *a* is the intercept of the line on the *y* axis,
and β is the slope of the line.

Equation 3.1 describes a direct linear relationship between *x* and *y* where the value of *y* can be precisely calculated from the value of *x*. The slope of the line can be described as the change in *y* which is associated with a unit change in *x*. For example, as *x* increases from 4 to 5 (a unit change), *y* increases by the value of β, the slope of the line. The slope of the line is also known as the *regression coefficient* and shows the effect that the explanatory variable has on the response variable. Figure 3.1 depicts the regression line for two variables which are perfectly linearly related and shows the regression coefficient, β, and the intercept, α.

Figure 3.1: An OLS regression model depicting a perfect linear relationship



In the social sciences, however, perfect relationships of the type shown in Figure 3.1 are the exception rather than the rule, as relationships are rarely direct and measurement rarely error-free. The best we can hope to do is to calculate a line of *best-fit* to approximately describe the relationship between variables *x* and *y*. For OLS regression, the most common method for calculating this line is to use the least-squares procedure which minimizes the sum of the squared deviations (also known as the error, or residual) of each data point from the line (for a full description of the least-square technique refer to an introductory statistics book such as Crawshaw and Chambers, 1984 or Hays, 1994). Equation 3.2 defines a line of best-fit ($y = \alpha + \beta x$) and includes a term which indicates the degree to which data points deviate from this line.

$$y = \alpha + \beta x + \varepsilon \qquad (3.2)$$

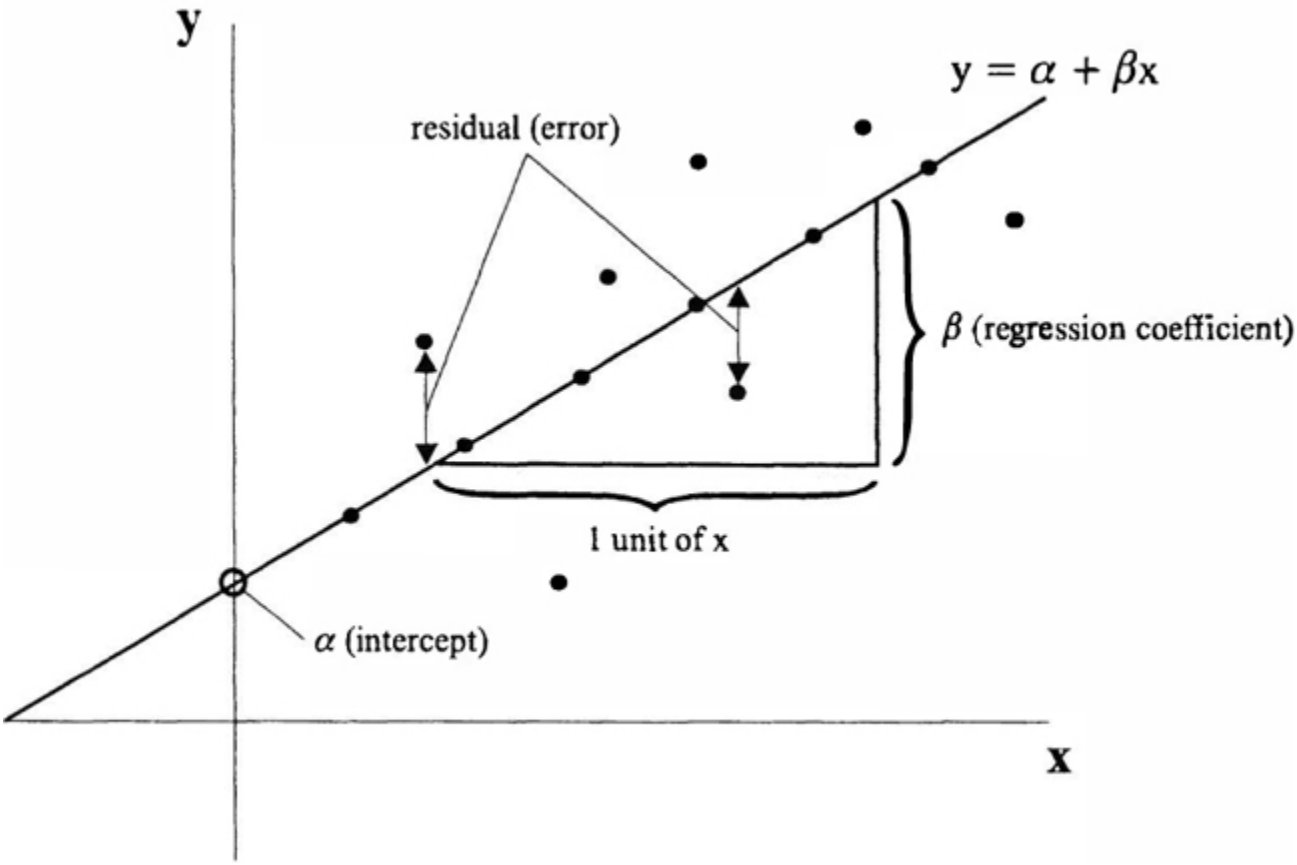
where ? represents the error.

The deviation, or error, is often represented as the difference between the observed value of *y* and the value of *y* predicted from the model ($\hat{y}$). The term $y–\hat{y}$ provides a measure of deviation, or, put another way, the amount of error when *y* is predicted using the regression model. It should be noted that the term ‘error’, as it is used here, does not indicate that any mistake has been made, it simply indicates that the relationship between the variables is not exact.

The interpretation of the parameters *α* and *β* in Equation 3.2 are slightly different to those provided for Equation 3.1. *α* now represents the *average* value of *y* when *x* = 0, whilst *β* represents the *average* change

in *y* which is associated with a unit change in *x*. Figure 3.2 shows the parameters *α* and *β* for a line of best-fit which models the relationship between two imperfectly related variables. The reader will note that the equation of the line does not include an error term since these are not explicitly estimated in regression models (error terms are, however, used in the construction of confidence and prediction intervals and in the calculation of goodness-of-fit statistics).

Figure 3.2: Line of best-fit



Using equations 3.1 and 3.2 it appears that one can predict the value of *y*, or at least the average value of *y*, for any given value of *x*. Whilst this is true mathematically, it is often unwise to predict values of a variable which are outside of the range of the observations. Ryan, 1997, distinguishes between *prediction*, where values of *y* are predicted within the observed range of the explanatory values, and *extrapolation*, where values of *y* are 'predicted' outside of the range of explanatory variables. For example, if a sample contained values of an explanatory variable that were between 0 and 25, it may not be appropriate to predict (extrapolate) the value of the response variable when the explanatory is equal to, say 60, as we do not know that our model will hold for such an extreme value.

Even though OLS regression is a linear modelling technique it can be used to accurately model non-linear relationships provided that one or both of the variables are transformed so that the relationship between them approximates to a straight line. Non-linear relationships which can be transformed to linear ones have been referred to as intrinsically or transformably linear. Failure to transform a non-linear relationship will not prevent

an OLS regression model being fitted to the data, but will result in a degraded analysis and adversely affect the model fit. A complete example of how an OLS regression model can be applied to non-linear data is provided in Section 3.2.

3.1.2 Confidence Intervals for β

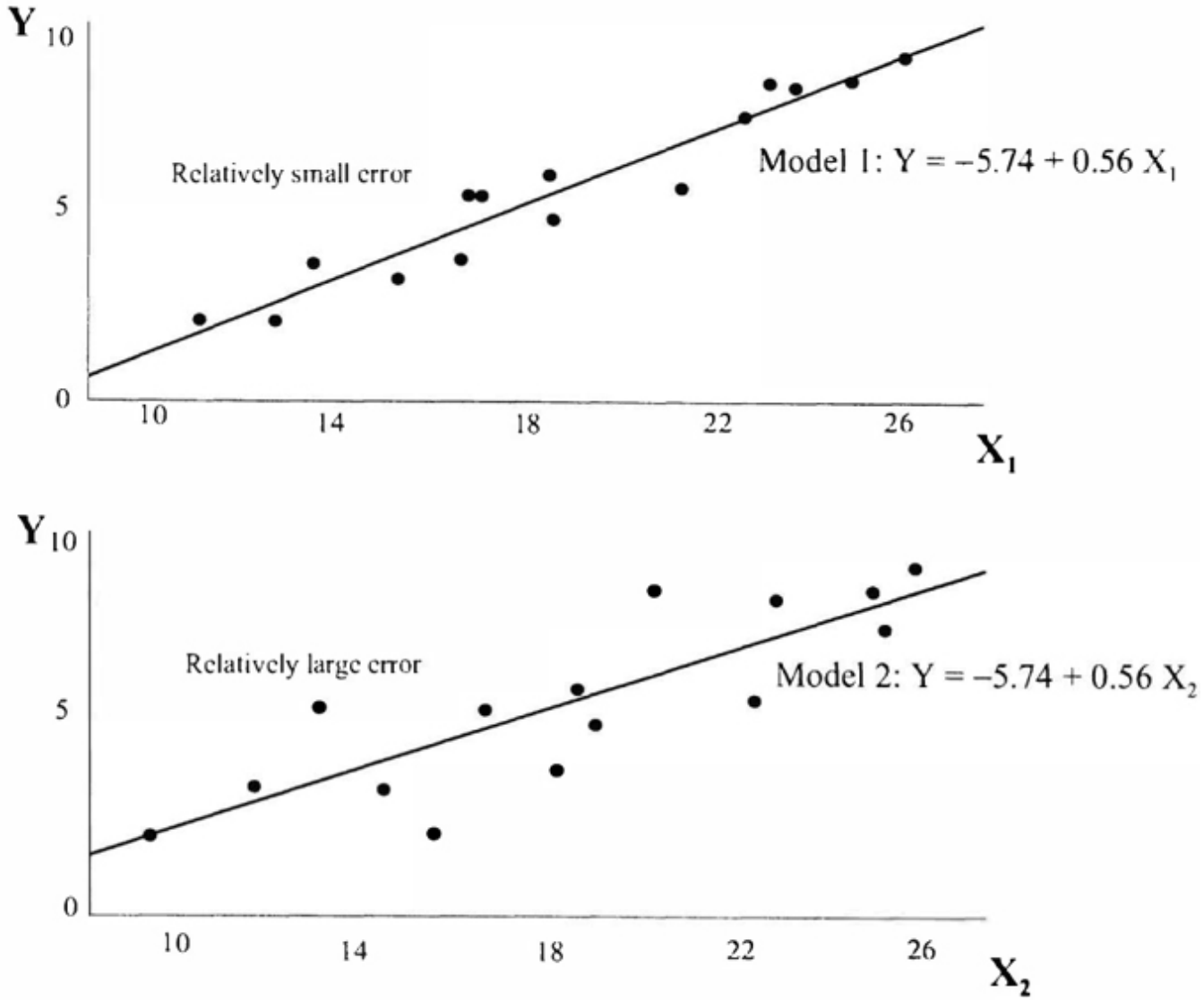
Although the value of β indicates the change in y which is associated with a unit change in x, its absolute magnitude does not indicate the 'strength' of the relationship between the variables. Consider the data presented in Table 3.1 and the accompanying scatterplots in Figure 3.3, for the three variables Y, X_1 and X_2.

Table 3.1: Data for Two Simple Regression Models

Y	X_1	X_2
0.52	10.50	11.09
2.09	13.19	12.91
4.52	16.62	14.09
1.93	14.94	15.29
0.63	12.17	16.15
4.52	16.85	17.06
2.52	16.35	18.34
5.19	18.34	18.72
4.02	18.40	19.03
8.21	23.21	20.07
4.77	21.30	21.87
7.97	23.81	22.27
8.24	25.02	23.98
7.01	22.68	24.19
8.92	26.28	24.75

The OLS regression line for the model $Y = \alpha+\beta X_1$ is identical (to 2 decimal places) to that for the model $Y = \alpha + \beta X_2$, even though the two scatterplots differ in the degree to which the data points are dispersed about the lines of best-fit. The relationship between Y and x_1 appears to be 'stronger' than the relationship between Y and x_2 (that is, β can be more precisely estimated), even though the value of β is the same in both cases. A useful exercise for determining the utility of the estimate of β is to identify the limits within which repeated samples can be expected to fall. These limits, or confidence intervals as they are commonly called, can be calculated using Equation 3.3.

Figure 3.3: Two simple regression models with different degrees of error



$$\text{confidence interval for } \beta = \hat{\beta} \pm t_{\alpha/2,n-k-1}(\text{s.e. } \hat{\beta}) \qquad (3.3)$$

where $\hat{\beta}$ is the estimated value of β,
$t_{\alpha/2,n-k-1}$ is the value of t given the confidence interval $\alpha/2$ and degrees-of-freedom $n-k-1$,
n is the number of cases used to construct the model,
k is the number of terms in the model (not including the constant),
and s.e. $\hat{\beta}$ is the standard error of $\hat{\beta}$.

The 'hat' symbol which appears above *β* indicates that this parameter is estimated. The term $t_{\alpha/2,n-k-1}$ provides the value of *t* for the confidence interval, *α*/2 (for example, a 95% two-tailed interval), with *n-k-1*

degrees-of-freedom. It should be noted that *k* indicates the number of *terms* in the model and not the number of *explanatory variables*. Although for this example, the number of terms in the model is equivalent to the number of *explanatory variables*, this isn't always the case since categorical variables may be represented using more than one term (see Section 3.3.7). It should also be noted that n refers to the number of cases used to construct the model and not the number of cases observed for a particular variable (cases with missing values for a particular variable are often completely removed from the analysis). The value of *t* can be derived using software or from statistical tables which provide critical values of the *t* distribution. For example, the value of *t* for a 95% two-tailed confidence interval with 19 degrees-of-freedom (that is, 21 observations and 1 term in the model) is equal to 2.093. Since the large sample (above about 30) approximation of *t* for 95% two-tailed confidence intervals is 1.96, Equation 3.3 can be simplified to Equation 3.4.

$$\text{confidence interval for } \beta = \hat{\beta} \pm 1.96(\text{s.e. } \hat{\beta}) \qquad (3.4)$$

where 1.96 is the large sample approximation of *t* for a two-tailed, 95% confidence interval.

Table 3.2: Statistics for the Regression Models Shown in Figure 3.3

					95% CIs	
	Coefficient	s.e.	t	P	Upper	Lower
Model 1						
X_1	0.562	0.041	13.813	0.000	0.474	0.650
(constant)	−5.738	0.782	−7.335	0.000	−7.428	−4.048
Model 2						
X_2	0.562	0.100	5.612	0.000	0.345	0.778
(constant)	−5.738	1.911	−3.002	0.010	−9.867	−1.609

Model 1: $Y = \alpha + \beta X_1$

Model 2: $Y = \alpha + \beta X_2$

The coefficients for variables x_1 and x_2 in Table 3.2 provide the values of β the slope of the line of best-fit) whilst the coefficients for the constants provide the values of a (the intercept of the line of best-fit on the *Y*-axis). Although identical regression equations are derived for both models ($Y = -5.738 + 0.562X$), Model I has a smaller standard error associated with the estimate of and β consequently has a smaller confidence interval. For Model I, we can say with 95% confidence that for each unit increase in x_1, *Y* is expected to increase somewhere between 0.474 and 0.650 (a range of 0.176). For Model 2, each unit increase in x_2 is expected to increase the value of *Y* somewhere between 0.345 and 0.778 (a range of 0.433). Clearly, a more accurate estimate of *Y* can be obtained using x_1 than can be obtained using x_2. These values were calculated using software (see Section 3.6), but could also have been calculated manually (as this is a relatively small sample we will use Equation 3.3 rather than Equation 3.4). For example, the confidence intervals for β which

are associated with x_1 can be calculated as

$$\text{confidence interval for fitted } y = \hat{y} \pm 1.96\sqrt{s_{\hat{y}}^2} \qquad (3.5)$$

which, allowing for rounding error, is the same result as that obtained using software (see Table 3.2).

Using these intervals, it is a simple matter to determine whether a significant linear relationship exists between the response and explanatory variables. Confidence intervals for β which include zero indicate no significant linear relationship as a unit change in x is associated with no change in y (x is therefore independent of y). For example, as neither of the confidence intervals for x_1 or x_2 include zero (see Table 3.2) both variables can be said to be significantly linearly related to the response variable at the 95% level. In OLS regression, the hypothesis that there is no linear relationship between the variables (i.e., that the slope of the line of best-fit is equal to zero) can be explicitly tested using the t statistic, which is calculated by dividing the regression coefficient by the standard error. The significance of $t_{n\text{-}k\text{-}1}$ can be determined using software or statistical tables. As a rough guide, the regression coefficient should be at least twice the value of the standard error for a statistically significant result.

3.1.3 Confidence Intervals for Fitted *y*

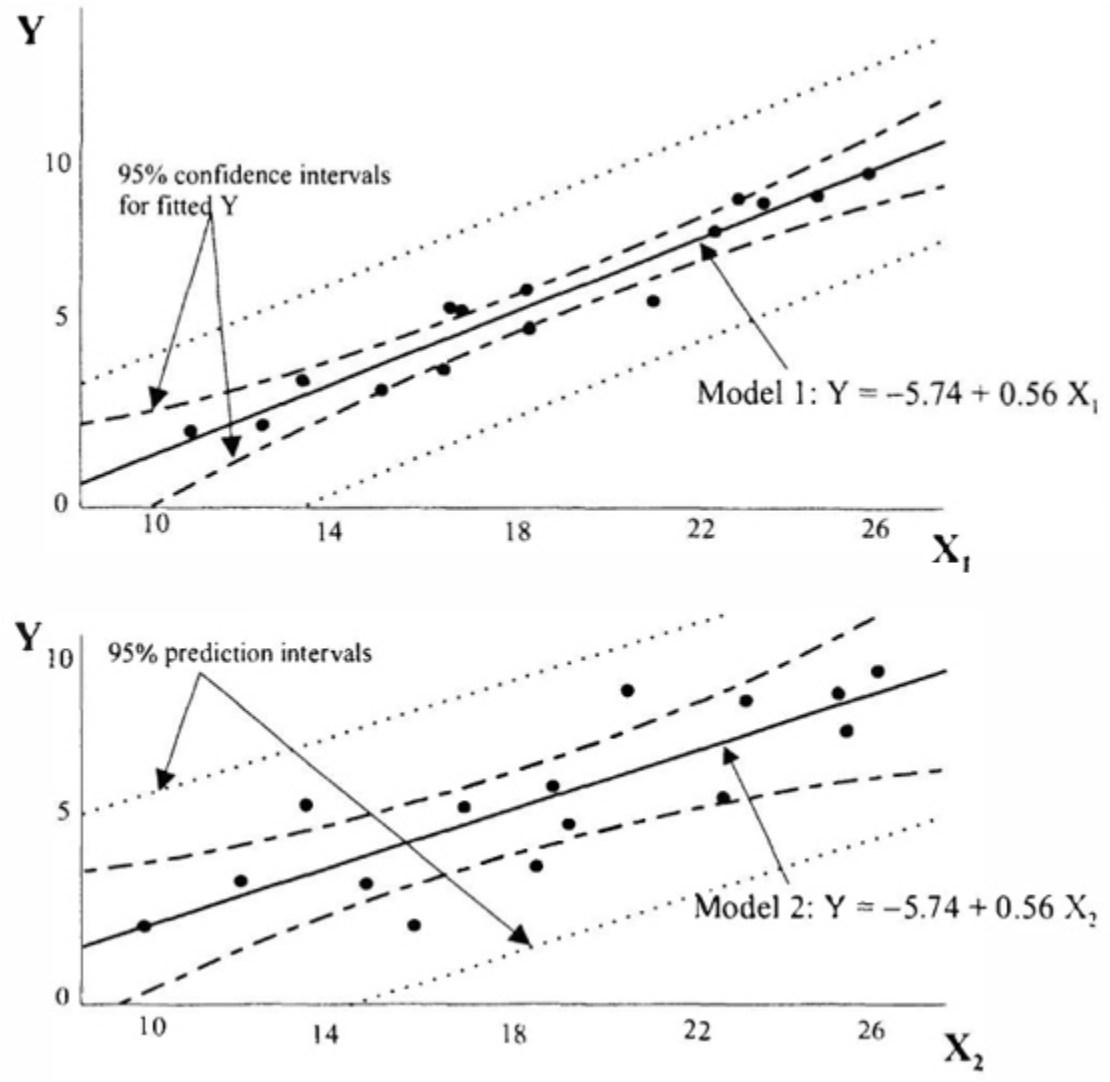
The fitted value for the response variable is the estimate of its mean value for a given value of the explanatory variable. Similar to the case of β, it is possible to construct confidence intervals for this mean fitted value. For large samples (n greater than about 30), 95% confidence intervals for fitted y can be calculated using Equation 3.5.

$$\text{confidence interval for fitted } y = \hat{y} \pm 1.96\sqrt{s_{\hat{y}}^2} \qquad (3.5)$$

where $\hat{y}$ is the fitted value of y,
1.96 is the large sample approximation of t for a two-tailed 95% confidence interval,
and $s_{\hat{y}}^2$ is the standard error of the mean prediction.

Figure 3.4 shows two regression models which have been fitted to the data presented in Table 3.1 along with the confidence intervals for fitted y. It is obvious from the graph that these intervals are not parallel, but are in fact curved. This is because fitted y can be more accurately predicted when x is close to its mean value. The size of the confidence interval depends on the distance of x from $\bar{x}$. The confidence intervals for fitted y given a number of different values of the variables x_1 and x_2 are provided in Table 3.3.

Figure 3.4: Confidence and prediction intervals for two regression models



3.1.4 Prediction Intervals

Equation 3.5 can be extended to allow probability statements to be made about the likely values of *individual* cases of the response variable given a certain value of the explanatory variable. This interval is not called a *confidence interval* since *y* is a random variable and not a parameter (Aitkin, Anderson, Francis & Hinde, 1989). The equation for this prediction interval has an extra term since the additional variation in the random variable needs to be allowed for. Ninety-five percent two-tailed prediction intervals for large samples can be calculated using Equation 3.6:

$$\text{prediction interval for } y = \hat{y} \pm 1.96\sqrt{s^2 + s_{\hat{y}}^2} \qquad (3.6)$$

where s^2 is the standard error of the estimate,

and $s_{\hat{y}}^2$ is the standard error of the mean prediction.

Figure 3.4 shows the prediction intervals calculated for two regression models. Similar to the confidence intervals for fitted *y*, these intervals are curved with values of *y* predicted more accurately when *x* is close to its mean value. The manual calculation of the expression $\sqrt{s^2 + s_{\hat{y}}^2}$ is quite involved and will not, therefore, be demonstrated here (see Ryan, 1997 and Norušis, 1994, for examples of how prediction intervals can be calculated manually). It is, however, unnecessary to manually compute prediction intervals for *y* as these can be generated using statistical software (see Section 3.6). Prediction intervals for *Y*, given a number of different values of *X* (some of which are included in the data in Table 3.1, and some of which are not), are provided in Table 3.3.

Table 3.3: Confidence and Prediction Intervals for Two Regression Models

	Fitted	95% CIs		95% PIs	
Value of X	Value of Y	Upper	Lower	Upper	Lower
Model 1					
10.50	0.162	−0.666	0.989	−1.646	1.969
18.34	4.567	4.150	4.982	2.907	6.226
31.87	12.168	9.941	12.070	9.078	12.933
Model 2					
7.30	−1.638	−4.249	0.972	−5.955	2.678
18.34	4.561	3.671	5.451	1.009	8.113
24.75	8.161	6.572	9.750	4.373	11.948

Model 1: $Y = \alpha + \beta X_1$

Model 2: $Y = \alpha + \beta X_2$

The interpretation of the intervals shown in Table 3.3 is similar to the interpretation of the confidence intervals for β which were discussed earlier. For example, when X_1 has a value of 10.5, the mean value of *Y* is predicted to lie between −0.666 and 0.989, and an individual case between −1.646 and 1.969. When X_1_and X_2 are both 18.34, the fitted values for *Y* are similar (the regression equations are nearly identical) but there are differences in the confidence and prediction intervals for the two models. Model 1 predicts a narrower range of values for *Y* and suggests that X_1 is a more useful variable than X_2 for predicting mean and

individual values of *Y*. From Table 3.3 it is easy to see that the prediction intervals are curved, as the optimal predictions are achieved when *x* is close to its mean value (the difference between the upper and lower prediction intervals are smaller for less extreme values of *x*). It is also clear that the confidence intervals for the mean value of *y* are narrower than the prediction intervals for individual values of *y*.

To summarize, one uses *confidence intervals* for fitted *y* when making statements about the predicted mean value of the response variable in further samples, whilst *prediction intervals* are used when predicting the value of a single case of the response variable.

3.1.5 Goodness-of-fit Measures

Confidence and prediction intervals provide an indication of the usefulness of a regression model, but do not provide an easily interpretable goodness-of-fit measure (that is, how well the regression model fits the data). Two statistics are discussed here, R^2, which provides some descriptive information about the model fit, and the *F* statistic, which provides a measure of significance.

The R^2 Statistic

R^2 provides descriptive information about the model fit and is calculated using Equation 3.7 which compares observed values of *y* with those predicted from the model. R^2 for a simple regression, where there is a single explanatory variable, is known as the coefficient of determination.

$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} \qquad (3.7)$$

where y is the observed value of y,
$\hat{y}$ is the value of y predicted from the model,
and $\bar{y}$ is the mean value of y.

In Equation 3.7, if $\hat{y}$ and *y* are the same (that is, the model perfectly predicts *y*), the numerator and the denominator assume identical values and $R^2 = 1$. If, on the other hand, the model provides no clues as to the value of *y*, then $\hat{y} = \bar{y}$ (that is, the predicted value of *y* remains constant and the explanatory variable therefore plays no part in determining the value of the response variable) and the expression $(\hat{y} - \bar{y})^2 = 0$ which in turn leads to an R^2 value of O. Values of R^2 therefore range from 0, which indicates no linear relationship, to 1, which indicates a perfect linear relationship.

R^2 is commonly interpreted as the percentage of the variability in *y* that is explained by *x* when it is used to predict *y*. In general terms, it provides an indication of how well the model fits the data. For simple regression,

R^2 indicates the strength of the linear relationship between *x* and *y*. For example, if a simple regression model has an R^2 value of 0.748 one can conclude that 74.8% of the variability in the response variable is accounted for by the explanatory variable. It should be noted, however, that although R^2 is widely used and accepted as a measure of model fit, it has a tendency to increase as the slope of the regression line increases and is not, therefore, a completely unbiased measure (see Barrett, 1974). Even with this limitation, R^2 is a useful statistic and is used extensively in this chapter to provide an easily understood estimate of model fit.

The *F* Statistic

Whilst R^2 provides an indication of the explanatory power of a model, it does not indicate the level of significance (that is, how likely it was that the R^2 value had been obtained by chance). To do this we need to test the hypothesis that the regression coefficient, β, equals zero. *A* test of this hypothesis is provided by the *F* test which can be calculated using the R^2 statistic, or, more appropriately, directly from the measure of deviance. Equation 3.8 (see Afifi and Clark, 1996 and Berry and Feldman, 1993) shows how *F* can be calculated from R^2.

$$F_{k,n-k-1} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \qquad (3.8)$$

where R^2 is the coefficient of determination,
n is the number of cases used to construct the model,
and k is the number of terms in the model (not including the constant).

If the value of *F* is not significant, the null hypothesis of no linear relationship between *x* and *y* is accepted. If, on the other hand, the value of *F* is significant the null hypothesis is rejected and the hypothesis that there is a significant linear relationship between *x* and *y* is accepted. You will note that for the simple case, the *F* test is equivalent to the *t* test since both evaluate the relationship $\beta = 0$. In fact, $\sqrt{F} = t$.

A more useful method of testing the hypothesis $\beta = 0$, is to derive the value of *F* from the residual sum of squares statistic (RSS), which is the deviance for a GLM with an identity link. RSS is the sum of the squared differences between the observed and predicted values of *y*, i.e., $(\sum(y - \hat{y})^2)$, and gives a measure of how much the observed data differ from predictions from the model. Equation 3.9 overleaf shows the general formula for calculating the value of *F* from the deviance measure RSS (see Afifi and Clark, 1996 and Francis, Greene and Payne, 1994).

This equation compares the amount of deviance in the null model ($y = \alpha$) with the amount of deviance in the model $y = \alpha + \beta x$. If the explanatory

$$F_{(df_{\text{null}}-df_{\text{model}}),df_{\text{model}}} = \frac{\text{RSS}_{\text{null}} - \text{RSS}_{\text{model}}}{(df_{\text{null}} - df_{\text{model}})(\text{RSS}_{\text{model}}/df_{\text{model}})} \quad (3.9)$$

where 'null' indicates the model $y = \alpha$,
'model' indicates the model $y = \alpha + \beta x$,
RSS is the residual sum of squares for the designated model,
and df is the degrees-of-freedom for the designated model.

variable does not enable a significantly better prediction to be made, RSS_{null} and $\text{RSS}_{\text{model}}$ will have similar values and the value of *F* will tend to be small. If, on the other hand, the explanatory variable has a large effect on the model, the expression RSS_{null} — $\text{RSS}_{\text{model}}$ will assume a relatively large value and *F* will also tend to be large. Although Equations 3.8 and 3.9 can both be used to calculate *F*, the latter is more useful as it is more general and enables not only the significance of the entire model to be determined, but through the computation of a *partial-F* statistic, enables the significance of individual and groups of variables to be assessed. This is particularly useful in multiple OLS regression and is discussed further in Section 3.3.5.

Table 3.4: Goodness-of-fit Measures for Two Regression Models

	Coefficient	R^2	t	F	P
Model 1					
X_1	0.562	0.936	13.813	190.80	0.000
Model 2					
X_2	0.562	0.708	5.612	31.34	0.000

Model 1: $Y = \alpha + \beta X_1$
Model 2: $Y = \alpha + \beta X_2$

Table 3.4 shows a number of statistics which assess the model fit of the two regression models computed for the 15 cases in Table 3.1. These values have been calculated from software, but could also have been calculated manually. For example, using Equation 3.8 the value of *F* for model 1 can be calculated from R^2 as 0.936/[0.064/(15–1–1)]' which equals 190.134 (allowing for rounding error, this is the same result as that obtained using software). *F* can also be derived using Equation 3.9. Given RSS_{null} = 112.715 and $\text{RSS}_{\text{model}}$ = 7.190, the value of *F* is equal to (112.715 − 7.190)/(7.190/13), which equals 190.796 (the calculation of *F* from RSS statistics is demonstrated in Section 3.6). Given that the general form of equation for calculating *F* is more useful (and also easier to understand), this book will standardize on the use of Equation 3.9 to calculate the *F* statistic.

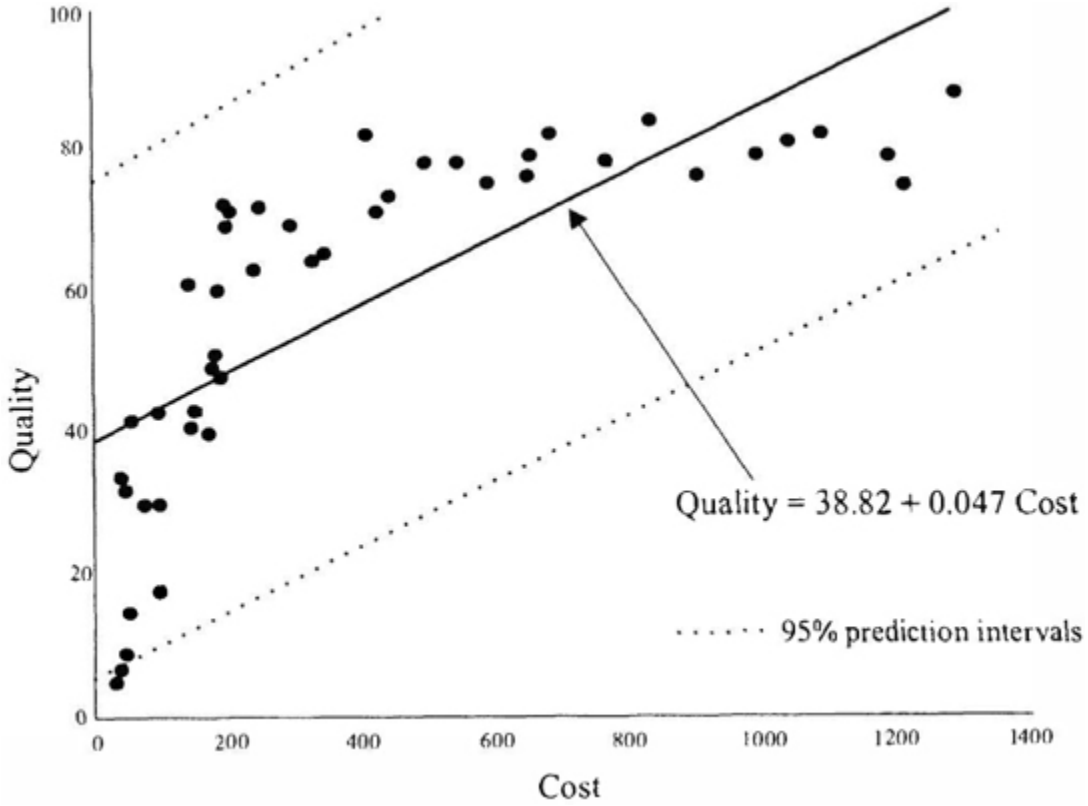
3.2 A Worked Example of Simple Regression

Table 3.5: The Cost and Sound Quality of Music Systems

Cost (£)	Quality	Cost (£)	Quality	Cost (£)	Quality
35	5	180	49	502	78
39	7	186	51	550	78
40	34	190	60	595	75
48	9	195	48	655	76
50	32	200	72	660	79
55	15	205	69	690	82
58	42	210	71	775	78
79	30	246	63	842	84
97	43	254	72	910	76
100	30	300	69	1000	79
100	18	335	64	1050	81
149	41	350	65	1099	82
150	61	415	82	1200	79
156	43	430	71	1225	75
175	40	448	73	1300	88

Table 3.5 contains hypothetical data showing the cost (£ sterling) and sound quality of 45 different music systems of varied price. Sound quality is recorded on a continuous scale and represents scores provided by a group of raters (the higher the score the better the sound reproduction).

Figure 3.5: Relationship between rated quality and cost of music systems

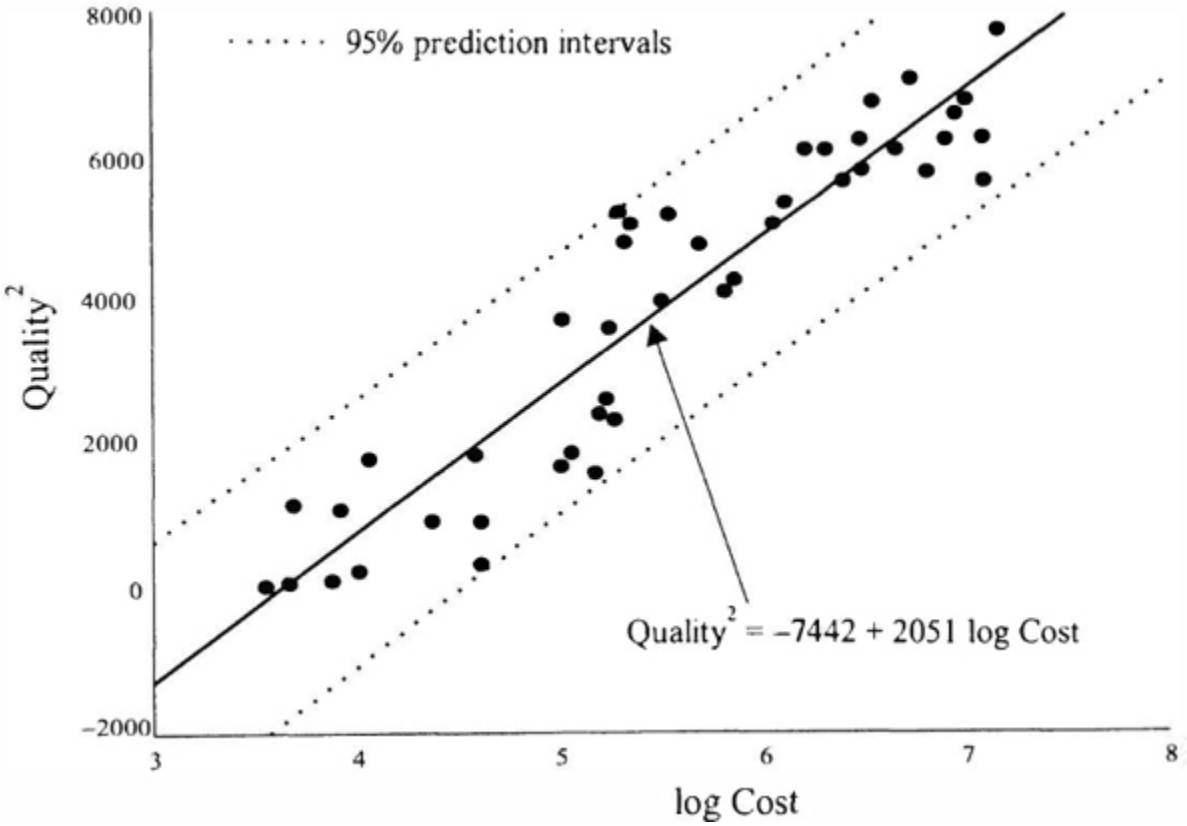


The scatter plot in Figure 3.5 displays the relationship between Cost and Quality for the music systems included in the study and shows that this relationship does not appear to be linear. There is a relatively large increase in sound quality between systems costing £50 to £300, but after this, for a given increase in the cost of the system, the increase in sound quality is much less dramatic. There appears to be a relationship between Cost and Quality which conforms broadly to a 'law of diminishing returns'. A simple OLS regression analysis of the data in Table 3.5 gives the model shown in Equation 3.10.

$$\text{Quality} = 38.820 + 0.047\,\text{Cost} \qquad (3.10)$$

Quality and Cost are significantly linearly related ($F_{1,43} = 53.25$; $P < 0.0005$) with an R^2 value of 0.553. Even though this is a highly significant result, the regression line does not appear to describe the relationship between Quality and Cost particularly well. It is obvious that a non-linear relationship is being described using a linear model (see Figure 3.5). In an attempt to improve this model, transformations were applied to the variables using some of the techniques described in Chapter 2. Cost was transformed using a natural log, whilst Quality was squared (see Equation 3.11). These transformations were chosen solely for the purpose of improving the model fit and resulted in a relationship between the variables which more closely approximated a straight-line (see Figure 3.6).

Figure 3.6: Relationship between Quality^2 and log Cost



$$\text{Quality}^2 = -7442.02 + 2051.20 \log \text{Cost} \qquad (3.11)$$

Figure 3.7: The transformed model

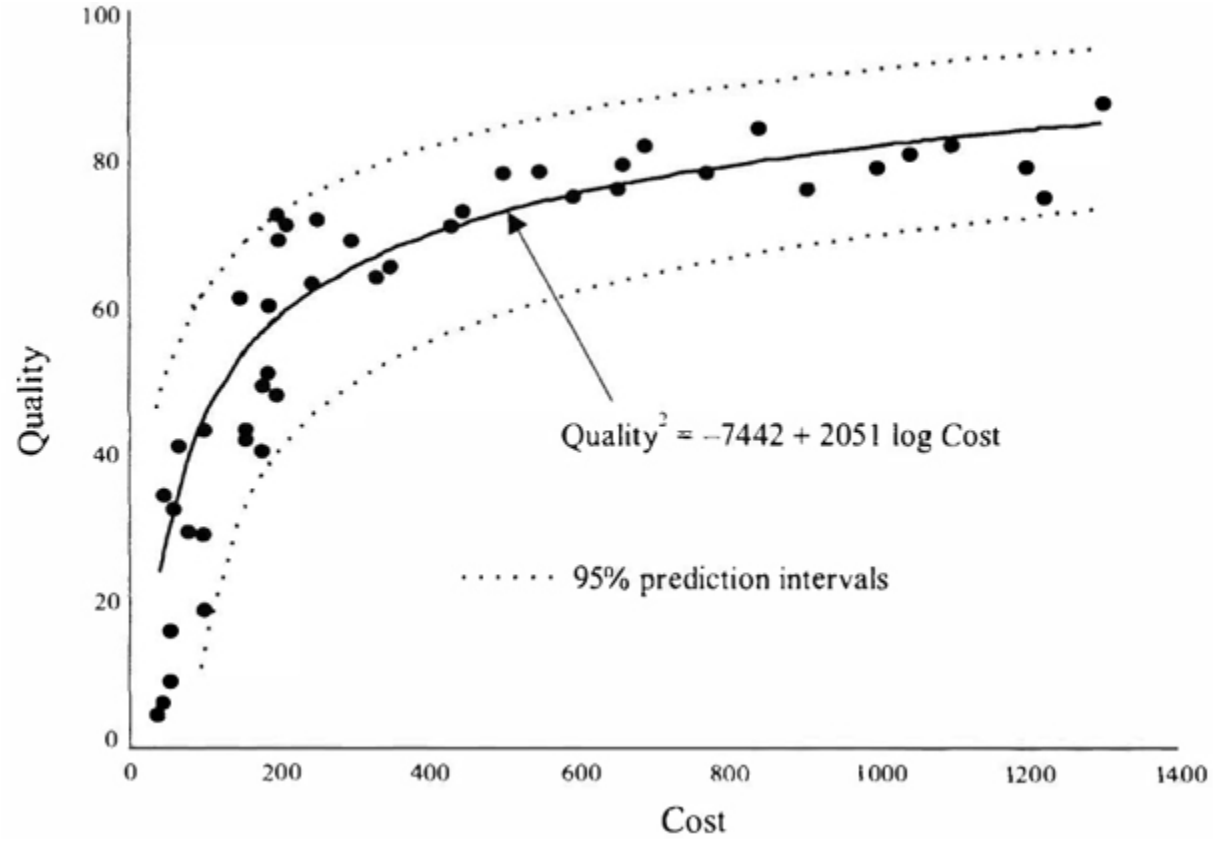


Table 3.6 shows clearly that the transformed model provides a higher value of R^2 and a more significant linear relationship between the variables than does the untransformed model. Redrawing Figure 3.6 using the original axes (Cost and Quality) demonstrates clearly that the transformed model provides a closer approximation to the data.

Table 3.6: Two Models Predicting Sound Quality

System Cost	Predicted Sound Quality	95% 2-tailed PIs		PI Range
		Lower	Upper	
Model 1				
£100	43.53	10.72	76.33	65.61
£385	56.94	24.39	89.49	65.10
£1200	95.30	61.18	129.43	68.25
Model 2				
£100	44.77	13.49	61.86	48.37
£385	69.06	54.40	81.11	26.71
£1200	84.27	72.48	94.60	22.12

Model 1: $\text{Quality} = 38.82 + 0.047\,\text{Cost}$
$F_{1,43} = 53.25,\ P < 0.0005,\ R^2 = 0.55$

Model 2: $\text{Quality}^2 = -7442 + 2051.2\,\log\text{Cost}$
$F_{1,43} = 266.96,\ P < 0.0005,\ R^2 = 0.86$

The prediction intervals for the transformed model are narrower making it a more useful model for predicting sound quality. The differences between the two models are shown in Table 3.6 which shows estimates for the sound quality of systems priced at £100, £385 and £1200 (all within the range of those sampled) which have been calculated using the transformed and untransformed models described above. Of particular note are the narrower prediction intervals associated with the transformed model which enables more accurate predictions to be made.

3.3 Multiple OLS Regression

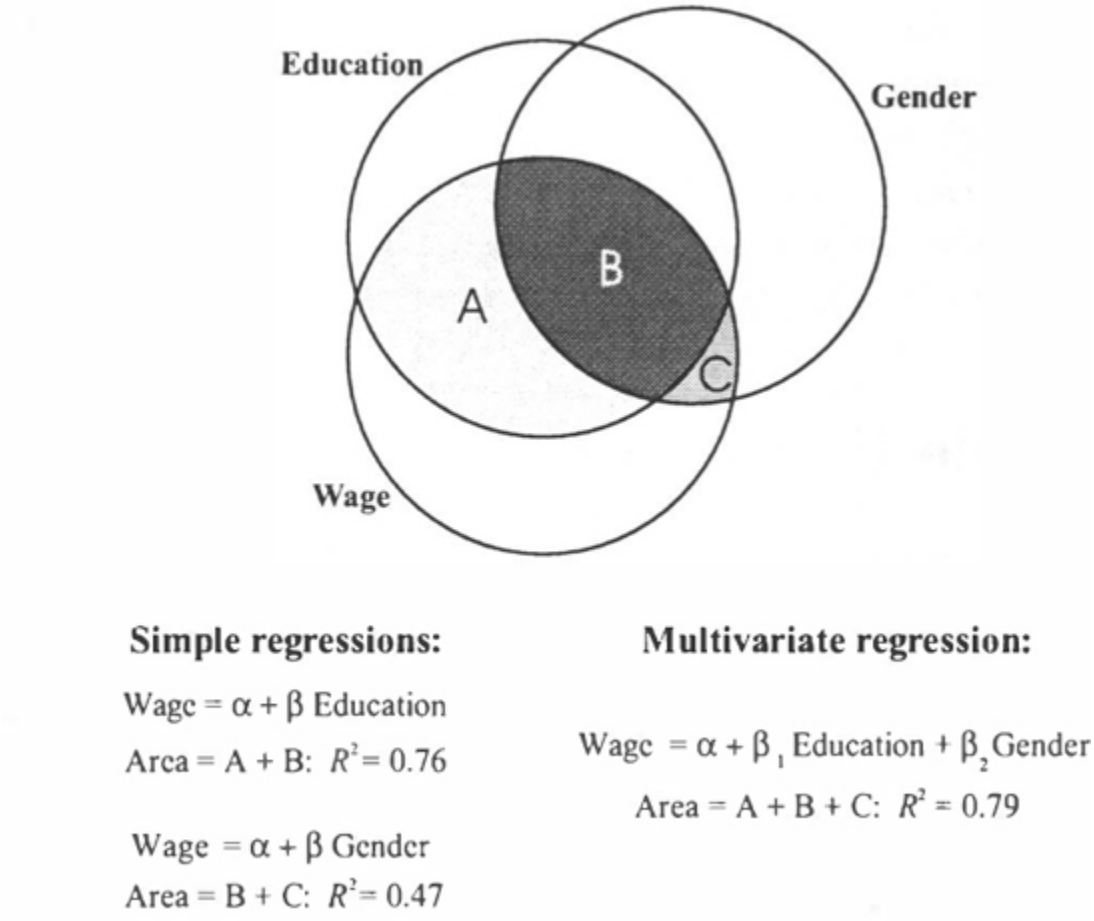
Multiple regression is a technique which can be used to investigate the relationship between a response variable and more than one explanatory variable. As with simple regression, it can be used to both identify significant relationships (explanation) and predict values of the response variable (prediction). The ability to analyse the effect of multiple variables is particularly useful in the social sciences as it is usually the case that more than one source of information is required to make adequate predictions. For example, wage may be dependent upon a number of factors including gender, education, ethnicity and experience, all of which have to be accounted for if wage is to be successfully modelled.

Although it is possible to investigate the relationship between a response variable and a number of explanatory variables using multiple simple regressions, this method is not appropriate if the explanatory

variables are interrelated. Multiple simple regressions do not take into account relationships between explanatory variables and as a result can provide a misleading picture of the data. This can be illustrated with the help of a hypothetical example where a particular company employs on the basis of educational achievement and not on the basis of gender[2]. If this company recruits in a region where males and females do not have equal access to education, it is likely that a simple regression model predicting wage from gender will show a significant relationship between the two. In this case, the relationship is not due to discrimination by the company on the basis of gender, but is due to a bias in the provision of education which results in greater access to schools and colleges for males. Males tend to be more highly educated and, consequently, better paid. The relationship between gender and wage is a consequence of the relationship between gender and education; a relationship which is not accounted for using simple regression analyses.

Figure 3.8 shows a graphical representation of the relationship between wage, education and gender. These variables are represented by circles with overlapping areas indicating the degree of correlation (this can be viewed as equivalent to the R^2 measure). The regression model Wage = α + β Education, has a model fit equivalent to areas A and B, and the regression model Wage = α + β Gender has a model fit equivalent to areas B and C. Since the $R2$ values of both models include the area B, we cannot simply combine them to provide a figure for the overall R^2 when wage is predicted using both explanatory variables. For example, combining the two R^2 values for the simple regression models gives an overall R^2 of 1.23 (equivalent to areas A+B+B+C), which suggests that over 100% of the variance in the response variable can be explained by the two explanatory variables. The extent to which wages can be predicted by both explanatory variables needs to be determined using a multiple regression model which takes account of both variables simultaneously and provides an overall R^2 value equivalent to the area A+B+C (which is equal to 0.79 in the example above).

Figure 3.8: Relationship between wage levels, education, and gender



Using multiple regression, one can calculate the effect that each explanatory variable has on the response variable whilst controlling for other variables in the model. In. the example above, the effect of education on wage is calculated whilst gender is held constant and the effect of gender on wage is calculated whilst education is held constant. The regression coefficients indicate the unique contribution made by each explanatory variable. In this case, the unique contributions of education and gender to predicting wage correspond to areas A and C respectively. It is clear from Figure 3.8 that gender has a relatively small effect on wage once its relationship with education has been taken into account.

3.3.1 The Regression Equation

Multiple OLS regression aims to arrive at an equation which describes the relationship between a continuous response variable and a number of explanatory variables. The multiple regression equation is almost identical to the simple regression equation except that it accounts for more than one explanatory variable. The general form of the equation is:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k \qquad (3.12)$$

where α is the average value of y when each $x = 0$,
x_1 to x_k are the values of k different explanatory variables,
and β_1 to β_k are partial regression coefficients.

In Equation 3.12, (*β* represents the average change in *y* which is expected to result from a change of one unit in *xwhen all other variables are held constant*. For example, if β_4 equals 1.98, *y* increases by an average of 1.98 each time x_4 increases by *1*. Similarly, if β_2 equals −37.1, *y* decreases by 37.1 for each unit increase in x_2. The partial regression coefficient (3 identifies the effect that *x* has on *y* independent of other variables in the model (that is, it identifies the unique contribution made by *x* in determining *y*). It should be noted that partial regression coefficients, β_k, are dependent upon the units in which variables are measured and cannot therefore be used directly to compare the relative importance of terms in a model — one must use confidence intervals for the coefficients, or significance tests for this purpose.

Interactions and Curvilinearity

Equation 3.12 assumes that the effects of the explanatory variables are additive. That is, the response variable is determined by so much of x_1, plus so much of x_2, plus so much of x_3, etc. This regression model assumes that the effect of any single explanatory variable remains the same across the range of other explanatory variables. The model, as it stands in equation 3.12, does not account for interactions between explanatory variables. It is, however, relatively easy to include interaction terms by including terms which are the product of the explanatory variables that are interacting (for example, an interaction between x_1 and x_2 can be represented in a model by including the term $x_1 \times x_2$). Equation 3.13 shows a regression model containing terms for two explanatory variables and an interaction between them. The ‘main effect’ of each variable is given by the terms $\beta_1 x_1$ and $\beta_2 x_2$, whilst the interaction between them is represented by the term $\beta_3 x_1 x_2$.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \qquad (3.13)$$

The regression coefficient β_3 indicates the change in *y* which is expected to occur as a result of a unit change in, $x_1 \times x_2$, whilst controlling for the other terms in the model. Although this definition is quite straightforward, it can be difficult to meaningfully interpret interactions (*y* is dependent on the product of X_1 and x_2). Interpreting interactions generally becomes more difficult as more explanatory variables are included. Interactions can contain as many terms as there are explanatory variables in the model. For example, a regression model containing five explanatory variables (x_1 to x_5) can contain an interaction term which is the product of all five variables ($x_1 \times x_2 \times x_3 \times x_4 \times x_5$). In addition to this 5-way interaction term, the model can also contain a number of 4-way, 3-way and 2-way interactions. The number of interactions which can be included increases greatly with the number of explanatory variables and consequently makes the regression model considerably more complex.

It is important to check interactions and include in the model those which are significant, even if interpretation proves difficult (in practice, however, many interactions tend to be insignificant and can be left out of the model; see Lewis-Beck, 1993). Significant interactions need to be included as these affect the parameters which are calculated for the other terms in the model. For example, a regression model with two explanatory variables x_1 and x_2, shows a significant interaction. This interaction needs to be included in the model (by including the term $x_1 \times x_2$) as it has consequences for the parameters calculated for x_1 and x_2. In this case, to appropriately interpret the effect that x_1 or x^2 have on the response variable, the interaction term needs to be included. It should be noted that the appropriate coding and interpretation of these terms is quite complex and a full discussion is beyond the scope of this chapter. Detailed accounts of how to code and interpret interactions can be found in Chapter 5 as well as a number of other texts (see, for example, Lewis-Beck, 1993; Hardy, 1993 and Jaccard, Turrisi and Wan, 1990). A worked example of a multiple regression with an interaction term is provided in Section 3.4.

The usefulness of OLS regression as a modelling tool is further enhanced since, in addition to interaction terms, regression models can include terms for non-linear and quadratic relationships. These relationships can be dealt with in much the same way as interactions, with additional terms (for example, x^2 and x^3) being added to the model (see Chapter 2 for a discussion of these types of relationships).

3.3.2 Confidence Intervals for?

As with simple regression, confidence intervals can be calculated to indicate the range within which the partial regression coefficients are expected to fall a certain proportion of the time (for example, 95% or 99% of the time). Ninety-five percent two-tailed confidence intervals for β can be calculated using Equation 3.3, which, for large samples can be simplified to Equation 3.14.

$$\text{Confidence interval for } \beta = \hat{\beta} \pm 1.96(\text{s.e. } \hat{\beta}) \qquad (3.14)$$

The confidence intervals associated with partial regression coefficients are interpreted in much the same way as the confidence intervals for simple regression coefficients which were discussed in Section 3.1.2. The only difference is that other terms in the model are controlled for. A partial regression coefficient of zero indicates that there is no linear relationship between that regressor and the response variable after controlling for all other terms in the model. Similarly, a confidence interval which includes zero indicates that there is no significant linear relationship between that term and the response variable. The significance of the relationship between individual partial regression coefficients and the response variable can be formally tested using the t statistic (with $n - k - 1$ degrees of freedom) to evaluate the hypothesis $\beta = 0$. If β is significantly different to zero, it can be concluded that there is a linear relationship between the explanatory and the response variable[3].

3.3.3 Confidence Intervals for Fitted *y*

Confidence intervals for the mean value of fitted *y* can be calculated in an identical way to that used in simple regression. Ninety-five percent two-tailed confidence intervals for a large sample can be derived using Equation 3.15.

$$\text{confidence interval for fitted } y = \hat{y} \pm 1.96\sqrt{s^2_{\hat{y}}} \qquad (3.15)$$

3.3.4 Prediction Intervals

The technique for calculating prediction intervals in multiple regression is identical to that used in simple regression and for large samples can be derived using Equation 3.16.

$$\text{prediction interval for } y = \hat{y} \pm 1.96\sqrt{s^2 + s^2_{\hat{y}}}. \qquad (3.16)$$

As with simple OLS regression, the size of the confidence and prediction intervals depend on the distance of x_k from $\bar{x}_k$. The more extreme the values of x_k used to predict individual or mean values of *y*, the larger the intervals will be. Similar to OLS regression, the confidence intervals for the fitted values of the mean of *y* are narrower than the prediction intervals for individual values of *y*. A demonstration of the calculation and use of prediction intervals in a multiple regression model is shown in Section 3.4.

3.3.5 Goodness-of-fit Measures

The goodness-of-fit of multiple regression models can be assessed using the R^2 and *F* statistics in much the same way as in simple regression.

The R^2 and R^2 Statistics

The goodness-of-fit of a multiple regression model can be indicated by the R^2 statistic which shows the proportion of the response variable that can be explained by *all* terms (or regressors) included in the model (see Equation 3.7). The R^2 statistic calculated for a model containing multiple terms is commonly known as the coefficient of multiple determination and is widely used. However, it is not an ideal indicator of model fit since each term introduced into a model increases the value of R^2 even if it has no influence on y, as each term will 'account for' at least one case. R^2 therefore provides an optimistic measure of model fit which tends to 1.0 as the number of terms included in the model increases relative to the number of cases. When there are as many terms in the model as there are cases, R^2 will always equal 1.0, no matter what the relationship between the response and explanatory variables. One solution to this problem is to calculate an *adjusted*R^2 statistic (R^2_a) which takes into account the number of terms entered into the model and does not necessarily increase as more terms are added. Adjusted R^2 can be derived using equation 3.17.

$$R_a^2 = R^2 - \frac{k(1 - R^2)}{n - k - 1} \quad (3.17)$$

where R^2 is the coefficient of multiple determination,
n is the number of cases used to construct the model,
and k is the number of terms in the model (not including the constant).

In a similar way to R^2, R_a^2 only provides a 'rough indication' of the model fit. In a discussion of the use of R_a^2, Draper & Smith (1981, page 92) conclude that it "might be useful as an initial gross indicator, but that is all". Given that neither statistic provides a 'perfect' measure of model fit, this book will use the more widely adopted R^2 statistic.

The F and Partial-F Statistics

An *F* test can be used to formally test the null hypothesis that there is no linear relationship between the response and all of the explanatory variables in the model (see Equation 3.18).

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 =, \ldots, = \beta_k = 0 \quad (3.18)$$

where β_1 to β_k are partial regression coefficients,
and k is the number of terms in the model.

In multiple regression, the value of *F* can be calculated using Equation 3.9 in exactly the same way as was demonstrated for simple regression and provides a measure of overall model fit. This equation is, however, of limited use in multiple regression as it does not enable the significance of individual terms in the model to be assessed. Equation 3.19 shows a modification to Equation 3.9 that allows nested models (a nested model is one which is a subset of another) to be compared through the calculation of a partial-F statistic.

$$F_{(df_{\mathrm{p}} - df_{\mathrm{p+q}}), df_{\mathrm{p+q}}} = \frac{\mathrm{RSS_p} - \mathrm{RSS_{p+q}}}{(df_{\mathrm{p}} - df_{\mathrm{p+q}})(\mathrm{RSS_{p+q}}/df_{\mathrm{p+q}})} \quad (3.19)$$

where 'p' indicates the smaller model,
'p+q' indicates the larger model,
RSS is the residual sum of squares for the designated model,
and *df* is the degrees-of-freedom for the designated model.

The partial-*F* statistic is important in OLS regression as it allows the significance of individual and groups of terms to be determined. For example, the unique effect that variable x_2 has on the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2$ can be determined by comparing this model with the model $y = \alpha + \beta_1 x_1$. The difference in the model

fit between the two, as determined by the partial-F statistic, indicates the effect that variable x_2 has on the response variable. Similarly, the effect that a group of terms has on the model fit can be determined by comparing a model which contains the group with a nested model which does not. For example, if a model containing three explanatory variables (*a, b* and *c*) shows a significant three-way interaction, the regression model predicting *y* can be represented as:

$$y = \alpha + \beta_1 a + \beta_2 b + \beta_3 c + \beta_4 ab + \beta_5 ac + \beta_6 bc + \beta_7 abc.$$

If the terms containing variable *α* are not highly significant, and are not theoretically crucial, it might be useful to remove all these parameters from the model (i.e., $\beta_1 a$, $\beta_4 ab$, $\beta_5 ac$ and $\beta_7 abc$). The effect on the model fit of removing all parameters containing variable *a* can be determined using a partial-*F* test computed using the full model above and the nested model:

$$y = \alpha + \beta_1 b + \beta_2 c + \beta_3 bc.$$

The resulting partial-*F* value shows the unique effect that all four terms considered together have on the model fit.

The partial-*F* statistic is used extensively in model selection and allows the process of model-building to be greatly simplified and accelerated. As it allows nested models to be compared, groups of terms can be assessed for significance which allows the significance of some or all terms which relate to one particular variable to be derived as well as the effect of, say, all two-way or three-way interactions (see Section 3.6 for an example of this in GLIM). In addition to this, partial-F allows the significance of categorical variables, which are represented as a number of separate terms, to be appropriately assessed. Given its importance in OLS regression, the use of the partial-F statistic is discussed in detail in Section 3.3.8 and demonstrated in Sections 3.4 and 3.6.

3.3.6 Multicollinearity

The technique of multiple regression allows more than one explanatory variable to be entered into a model. However, there are some considerations concerning which variables may be entered. Perhaps the most important of these is *multicollinearity*, a term used here to describe a situation where an explanatory variable is related to one or more of the other explanatory variables in the model. If these relationships are perfect or very strong, the calculation of the regression model and the appropriate interpretation of the results can be affected. In the case where one explanatory variable can be precisely predicted from one or more of the other explanatory variables (perfect multicollinearity), the analysis fails as a regression equation cannot even be formulated. When a relationship is strong, but not perfect (high multicollinearity), the regression equation can be formulated, but the parameters may be unreliable. Parameters which are unreliable can change dramatically as a result of relatively minor changes in the data set with the addition or deletion of a small number of observations exerting a large influence on the regression equation and, subsequently, on the interpretation of the results.

The consequences of multicollinearity depend, to some degree, on the objectives of the analysis. If the goal is prediction, then multicollinearity need not present much of a problem, as it primarily affects the calculated importance of the explanatory variables. Even though the interpretation of the regression coefficients associated with the explanatory variables may be suspect, the response variable may still be able to be accurately predicted. However, if the goal is explanation (that is, the aim is to identify the strength of relationships between individual explanatory variables and the response variable), the presence of a high degree of multicollinearity poses a serious problem for the correct interpretation of the results. When conducting a multiple regression, one has to identify when multicollinearity is likely to present a problem and a strategy to deal with it must be decided upon.

Perfect Multicollinearity

Perfect multicollinearity occurs when an explanatory variable can be precisely predicted from other explanatory variables in the model. When this happens, the variable contributes no unique information to the model and is therefore redundant. The inclusion of one or more redundant explanatory variables in a regression model is problematic as it is not possible to determine the parameters associated with these variables and, consequently, a regression equation cannot even be formulated. This problem can be demonstrated by looking at a three-variable relationship, which can be represented algebraically as a plane $y = \alpha + \beta_1 x + \beta_2 z$ (the three variables are y, x and z). If one of these explanatory variables is redundant (say, $x = 5z$) then y can be described in terms of a single variable (x or z), which is represented algebraically as a line ($y = \alpha + \beta x$). The regression procedure attempts to calculate parameters for x and z, but since there is merely information about one of them, only one coefficient can be computed and the regression technique breaks down (for a detailed discussion of this see Berry and Feldman, 1993 and Maddala, 1992). In essence, when there is perfect multicollinearity, the regression parameters cannot be formulated as the multiple regression procedure attempts to fit an equation which has more dimensions than are present in the data.

In practice, perfect multicollinearity is not usually a problem as it is quite rare and can be readily detected. In fact, many statistical analysis packages automatically alert the user to its presence. A more serious problem for the analyst is the presence of high multicollinearity where a regression model can be formulated, but the parameters associated with some of the explanatory variables may be unreliable.

High Multicollinearity

When explanatory variables are highly (but not perfectly) related to one or more of the other explanatory variables in the model it becomes difficult to disentangle the separate effect of each variable. As a variable which shows a high degree of multicollinearity provides little unique information, the regression coefficient associated with it is also based on limited information and therefore tends to have a large standard error (for detailed discussions on this refer to Afifi and Clark, 1996 and Edwards, 1985). In such cases the regression parameters are unlikely to accurately reflect the impact that x_i has on y in the population.

The problems associated with high multicollinearity can be demonstrated using hypothetical data which

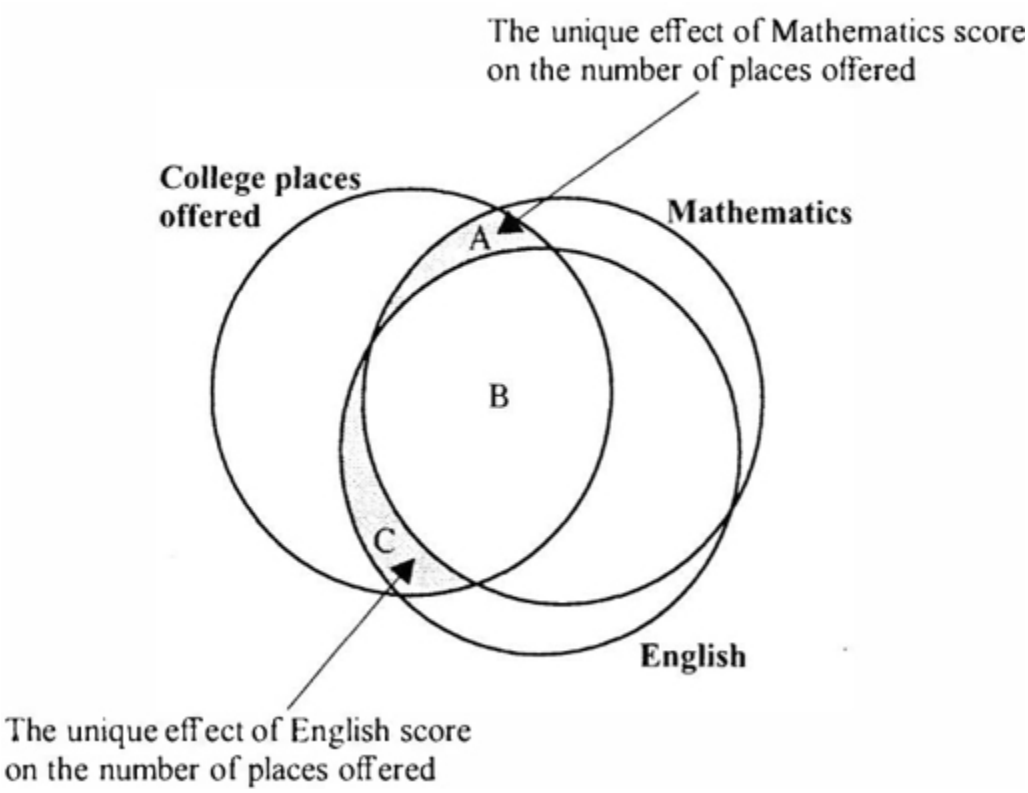
shows the relationship between the number of college places offered to students and marks obtained in two compulsory subjects, English and Mathematics (see Table 3.7)[4].

Table 3.7: Exam Marks and Offers of College Places

Number of Colleges Offering Places	English (%)	Mathematics (%)
0	22.0	17.0
1	32.5	34.5
2	38.0	18.0
3	39.5	46.0
4	52.0	48.0
5	44.5	35.0
6	49.5	43.5
7	72.5	70.5
8	61.5	67.5
9	85.5	74.5
10	68.5	87.6

One would expect there to be a strong relationship between the number of college places a student is offered and the student's marks in English and Mathematics as the decision to offer a place at a college is based largely on the student's academic performance. One would also expect a student's mark in one subject to be strongly related to their mark in the other subject, as good students tend to score relatively highly in both. This three-variable relationship is shown in Figure 3.9 and the associated regression model in Equation 3.20 overleaf (for the simple purpose of demonstration this model does not include an interaction term). The model appears to provide a good prediction of the number of college places offered to a student as indicated by the F and R^2 statistics ($F_{2.8} = 32.520$; $\mathbf{P} < 0.0005$: $R^2 = 0.890$) which corresponds to area A + B + C, in Figure 3.9.

Figure 3.9: The relationship between exam marks and the number of college places offered



College places offered = −2.79 + 0.086 English + 0.068 Mathematics (3.20)

From Figure 3.9 it can be seen that the *unique* contributions made by each of the explanatory variables to the number of college places offered is relatively small. When controlling for marks in Mathematics, marks in English only contribute a small amount to the model fit (area C). Similarly, when controlling for marks in English, marks in Mathematics only contribute a small amount to the model fit (area A). The results in Table 3.8 confirm this and show that the unique contribution of each of the explanatory variables when they are both entered into the model (Model l) is not significant, as shown by the *t* statistics. It appears clear from the *F* and R^2 statistics that Model l provides a good fit, even though neither of the explanatory variables are significant. This, perhaps unexpected, result is due to the high degree of multicollinearity between the explanatory variables. Logically we might expect marks in both English and Mathematics to be strongly related to the number of college places offered as places are offered mainly on the basis of academic performance. This is what we find when simple regression models are calculated using single subjects to predict college places (Models 2 and 3). The resulting models fit almost as well as the model which uses both variables, but the regression parameters for the explanatory variables are now highly significant. We can see that the presence of multicollinearity has not really affected the predictive power of the model, but has serious implications for the interpretation of the importance of the explanatory variables.

Table 3.8: Modelling the Number of College Places Offered to Students

	Coefficient	s.e.	t	P
Model 1				
English	0.086	0.046	1.863	0.100
Mathematics	0.068	0.038	1.805	0.109
(constant)	−2.790	1.148	−2.431	0.041
Model 2				
English	0.161	0.023	7.028	0.000
(constant)	−3.276	1.248	−2.625	0.028
Model 3				
Mathematics	0.131	0.019	6.951	0.000
(constant)	−1.473	1.021	−1.443	0.183

Model 1: Places offered $= \alpha + \beta$ English $+ \beta$ Mathematics
$F_{2,8} = 32.52,\ P < 0.0005,\ R^2 = 0.089$

Model 2: Places offered $= \alpha + \beta$ English
$F_{1,9} = 49.394,\ P < 0.0005,\ R^2 = 0.846$

Model 3: Places offered $= \alpha + \beta$ Mathematics
$F_{1,9} = 48.313,\ P < 0.0005,\ R^2 = 0.843$

Identifying Instances of Multicollinearity

Some instances of multicollinearity can be identified by inspecting pair-wise correlation coefficients. Relationships between explanatory variables which are of the order of about 0.8 or larger indicate a level of multicollinearity that may prove to be problematic[5]. In the above example, the correlation between English and Mathematics scores is 0.897 indicating that multicollinearity may be a problem for these data. Whilst this approach is quite adept at identifying problem relationships between pairs of explanatory variables, it cannot always identify those instances where a combination of more than one variable predicts another. These relationships can, however, be determined using R^2 values to show the degree to which each explanatory variable can be explained using the other explanatory variables in the model. As with pair-wise correlations, we cannot say with any certainty how high the value of R^2 must be before multicollinearity is viewed as a cause for concern, but typically, values of about 0.8 or higher are taken as being indicative of a degree of multicollinearity which may be problematic. In the example above, if we predict a student's Mathematics score using their English score we obtain a regression model with an R^2 value of 0.804, which indicates that a

problematic level of multicollinearity may be present.

Calculating individual R^2 values for each explanatory variable in the model is a useful method of identifying instances of multicollinearity, but can be quite a lengthy process if there are a number of variables. It is, however, not necessary to manually compute these R^2 values as a number of analysis packages provide equivalent information through the 'tolerance' and 'variance inflation factor' (VIF) statistics shown in equations 3.22 and 3.21.

$$\text{Tolerance}(\beta_i) = 1 - R_i^2 \tag{3.21}$$

where β_i is the regression coefficient for variable i,
and R_i^2 is the squared multiple correlation coefficient between x_i and the other explanatory variables.

$$\text{VIF}(\beta_i) = \frac{1}{\text{Tolerance}} \tag{3.22}$$

We can see from equation 3.22 that an R^2 value of 0.8 will result in a VIF value of 5 and a tolerance value of 0.2. Any explanatory variables which have a VIF value of 5 or more, or a Tolerance of 0.2 or less, are therefore of interest as they show a degree of multicollinearity which could be problematic. Table 3.9 shows the regression analysis of the example data set with VIF and tolerance values of a high enough level to be of concern. It should be noted that as there are only two explanatory variables, the statistics for both variables are the same.

Table 3.9: VIF and Tolerance Statistics

	Coefficient	s.e.	t	P	Tolerance	VIF
English	0.086	0.046	1.863	0.100	0.196	5.099
Mathematics	0.068	0.038	1.805	0.109	0.196	5.099

$$\text{Places offered} = \alpha + \beta_1 \text{English} + \beta_2 \text{Mathematics}$$
$$F_{2,8} = 32.52,\ P < 0.0005,\ R^2 = 0.890$$

The tolerance and VIF statistics are based on the R^2 measure and therefore assume that the data are continuous. It is, however, possible to use these statistics on discontinuous data provided that the variables have been coded appropriately (see Section 3.3.7). This greatly increases the usefulness of the techniques as it enables problematic relationships between all types of variables to be identified. It should be noted, however, that these statistics can only give a rough indication of which relationships may be problematic, they do not provide any proof that multicollinearity will be a problem, nor do they identify all instances of problematic relationships. The tolerance and VIF statistics merely provide a convenient method for identifying

at least some of the relationship of concern.

Dealing with Multicollinearity

There are a number of ways in which multicollinearity can be reduced in a data set. These methods include:

1. Collect more data

   As multicollinearity is a problem which results from insufficient information in the sample, one solution is to increase the amount of information by collecting more data. As more data is collected and the sample size increases, the standard error tends to decrease which reduces the effect of multicollinearity (see Berry and Feldman, 1993). Although increasing the amount of data is an attractive option and one of the best methods to reduce multicollinearity (at least when the data set is relatively small), it is in many instances, not practical or possible, so other less attractive methods need to be considered.

2. Collapse variables

   One option to reduce the level of multicollinearity is to combine two or more explanatory variables which are highly correlated into a single composite variable. This approach is, however, only reasonable when the explanatory variables are indicators of the same underlying concept. For example, using the data in Table 3.7, it makes theoretical sense to combine the two explanatory variables (marks in English and Mathematics) into a single index of academic performance. This single index could simply be the sum of the two scores, or the average score for the two subjects. The use of a composite variable in the regression model enables one to assess the contribution made by academic performance to the number of college places offered, without the problem of a high degree of multicollinearity which existed between English and Mathematics scores.

   The process of combining variables into latent variables (or factors as they are sometimes called) is not always as straightforward as in theexample shown above, where two variables were related in quite an obvious way and could be combined easily into a single index. If there are a number of variables which are inter-related, it might be appropriate to first identify any latent variables in the sample using factor analysis and then enter these into the regression model. The technique of factor analysis is discussed in detail in Chapter 6.

3. Remove variables from the model

   When it is not possible to collapse highly related explanatory variables into a composite variable, one may delete one or more variables to remove the effect of

multicollinearity. This option, whilst being one of the easiest to accomplish practically, can be problematic if the variable measures some distinct theoretical concept which cannot be easily dismissed from the model. It should be noted that the removal of a relevant explanatory variable from a model can cause more serious problems than the presence of high multicollinearity (the removal of important variables may result in a model which is mis-specified, see Berry and Feldman, 1993). It is, therefore, generally unwise to remove explanatory variables from a regression equation merely on the grounds that they show a high degree of multicollinearity.

In general, the most reasonable method of dealing with multicollinearity is to collect more data and, where possible, collapse a number of variables into composite or latent variables, provided that they make theoretical sense. If no more data can be collected, the variables cannot be incorporated into a composite variable, and the highly related variables are deemed to be a necessary part of the model (and therefore cannot be removed), then one might just have to recognize its presence and live with its consequences (the consequence being that it is not possible to obtain reliable regression coefficients for all of the variables in the model).

3.3.7 Dummy Variable Coding Discontinuous Data

One of the requirements of OLS regression is that all variables entered into the model are measured on a continuous scale, a requirement which is, unfortunately, not met by all social science data. It is possible, however, to include dichotomous, or binary, data in a model since the OLS regression procedure treats this as continuous data which can only assume one of two values. The ability to include dichotomous data enables variables such as male-female, dead-alive, rich-poor, passed-failed, high-low etc., to be used as explanatory variables[6]. Multi-category categorical explanatory variables, such as drinking levels (high, medium and low), location (Europe, North America, South America, Africa), educational level (unqualified, high school, university), and different treatments (treatment 1, treatment 2, treatment 3, etc.) can also be included if they are dichotomized. The ability to include dichotomous and multi-category data in OLS regression models greatly increases the usefulness of the technique. The process of transforming discontinuous data into a form which can be entered into a regression model is called dummy coding. There are a number of methods of dummy coding data, however, only two of the more common methods, indicator and deviation coding, will be discussed in detail here.

Indicator Coding

Indicator coding is perhaps the easiest dummy coding method as it involves simply transforming data into a number of dichotomies. The dichotomy *must* be coded 0 and 1 (either explicitly in the data set, or internally by software) to indicate the presence or absence of a particular attribute. For example, in the case of gender, if a code of 0 refers to ‘not female’, a code of 1 refers to ‘female’. Alternatively, if a code of 1 refers to ‘male’, a code of 0 refers to ‘not male’[7]. It is not appropriate to code the data using other numbers, such as 1

and 2 (which is perhaps more intuitive for coding purposes) as the regression procedure attaches a specific meaning to these numbers (the group coded 2 will be twice the value of the group coded 1). By using the values 0 and 1 we are merely describing the presence or absence of a particular attribute, rather than defining its level.

Table 3.10 shows a categorical variable (the different categories are designated by the letters A, B, C and D) which has been recoded into a series of dichotomies. In the table, the original variable is represented as four separate dummy variables, each indicating the presence or absence of a particular category. For example, dummy variable 1 only records whether or not the original variable is equivalent to A, dummy variable 2 only records whether or not the original variable is equivalent to B, and so on. The information contained in the original variable is now represented by a number of discrete dummy variables.

The use of dummy variables in an OLS regression model is not completely straightforward because the inclusion of all of them at the same time leads to a situation where perfect multicollinearity exists (refer to section 3.3.6). For example, dummy variable 4 in Table 3.10 can be perfectly predicted from dummy variables 1, 2 and 3 and is therefore redundant. If any of the variables 1, 2 or 3 have the value of 1, then variable 4 necessarily takes the values of 0, and if variables 1, 2 and 3 all have the value of 0, then variable 4 necessarily takes the value of 1. In general, if we have j categories, a maximum of j — 1 dummy variables can be entered into a model. The dummy variable which is omitted is called the reference category and is the category against which other dummy

Table 3.10: The Indicator Method of Dummy Variable Coding

Original Variable	Dummy Variables 1	2	3	4
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1

variables are compared. This is also known as *aliasing* a variable. It should be noted that the choice of reference category is often quite arbitrary, although sometimes there will be reasons that a particular reference category is chosen. For example, when comparing a number of treatments for a particular illness, it might make sense to compare each with the standard treatment currently used to treat the disease (see Hardy, 1993, for a more in-depth discussion of reference category choice). The choice of reference category does not affect the model fit as this remains the same no matter which category is designated as the reference. Changing the reference category merely alters the way the differences are apportioned between the dummy variables. The setting of one level of a categorical variable to be the reference category is commonly called *aliasing*.

The use of dummy variables in OLS regression is demonstrated using data collected as part of a study of the quality of statements elicited from young children. Details of the data collection methods and overall design of the study can be found in Hutcheson, Baxter, Telfer and Warden, 1995. It should be noted that the original data set has been changed somewhat and that some variables have been added to enable certain analysis techniques to be demonstrated clearly. The data in Table 3.11 includes variables which record statement quality, the child's gender, age and maturity, how coherently they gave their evidence, the delay between witnessing the incident and recounting it, the location of the interview (the child's home, school, a formal interview room, and an interview room specially constructed for children), and whether or not the case proceeded to prosecution.

Table 3.11: The Quality of Children's Testimonies

Age	Gender	Location	Coherence	Maturity	Delay	Prosecute	Quality
5–6	male	3	3.81	3.62	45	no	34.11
5–6	female	2	1.63	1.61	27	yes	36.59
5–6	male	1	3.54	3.63	102	no	37.23
5–6	female	2	4.21	4.11	39	no	39.65
5–6	male	3	3.30	3.12	41	no	42.07
5–6	female	3	2.32	2.13	70	yes	44.91
5–6	female	4	4.51	4.31	72	no	45.23
5–6	female	2	3.18	3.08	41	no	47.53

Age	Gender	Location	Coherence	Maturity	Delay	Prosecute	Quality
5–6	male	1	2.66	2.72	13	no	45.81
5–6	female	3	4.70	4.98	29	no	49.38
5–6	male	2	4.31	4.21	39	yes	49.53
5–6	female	3	3.46	3.54	47	no	47.51
5–6	female	2	3.42	3.33	31	yes	50.54
5–6	male	1	3.08	3.07	82	no	51.25
5–6	female	4	4.04	4.12	29	no	51.63
8–9	male	3	3.12	3.01	51	yes	52.02
5–6	female	3	3.23	3.08	78	no	52.39
5–6	male	1	2.63	2.62	81	no	54.49
8–9	male	3	3.02	3.00	71	no	54.64
5–6	female	2	3.62	3.53	43	no	55.27
8–9	male	1	1.54	1.21	55	yes	55.47
5–6	male	3	2.79	2.74	45	no	55.56
8–9	male	1	2.76	2.71	27	no	56.47
8–9	female	1	4.63	4.66	74	no	56.72
5–6	male	3	3.35	3.07	88	yes	57.07
5–6	female	1	2.63	2.82	29	no	57.76
8–9	female	2	2.77	2.71	56	yes	57.87
8–9	male	4	3.26	3.91	55	yes	58.33
5–6	male	1	2.99	2.87	75	no	58.73
5–6	male	3	3.59	3.09	61	no	58.84
8–9	female	2	2.43	2.39	76	yes	59.19
8–9	male	4	2.41	2.38	45	no	59.64
8–9	male	1	2.37	3.36	90	yes	59.86
8–9	male	2	3.92	3.98	92	no	59.97
5–6	male	3	3.63	3.72	44	no	60.81
5–6	female	1	2.30	2.21	45	yes	60.88
8–9	female	4	1.56	1.92	16	no	61.83
8–9	female	2	2.92	2.81	23	yes	61.98
8–9	female	3	4.11	1.72	63	yes	62.09
5–6	female	1	2.49	2.51	87	no	62.54
5–6	female	2	2.41	2.04	55	no	62.84
8–9	female	3	2.56	2.65	66	yes	63.38
8–9	male	1	1.95	2.03	8	yes	63.67
8–9	male	2	2.52	2.49	75	no	64.34
8–9	female	3	2.15	2.09	81	yes	64.37
5–6	male	4	2.78	2.79	9	yes	65.07
5–6	female	1	2.95	3.01	39	no	65.77
8–9	male	4	1.89	3.45	68	yes	65.93
8–9	female	3	3.62	3.56	26	yes	66.60
8–9	female	2	2.16	2.91	10	no	67.08
8–9	male	1	3.65	3.91	52	yes	68.33
8–9	male	4	2.32	2.33	19	yes	68.44
8–9	male	1	1.34	1.42	77	no	68.90
8–9	female	1	2.02	2.11	25	yes	69.59
8–9	male	4	1.76	1.72	86	no	69.89
5–6	female	4	1.78	1.77	45	yes	70.45
8–9	female	2	2.52	2.54	2	yes	70.71
8–9	male	4	1.83	1.73	29	yes	71.40
8–9	male	4	1.51	1.73	45	yes	71.83
8–9	female	3	3.08	2.08	14	no	72.30
5–6	female	4	1.03	0.98	25	yes	72.61
8–9	female	1	1.76	1.65	21	no	73.99
8–9	male	4	2.00	1.94	17	yes	74.18

Table 3.12 shows the results of an OLS regression calculated on some of the data presented in Table 3.11. This analysis uses a selection of the available data and is only meant to provide an illustration of the use of dummy variable coding and not a full model of statement quality[8]. The response variable is the quality of statements elicited from children, whilst the explanatory variables are the delay in days from the incident to the interview and the location of the interview (the four different locations have been coded into three dummy variables with the special interview room acting as the reference category). The coefficients show the effect that each explanatory variable has on the response variable (each dummy variable can be considered as a separate explanatory variable). For example, the coefficient for the variable Delay shows that for each unit increase in the delay (an increase of 1 day), the quality of the statement elicited from a child decreases by an average of 0.099; a relationship that is not significant ($t_{65} = -1.897$; $P = 0.062$). When j — 1 dummy variables are included in the model (as they are here) each dummy variable indicates the effect of one of the locations compared to the reference category. For example, the parameters for the dummy variable Location-Home indicate that the quality of information elicited from children who are interviewed at home is 6.043 points less than the quality of information elicited from children who are interviewed in the special interview room (the reference category). The value of the **t** statistic suggests that this difference is not significant at the 0.05 level ($t_{65} = -1.664$; **P** = 0.101). Similarly, the parameters associated with the dummy variable Location-Formal indicate that the quality of information elicited from children who are interviewed in a formal interview room is 8.938 points less than those children interviewed in the special interview room, a difference which is significant at the 0.05 level ($t_{65} = -2.469$; $P = 0.016$).

Table 3.12: Modelling Statement Quality Using Delay and Location (indicator dummy-coded)

	Coefficient	s.e.	t	P
Delay	−0.099	0.052	−1.897	0.062
Location dummy variables				
Location-Home	−6.042	3.630	−1.664	0.101
Location-School	−7.042	3.550	−1.983	0.052
Location-Formal	−8.938	3.621	−2.469	0.016
(constant)	70.790	3.137	22.565	0.000

$$\text{Model: Quality} = \alpha + \beta_1 \text{Delay} + \beta_2 \text{Location-Home} + \beta_3 \text{Location-School} + \beta_4 \text{Location-Formal}$$

$$F_{4,65} = 3.455,\ P = 0.013,\ R^2 = 0.175$$

The reason that the parameters for each of the dummy variables in Table 3.12 provides a comparison between the category coded 1 and the reference category is that the other location categories are controlled for. For example, although dummy variable Location-Home only indicates the presence or absence of a home interview, the parameters for this variable relate to the comparison between interviews conducted at home with the reference category, and not, as might be expected from the dummy variable coding, a comparison

between interviews conducted at home and those conducted away from the home. This rather counter-intuitive meaning of the parameters comes about because when we include the other dummy variables in the model, these control for the other interview venues.

The regression coefficients associated with dummy variables only provide a comparison with the reference category when j — 1 dummy variables are included in the model. If fewer than j — 1 dummy variables are included, the interpretation of the parameters change. For example, if only the dummy variable Location-Home is included in the model, the parameters for this variable indicate the difference between an interview conducted at home and an interview conducted away from home. If dummy variables Location-Home and Location-School are the only ones to be included in the model, the parameters for Location-Home indicate the difference between interviews conducted at home to ones conducted in the formal and the special interview rooms (by including Location-School in the model, the effect of being interviewed at school has been controlled for). When interpreting the parameters associated with dummy variables one needs to check how many have been included in the model as this affects the interpretation of the regression parameters. This is quite an important point, particularly when automatic model selection procedures are used, since dummy variables are not always entered into the model as a group.

The choice of reference category does not affect the overall model fit as long as j — 1 dummy variables are included in the model. This is demonstrated in Table 3.13 where models 1 and 2 show the parameters associated with j — 1 dummy variables when different reference categories are used. The only difference between the models is that the effect of location is apportioned differently amongst the dummy variables. The model fit is, however, affected when dummy variables are removed, a result which is demonstrated by comparing models 1 and 2 with models 3 and 4.

Deviation Coding

When investigating the effect of a categorical variable, it is not always appropriate, or even desirable, to compare each dummy variable against a reference category as is the case with indicator coding. In such circumstances, deviation or effects coding may be used to compare each dummy variable to the group average (see Table 3.14).

Table 3.13: A Comparison of Four Models of Statement Quality Using Delay and Location (dummy-coded using indicator method)

	Coefficient	s.e.	t	P
Model 1				
Delay	−0.099	0.052	−1.897	0.062
Location dummy variables				
Location-Home	−6.042	3.630	−1.664	0.101
Location-School	−7.042	3.550	−1.983	0.052
Location-Formal	−8.938	3.621	−2.469	0.016
(constant)	70.790	3.137	22.565	0.000
Model 2				
Delay	−0.099	0.052	−1.897	0.062
Location dummy variables				
Location-Home	2.896	3.444	0.841	0.404
Location-School	1.896	3.571	0.531	0.597
Location-Special	8.938	3.621	2.469	0.016
(constant)	61.852	3.738	16.547	0.000
Model 3				
Delay	−0.132	0.052	−2.539	0.014
Location dummy variables				
Location-Home	−1.125	3.150	−0.357	0.722
Location-School	−2.624	3.182	−0.824	0.413
(constant)	67.852	2.991	16.547	0.000
Model 4				
Delay	−0.129	0.052	−2.483	0.016
Location dummy variable				
Location-Home	−0.310	2.983	−0.104	0.918
(constant)	66.707	2.715	24.566	0.000

Model 1: $F_{4,65} = 3.455,\ P = 0.013,\ R^2 = 0.175$
Model 2: $F_{4,65} = 3.455,\ P = 0.013,\ R^2 = 0.175$
Model 3: $F_{3,66} = 2.391,\ P = 0.077,\ R^2 = 0.098$
Model 4: $F_{2,67} = 3.263,\ P = 0.044,\ R^2 = 0.089$

The coding scheme used for deviation coding is similar to that used for indicator coding (see Table 3.10)

except for the way in which the reference category is identified. Using indicator coding, the reference category is always coded 0, but in deviation coding the reference category is explicitly coded as −1. Using the deviation coding method, the contrast indicated by each dummy variable is between the group coded 1 and the reference group. For example, in Table 3.14, the dummy variable Location-Home provides a contrast between interviews conducted at the child's home and those conducted in the special interview room, whilst the dummy variable Location-School provides a contrast between interviews conducted at the child's school and those conducted in the special interview room. When *j* — 1 dummy variables are included in the model, the parameters associated with each dummy variable indicate a comparison between the group coded 1 and the average of all of the groups.

Table 3.14: Deviation Dummy Variable Coding of Location

Location of Interview	Dummy Variables 1	2	3
Child's Home	1	0	0
Child's School	0	1	0
Formal Interview Room	0	0	1
Special Interview Room	−1	−1	−1

Table 3.15 shows a similar analysis to that shown in Table 3.12, however, this time the location variable has been dummy coded using the deviation coding method. Although the parameters associated with the variable 'delay' are identical in value and meaning to those in Table 3.12, the interpretation of the parameters associated with the dummy variables have changed. The parameters associated with the dummy variable Location-Home indicate the difference in the quality of interviews conducted in the child's home to the average quality of interviews from all of the interview venues. In this case we can say that interviews conducted at home are 0.537 of a point poorer than the average for all of the groups, a difference which is not significant (t_{65} = −0.248; P = 0.805). Similarly, the parameters associated Location-School show that those interviews conducted at the child's school are 1.536 points poorer than the average, a difference which is also not significant (t_{65} = −0.70 3; P = 0.485).

Using deviation coding, the comparison is between each dummy variable category and the average value. Therefore, the regression coefficients for the dummy variables remain constant no matter which category is chosen as the reference. This can be clearly seen in Table 3.15. In a similar fashion to indicator coding, we have to note how many dummy variables have been entered into the equation. As it is difficult to correctly interpret the regression coefficients when fewer than *j* – 1 categories are entered into the model, it is recommended that when deviation coding is used, all *j* — 1 categories are entered to ensure that the comparison is with the average of all groups and not a subset.

Tables 3.13 and 3.15 show similar models calculated using different dummy variable coding schemes. A quick inspection of these tables demonstrates that the type of dummy coding used does not affect the overall model

fit, it merely affects the distribution of the differences between the dummy variables. Looking at models 1 and 2, where *j* — 1 dummy variables have been included, we can see that the overall model fit remains the same, as do the parameters associated with the variable Delay. The only statistics to change are those associated with the dummy variables.

Table 3.15: A Comparison of Four Models of Statement Quality Using Delay and Location (dummy-coded using deviation method)

	Coefficient	s.e.	t	P
Model 1				
Delay	−0.099	0.052	−1.897	0.062
Location				
Location-Home	−0.537	2.166	−0.248	0.805
Location-School	−1.536	2.184	−0.703	0.484
Location-Formal	−3.433	2.159	−1.590	0.117
(constant)	65.285	2.720	24.003	0.000
Model 2				
Delay	−0.099	0.052	−1.897	0.062
Location				
Location-Home	−0.537	2.166	−0.248	0.805
Location-School	−1.536	2.184	−0.703	0.484
Location-Special	5.506	2.226	2.474	0.016
(constant)	65.285	2.720	24.003	0.000

Model 1: $F_{4,65} = 3.455,\ P = 0.013,\ R^2 = 0.175$

Model 2: $F_{4,65} = 3.455,\ P = 0.013,\ R^2 = 0.175$

Dummy Coding Ordered Categorical Data

Explanatory variables with ordered categories can be input into regression models by utilizing dummy coding procedures, however, this often incurs a ‘cost’ as the dummy coding techniques presented above do not retain information about order. For example, in the case of a variable which codes for four levels of drinking behaviour (high, medium, low, and abstinent), if indicator or deviation coding is used, the individual categories are treated as being unrelated, even though this is clearly not the case. Using these coding methods, information about the order of the categories is lost and the analysis consequently loses some power.

In circumstances where it is important to take account of the ordered nature of the data (for instance, when the effect is relatively small) it is advisable to use a scoring method where this information is retained (for example, in integer coding, 1, 2, 3…). If a continuous variable has been collapsed, then the mean value of the

original variable for each category can be used to score the ordered categories. A more detailed treatment of ordinality for both explanatory and response variables is given in Chapter 5.

3.3.8 Model Selection

Consider the two models presented in Table 3.16, which have been calculated using the data in Table 3.11 (for simplicity, only the main effects have been included in these models). Model 1 uses the variables Delay and Gender to model Quality, whilst the nested Model 2 only uses Delay. Although Model I has a larger R^2 value, which we would expect as it contains a greater number of terms, the F statistics show that the smaller model actually provides a more significant linear prediction of Quality. The inclusion of Gender in the model does not improve the prediction of Quality and can therefore be omitted without any significant loss of 'power'. Ideally, only those variables which contribute significantly to the prediction of the response variable should be retained. The removal of unimportant variables results in a simpler model which helps in interpretation and often provides a clearer insight into the way the response variable varies as a function of changes in the explanatory variables. In general, a good model should enable an accurate prediction to be made of the response variable, but only contain those explanatory variables which play a significant role.

Table 3.16: Model Selection

	Coefficient	s.e.	t	P
Model 1				
Delay	−0.133	0.051	−2.590	0.012
Gender	−1.108	2.578	−0.430	0.669
(constant)	67.372	3.144	21.429	0.000
Model 2				
Delay	−0.130	0.050	−2.571	0.012
(constant)	66.677	2.680	24.881	0.000

Model 1: $\text{Quality} = 67.372 - 0.133(\text{Delay}) - 1.108(\text{Gender})$
$F_{2,67} = 3.358,\ P = 0.041,\ R^2 = 0.091$

Model 2: $\text{Quality} = 66.677 - 0.130(\text{Delay})$
$F_{1,68} = 6.611,\ P = 0.012,\ R^2 = 0.089$

Criteria for Including and Removing Variables

Decisions about which variables may be entered or removed from a model can be made on the basis of the partial-F statistic. Using Equation 3.19, nested regression models can be compared to assess the effect that individual (or groups of) explanatory variables have on the response variable. For example, to assess

the effect that Gender has on Quality (whilst controlling for Delay), one can compare the model Quality = $\alpha+\beta_1$ Delay + β_2Gender with the model Quality = $\alpha+\beta_1$ Delay. The value of partial-*F* indicates the effect that Gender has on Quality when Delay is controlled for. To calculate partial-*F*, the residual sum of squares and the degrees-of-freedom for the two regression models need to be calculated. From software, these are computed to be:

$$\text{Quality} = \alpha + \beta_1 \text{Delay} + \beta_2 \text{Gender} \qquad RSS = 7648.585,\ df = 67$$

$$\text{Quality} = \alpha + \beta_1 \text{Delay} \qquad RSS = 7669.682,\ df = 68$$

which, when entered into Equation 3.19 gives the value of partial-*F* as

$$F_{68-67,68} = \frac{7669.682 - 7648.585}{(68-67)(7648.585/67)}$$

$$F_{1,68} = 0.1848; P = 0.669.$$

The effect that Gender has on the model is shown by the partial-F statistic. In this case, as the difference in the number of terms between the two models is equal to 1 the partial-*F* statistic is equivalent to the t statistic, which is commonly provided by software (in fact, $\sqrt{F} = t$). For the model above it can be seen that this is the case as $\sqrt{0.1848}$ does indeed equal 0.430 (see Table 3.16). The *t* statistic therefore shows the unique contribution of individual terms in the model and can be used to determine which terms may be retained or removed during the process of model-building. When more than one term is added or removed from a model in a single step, the partial-F statistic can be used to determine significance (an example of the use of *t* and partial-F values in model-building is provided in Section 3.6).

3.3.9 Automated Model Selection

Automated selection procedures can be used to make decisions about whether terms are included or excluded from a regression model on statistical grounds according to how much the variables contribute to predicting the response variable. Ideally, such decisions should be based on theoretical as well as statistical grounds, however, it is sometimes convenient to use automated procedures. Whilst such a technique of model-building is relatively quick and efficient at deriving a model which provides a good prediction of the response variable, it does not always provide a model which is adequate for explanatory purposes. Agresti makes the point that …

> Computerized variable selection procedures should be used with caution. When one considers a large number of terms for potential inclusion in a model, one or two of them that are not really important may look impressive simply due to chance. For instance, when all the true effects are

> weak, the largest sample effect may substantially overestimate its true effect. In addition, it often makes sense to include certain variables of special interest in a model and report their estimated effects even if they are not statistically significant at some level.
>
> Agresti, 1996; p. 129.

An additional problem with automated variable selection procedures is that most software packages will not preserve the hierarchy of terms when interactions are examined. For example, a main effect may be removed from the model even though an interaction including that variable is included. This is problematic since to appropriately interpret an interaction one needs to include the main effects in the model. We recommend that if an automated model selection procedure is used, it should be used with caution, particularly when interactions are being considered, and only in conjunction with theoretical considerations about the most useful form of the model.

Even with these reservations, automated selection procedures are widely used. For this reason, three of the more common methods are discussed below, forward selection, backward deletion, and stepwise selection. It should be noted that none of these selection procedures is 'best' in any absolute sense. They merely identify subsets of variables which, for the sample, are good predictors of the response variable.

Forward Selection

The forward selection method of automated model-building selects terms to enter into the model singularly, on the basis of relative importance, as determined by the partial-*F* test. The first term to be entered into the model is the one which, if added, results in the most significant change in the value of *F* (as determined by the partial-*F*, or equivalent *t* statistics), provided that this meets an 'entry criterion' (for variable entry, this is usually set at $P < 0.05$). Once a term has been added to the model, the regression is recalculated and partial-*F* values obtained for all terms still to be considered for entering into the model. Of these terms, the one which would result in the most significant change to the *F* statistic is entered into the model (provided that it is above the criterion) and the regression recalculated. This procedure continues until all terms are either included in the model or until no more reach the required level of significance.

Backward Elimination

The backward elimination technique of model building is very similar to forward selection except that the starting model is one where all explanatory variables are entered and terms are then removed from the model sequentially. At each step in the process, the term which, if removed, results in the smallest significant change in the value of *F* (as denoted by the partial-*F* or *t* statistics), is removed from the model — provided that it has reached a 'removal criterion' (in backward elimination the removal criterion is usually set at $P = 0.1$). After each term is removed, the regression equation is recalculated and those terms left in the model are re-examined to see if any contribute less than the criterion level (as determined by partial-*F*). This process continues until all terms have been removed from the model, or until no more reach the criterion for removal.

Stepwise Selection

Stepwise selection is simply a combination of forward selection and backward elimination and is one of the most commonly used methods of automated variable selection. Stepwise selection builds a model in much the same way as in forward selection except that at each step, rather than just assessing which terms can be added to the model (using the forward selection procedure), those which are currently included are tested to see if any can be removed (using the backward elimination procedure). The advantage of this procedure is that those terms whose importance diminishes as additional terms are added can be removed.

Stepwise selection proceeds in the same way as forward selection until two terms have been entered into the model. At this point, both are examined to see if they still meet the criteria for retention. If either of them do not, the least significant one is removed and the regression recalculated. This process of removing terms continues until all are above the criterion for retention (P < O.1). The model selection procedure then proceeds to test whether any of the terms currently not in the model reach the criterion for entry ($P < 0.05$). If any do reach the criterion, the most significant one is entered into the model and the regression is recalculated. The process is then repeated with all terms in the model checked for significance and if any are below the criterion for retention they are removed. This procedure of entering and removing terms continues until no more variables reach the entry criterion.

The automated model selection procedure for a number of software packages rely on the inputting of terms into the model individually. It should be noted that this can cause some problems for categorical variables which have been dummy coded, since these will often require to be treated as a group so that the regression parameters can be appropriately interpreted (see Section 3.6 for a demonstration of how categorical variables can be defined as a group using SPSS). As grouped terms often have to be dealt with using manual modelling procedures, the use of automated model selection procedures appears less attractive. Given this and earlier comments about automated selection techniques, they should be used with a degree of caution.

3.4 A Worked Example of Multiple Regression

This section shows a complete example of how OLS regression may be used to analyse a data set. It should be noted that this example is only designed to provide a demonstration of how the techniques described in this chapter may be used and does not provide ‘the’ correct way to analyse these data. The data to be analysed are the child witness data shown in Table 3.11 which have already been used for demonstration purposes earlier in this chapter. These data contain four continuous variables, Quality of statement, Coherence of evidence, Maturity of child, and the Delay between the child witnessing the incident and recounting it, three dichotomous variables, Age (5–6 and 8–9- year-olds), Gender and whether or not the cases led to a Prosecution[9], and one categorical variable, Location, which indicates where the interview was conducted (at the child's home, school, in a formal interview room, or in a room specially designed for children). All the dichotomous variables were dummy coded 0 and 1 with the code of 1 indicating 8–9-year-old

and male children. Location was coded into three dummy variables using the indicator coding method with the formal interview room chosen as the reference category (this can be considered to be the 'standard' location of interviews and was therefore ideal to use as a reference category). For simplicity we will assume that none of the variables require transforming, that there are no outliers, and that the residuals do not give any cause for concern (see Chapter 2).

A useful first step in the analysis of these data is to determine if there are any associations between the explanatory variables which could give cause for concern. Table 3.17 shows that the variables Coherence and Maturity have VIF and tolerance values above recommended limits. It is quite easy to hypothesize why this might be the case for these data, for Coherence and Maturity are likely to be rated similarly since a child rated as being relatively mature is likely to have been perceived this way, at least partly as a result of the coherence of his or her evidence. A highly significant pair-wise correlation between these variables ($r = 0.91$; $P < 0.0005$) appears to confirm this view. This problematic association was 'solved' in this example by removing Maturity from the analysis as, of the two, this variable was perhaps the more difficult to rate accurately. The removal of Maturity from the model shown in Table 3.17, will not make a significant difference to the model fit ($t_{60} = -0.774$;p = 0.442) and once removed reduces the tolerance and VIF values for Coherence to more acceptable levels (on a recalculation of the initial statistics the tolerance and VIF values for Coherence are reduced to 0.743 and 1.346, respectively).

Table 3.17: Initial Statistics

	Coefficient	s.e.	t	P	Tolerance	VIF
Age	10.521	2.220	4.739	0.000	0.817	1.223
Coherence	−0.587	3.395	−0.173	0.863	0.129	7.754
Delay	−0.062	0.043	−1.441	0.155	0.847	1.181
Gender	2.116	2.162	0.979	0.331	0.856	1.169
Location dummy variables						
Location-Home	1.553	3.024	0.514	0.609	0.572	1.747
Location-School	0.277	2.993	0.092	0.927	0.607	1.647
Location-Formal	6.260	3.302	1.896	0.063	0.499	2.006
Maturity	−2.400	3.100	−0.774	.442	0.148	6.779
(constant)	63.109	6.038	10.452	0.000		

In order to demonstrate the use of an interaction term in OLS regression, one is constructed here and entered into the model. The interaction chosen here is that between Age and Coherence and is represented by the variable Age × Coh which is the product of Age and Coherence. Although there are many interaction terms which could have been entered, all of them are, however, insignificant. The interaction term AgexCoh is chosen merely to demonstrate how such a term may be included in an OLS regression model. Once all the variables to be considered for the model have been appropriately coded and any problematic levels of multicollinearity dealt with, insignificant terms may be removed from the model. For the purpose of

demonstration, a stepwise selection procedure is used here to select a subset of terms to model Quality. A proposed model is shown in Table 3.18 and shows that Age, Coherence, Delay, and the three dummy variables representing Location have been retained in the model[10]. Using the *t* statistics, the contribution of each of the dummy variables to the model can be determined. The *t* statistics will not, however, provide the significance of Location overall. To do this we need to compute a partial-*F* statistic. The residual sum of squares statistic for the model which includes all of the location dummy variables is 4389.317 with 63 degrees-of-freedom and the nested model where the dummy variables are removed has a residual sum of squares equal to 4639.385 with 66 degrees-of-freedom. Entering these data into Equation 3.9 partial-*F* is computed as:

$$F_{66-63,63} = \frac{4639.385 - 4389.317}{(66 - 63)(4389.317/63)}$$

$$F_{3,63} = 1.196; P = 0.319.$$

Table 3.18: Model Selected Using Stepwise Selection

	Coefficient	s.e.	*t*	*P*
Variables included in the model				
Age	9.885	2.155	4.573	0.000
Coherence	−3.112	1.406	−2.214	0.030
Delay	−0.072	0.042	−1.701	0.094
Location dummy variables				
Location-Home	0.532	2.891	0.184	0.854
Location-School	0.053	2.902	0.018	0.986
Location-Special	4.880	3.087	1.581	0.119
(constant)	65.999	5.573	11.842	0.000
Variables not included in the model				
Gender	0.104		1.064	0.291
age*coh	0.425		1.301	0.198

$F_{6,63} = 9.631,\ P < 0.0005,\ R^2 = 0.478$

On this evidence, Location does not make a significant contribution to the model and can therefore be removed. The final model which contains only the variables Age, Coherence and Delay is shown in Table 3.19.

Table 3.19: Final Model

			95% CIs for β			
	Coefficient	s.e.	Lower	Upper	t	P
Age	9.862	2.152	5.566	14.157	4.583	0.000
Coherence	−3.786	1.319	−6.418	−1.153	−2.871	0.005
Delay	−0.086	0.041	−0.167	−0.006	−2.133	0.037
(constant)	69.878	4.550	60.794	78.963	15.358	0.000

$F_{3,66} = 17.905,\ P < 0.0005,\ R^2 = 0.449$

Once a final model has been derived, the regression parameters can be interpreted. From Table 3.19 it can be seen that as Age increases from 5–6 to 8–9 years (a unit increase) Quality improves by an average of 9.862 (P < 0.0005). Furthermore, we can say with 95% confidence that we can expect the improvement to be between 5.566 and 14.157. Compared to 5–6-year-old children, 8–9-year-olds are expected to provide statements which are at least 5.6% and at most 14.2% higher quality. Similarly, as the coherence score for the child increases by one unit, Quality decreases by an average of 3.786. As the initial coding scheme assigned low values to children who were coherent and high values to those who were incoherent, this result means that the children who are rated as more coherent provide higher quality statements. Similar interpretations of the regression coefficients and associated confidence intervals can be applied to Delay.

One can also calculate confidence intervals for the fitted values of mean *Y* and the prediction intervals. For example, a child aged 5–6 with a coherence rating of 3.81 and a delay of 45 days can be expected to provide a statement of quality 51.565. This result can be derived using software, or manually using the equation:

Quality = 69.878 + 9.862 Age — 3.786 Coherence — 0.086 Delay.

From software, the confidence intervals for the fitted values of mean *Y* are 48.026 to 55.104, and the prediction intervals are 34.456 to 68.675. For a large sample of 5–6-year-old children with a coherence rating of 3.81 and a delay of 45 days (if such a sample could be found), one would expect the mean statement quality to be at least 48.026 and at most 55.104. An individual with the same characteristics one would expect to provide a statement with a quality of at least 34.456 and at most 68.675. The software commands to obtain these statistics can be viewed in Section 3.6.

This worked example has demonstrated how OLS regression might be used to analyse a selection of continuous, dichotomous and categorical variables. Problematic levels of multicollinearity were first identified and dealt with, interaction terms were computed and entered into the model (explicitly defining interaction terms in this way is a requirement of some, but not all software packages) and then an automatic selection process (stepwise) used in conjunction with manual modelling techniques (for categorical data) was used to derive a final model. In practice, it would also be wise to investigate the model residuals (see Chapter 2), but

to save space in this example, these were assumed not to give any cause for concern. The SPSS and GLIM software commands for the above procedures are provided in Section 3.6.

3.5 Summary

OLS regression provides a model-building approach to the analysis of continuous data. It is a generalized linear model which directly maps the random component of a model (the continuous response variable) onto the systematic component (the explanatory variables expressed as a linear function). OLS regression explicitly models continuous variables and although it is most easily demonstrated using continuous explanatory variables, the application of dummy coding methods enables dichotomous and categorical explanatory variables to also be included in the model. By this means, OLS regression allows one to carry out analysis of variance and analysis of covariance which have traditionally been treated as separate techniques. This greatly increases the utility of the technique and has no doubt contributed to its popularity.

Since OLS regression is a form of generalized linear model, it shares much common ground with the related techniques of logistic regression and loglinear analysis, which are used to model dichotomous and categorical data. Due to the common theoretical underpinnings of these models, many of the methods discussed in this chapter can be applied to other generalized linear modelling techniques described in Chapters 4 and 5. The form and interpretation of the OLS regression equation, dummy variable coding and model-building have direct relevance to the techniques used in logistic regression and loglinear analysis and should be viewed as being complimentary to these chapters. The material in this chapter provides much of the information required to understand and utilize logistic regression and loglinear analysis and can therefore be regarded as an introduction to these techniques.

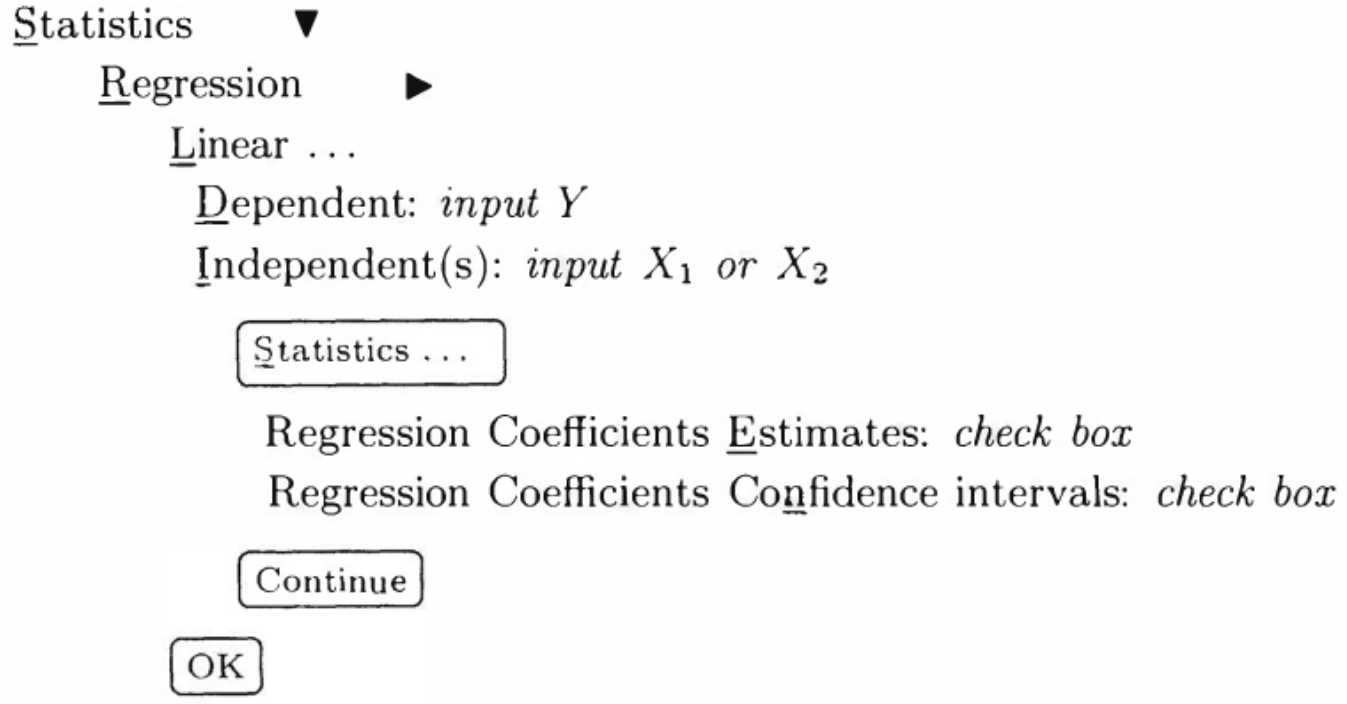
3.6 Statistical Software Commands

3.6.1 SPSS

Detailed discussions of the SPSS procedures outlined in this section can be found in Norušis (1994) and in the appropriate SPSS manuals (see, for example, SPSS Inc., 1996a).

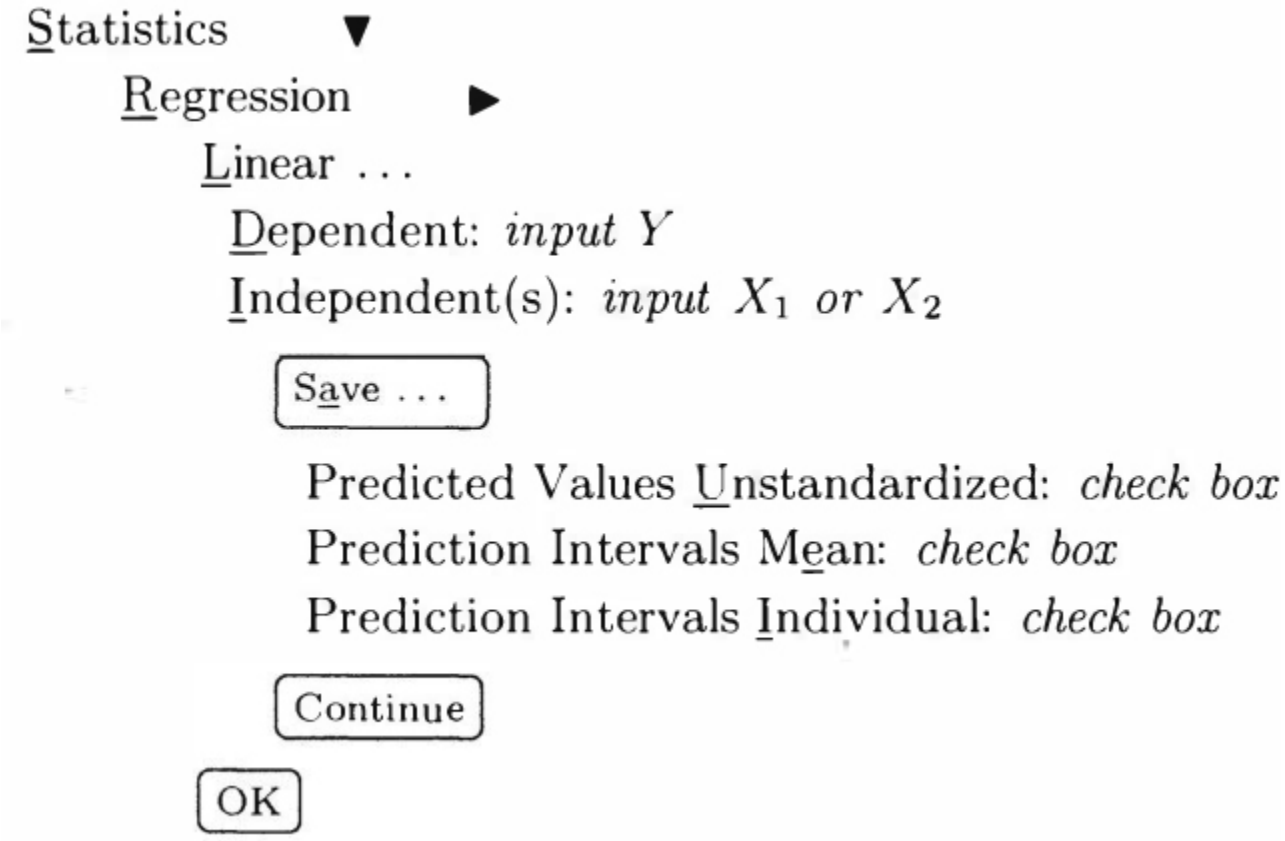
Computing Regression Coefficients and Associated Confidence Intervals

Input the data from Table 3.1 in three columns, exactly as shown in the table. The regression coefficients and associated confidence intervals can be derived using the commands:



Predicting Values of *y* and Associated Confidence and Prediction Intervals

Input the data from Table 3.1 in three columns, exactly as shown in the table. The following commands can be used to generate statistics relating to the response variable.



Running the above regression model will create five new variables; the predicted value of Y (pre_1), the upper and lower confidence intervals for the fitted mean value of Y (lmci_1 and umci_1), and the upper and lower prediction intervals for individual values of Y (lici_1 and uici_1). You will note that predictions are only given for those points included in the data set. It is possible, however, to obtain predictions for *Y* given any value of the explanatory variable by inputting the value of *X* into the spreadsheet. For example, the prediction intervals for *Y* when X_1 is equal to 29.8 can be calculated by the following method:

- Enter a new case for variable X_1 and give it the value of 29.8 (i.e., add the value 29.8 onto the bottom of the column of values for X_1).
- **Do not** enter any value for the Y variable.
- Run the regression as above and the predicted values for all cases will be saved including the values for the new case.

This procedure works in SPSS as cases are removed from the analysis listwise. The data point added will, therefore, play no part in the calculation of the regression model, but the programme will nevertheless calculate a value of *Y* when X_1 is equal to 29.8. In this case, the predicted value of *Y* is 11.005, the prediction intervals are 9.08, 12.93, and the confidence intervals for *Y* are 9.94, 12.07.

Calculating the Residual Sum of Squares (RSS) Statistics

The value of *F* for the models shown in Table 3.4 can be calculated using Equation 3.9. To do this manually, we need to compute the value of RSS for the null and final models using OLS regression (see above). For the model $Y = \alpha + x_1$, SPSS provides the statistics shown in Table 3.20.

Table 3.20: Statistics for Calculating *F*

	Sum of Squares	Degrees of Freedom	Mean Squares
Regression	105.525	1	105.525
Residual	7.190	13	0.553
Total	112.715	14	

From Table 3.20, RSS_{null} = 112.715 (the total deviance in the model) and RSS_{final} = 7.190 (the deviance in the model that includes the explanatory variable). df_{null} = 14 (n — *k* — 1, where *k* = 0) and df_{final} = 13 (n — *k* — 1, where *k* = 1). Inputting these values into Equation 3.9, the value of *F* can be calculated as:

$$F_{1,13} = \frac{112.715 - 7.190}{(7.190/13)} = 190.796$$

RSS statistics are also required for computing partial-*F* statistics when comparing nested models (see

Equation 3.19). Table 3.21 shows how the RSS statistics were derived for the model shown in Table 3.18. To obtain these statistics, one needs to compute two regression models (see above), one model which includes the locations and one which does not.

From Table 3.21, RSS_{p+q} = 4389.317 (the deviance in the larger model) and RSS_p = 4639.385 (the deviance in the smaller model). df_{p+q} = 63 (n — *k* — 1, where *k* = 6) and df_p = 66 (n — *k* — 1, where *k* = 3). Inputting these values into Equation 3.19, the value of partial-*F* is calculated as:

$$F_{66-63,63} = \frac{4639.385 - 4389.317}{(66 - 63)(4389.317/63)}$$

$$F_{3,63} = 1.196$$

Calculating Tolerance and VIF Values

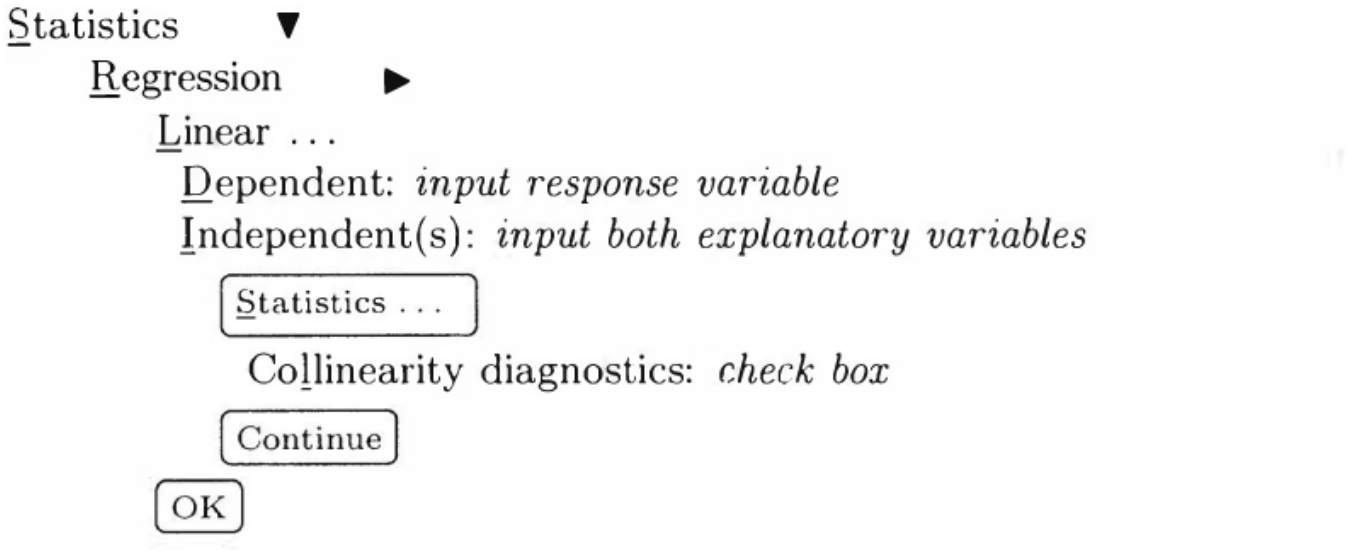
Enter the data from Table 3.7 into an SPSS spreadsheet. To calculate the tolerance and VIF statistics, use the commands:

Table 3.21: Statistics for Calculating Partial-F

	Sum of Squares	Degrees of Freedom	Mean Squares
The larger model (p+q)			
Regression	4025.988	6	670.998
Residual	4389.317	63	69.672
Total	8415.306	69	
The smaller model (p)			
Regression	3775.920	3	1258.640
Residual	4639.385	66	70.294
Total	8415.306	69	

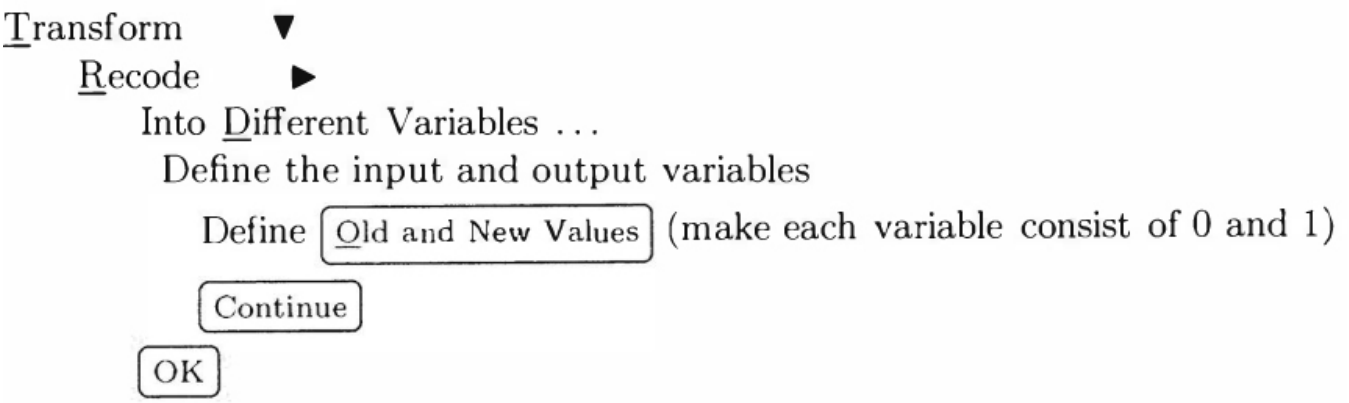
Model p+q: $\alpha + \beta_1 \text{Age} + \beta_2 \text{Coherence} + \beta_3 \text{Delay} + \beta_4 \text{Location}$ (3 dummy variables)

Model p: $\alpha + \beta_1 \text{Age} + \beta_2 \text{Coherence} + \beta_3 \text{Delay}$

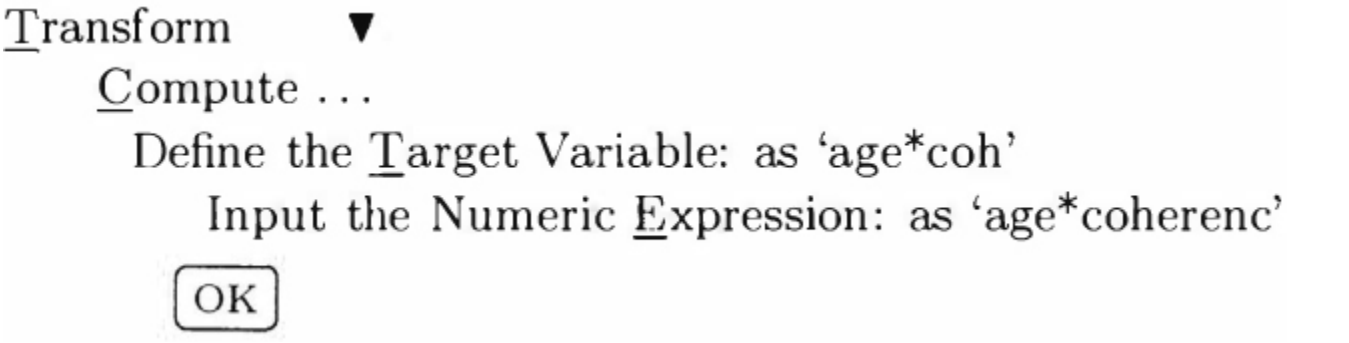


Specific Code Relating to the Example in Section 3.4

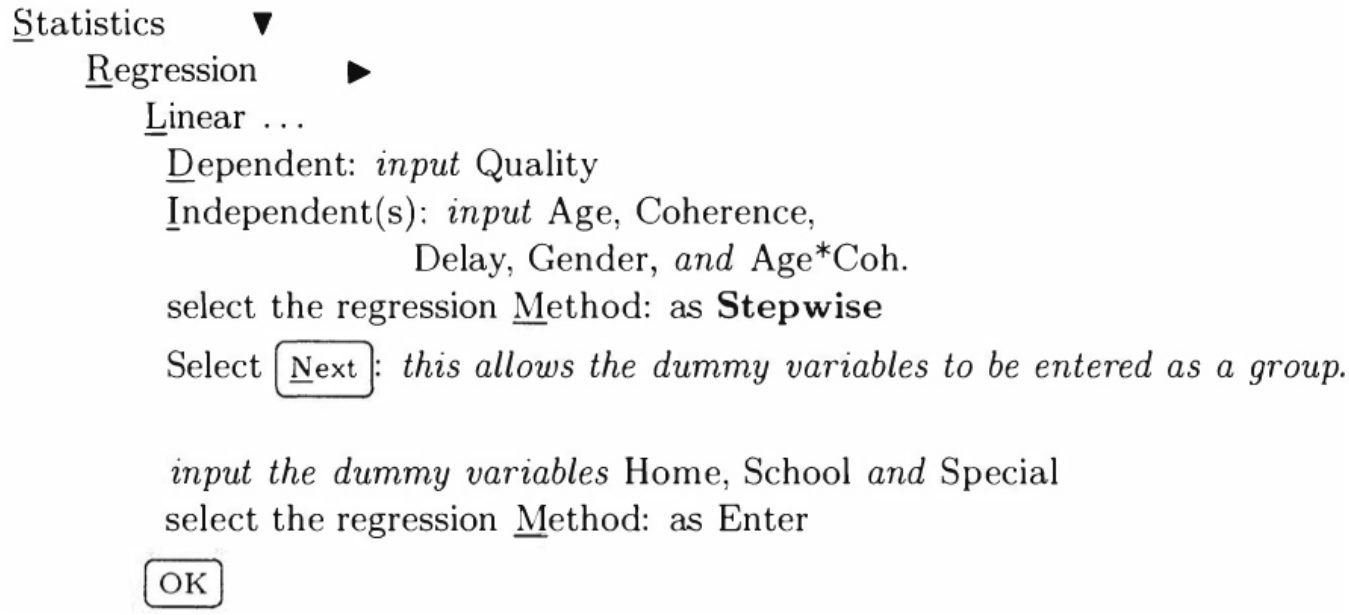
Input the data from Table 3.11 into an SPSS spreadsheet. To define the dummy location variables, one can recode the location variable using the indicator coding method:



The interaction between the variables ‘age’ and ‘coherence’ can be defined using the commands:



The stepwise regression model shown in Table 3.18 can be computed using the commands:



3.6.2 GLIM

In order to run the GLIM examples presented below, type in the following command from within GLIM: $INPUT 'FILENAME. EXT'*$*, replacing FILENAME.EXT with that of each example file. To aid the reader, we have commented the commands extensively within each file. All macros which have been used are included at the end of this section.

Computing Regression Coefficients and Associated Confidence and Prediction Intervals

```
$echo on $
!       Example from the data set in Table 3.1
!       First input any macros that we require.
$input %plc 80 NORMAC $ $echo on $
$input 'ols.mac' $
!       Read in the data set.
$slen  15 $
$data Y X1 X2 $
$read
0.52    10.50   11.09
2.09    13.19   12.91
etc.
etc.
7.01    22.68   24.19
8.92    26.28   24.75
```

```
$
!       Declare the response variable.
$yvar Y$
!       Fit the first model.
$fit X1 $
!       Display parameter estimates and standard errors.
$dis e $
!       Call macro to calculate a t-test and the 95%
!       Confidence Intervals for each parameter.
$use ttest $
!       Fit and examine the second model.
$fit X2 $
$dis e $
$use ttest $
!       Predict fitted values for all original values of the explanatory
!       variable, for the first model.
$fit X1 $
$predict X1=X1 $
!       Call macro to tabulate the fitted values, confidence intervals for
!       fitted Y, and the prediction intervals, with the original values of
!       the explanatory variable.
$look X1 $
$use cipiori $
!       Predict fitted value for a new value of the explanatory variable.
$predict X1=29.8 $
!       Call macro to display fitted value, confidence interval for
!       fitted Y, and the prediction interval, with a new value of
!       the explanatory variable.
$use cipinew $
!       Predict fitted values, 95% Confidence Intervals for Fitted Y,
!       and the Prediction Intervals, with all the original values of
!       the explanatory variable, for the second model.
$fit X2 $
$predict X2=X2 $
$look X2 $
$use cipiori $
!       Now calculate these for a new value of X2.
$predict X2=29.8 $
$use cipinew $
!       Calculate R-squared and and F-test for the first model.
$fit X1 $
!       Call macro to calculate r-squared.
$use rsq $
!       Call macro to set rss and df for fitted model.
$use rssdf1 $
!       Remove the tested model term.
$fit -X1$
!       Call macro to set rss and df for reduced model.
$use rssdf2$
!       Call macro to calculate F-test.
$use ftest $
```

```
$use rsq $
$use rssdf1 $
$fit -X2 $
$use rssdf2 $
$use ftest $
$return $
```

Calculating Tolerance and VIF Values

```
$echo on $
!       Example from the data set in Table 3.7
!       First input any macros that we require.
$input %plc 80 NORMAC $ $echo on $
$input 'ols.mac' $
!       Read in the data set.
$slen  11 $
$data PLACES ENGLISH MATHS $
$read
0       22      17
1       32.5    34.5
etc.
etc.
9       85.5    74.5
10      68.5    87.6
$
!       Declare the response variable.
$yvar PLACES $
!       Fit the model to be tested.
$fit ENGLISH+MATHS $
!       Call macro to calculate r-squared.
$use rsq $
!       Display parameter estimates and standard errors.
$dis e $
!       Call macro to calculate a t-test of each parameter.
$use ttest $
```

```
$fit MATHS $
$use rsq $
$dis e $
$use ttest $
$use rssdf1 $
$fit -MATHS $
$use rssdf2 $
$use ftest $
!       Calculate VIF and Tolerance values for explanatory variables.
!       First calculate VIF and Tolerance for ENGLISH.
$yvar MATHS $
$fit ENGLISH $
$use tolvif $
!       Then calculate VIF and Tolerance for MATHS.
$yvar ENGLISH $
$fit MATHS $
$use tolvif $
$return $
```

Analysis of Child Witness Data Presented in Section 3.4.

```
$echo on $
!       Example from the data set in Table 3.11
!       First input any macros that we require.
$input %plc 80 NORMAC $ $echo on $
$input 'ols.mac' $
$slen  70 $
$factor age 2 gender 2 location 4 $
$data
AGE     GENDER  LOCATION
COHERENC        DELAY
MATURITY        QUALITY PROSECUT $
$read
1 1 3 3.81 45 3.62 34.11  0
1 2 2 1.63 27 1.61 36.59  0
    etc.
2 2 2 1.94 46 1.99 80.67  1
2 1 4 1.89 15 1.87 83.15  1
$
$yvar QUALITY $
!       Note that prosecut is a response variable used in the
!       Logistic Regression chapter, and is not used in this analysis
```

```
!       Set reference catagory for LOCATION to level 3.
$factor location 4 (2) $
!       Fit and test the DELAY+LOCATION model
$fit DELAY+LOCATION
$use rsq $
$use ttest $
$use rssdf1 $
$fit ~(DELAY+LOCATION) $
$use rssdf2 $
$use ftest $
!       Calculate VIF and Tolerance for each continuous explanatory
!       variable as in Table 1.20.
!       Note that the declaration of a factor as the response variable
!       is not supported with the macro viftol since GLIM uses internal
!       dummy coding of factors ---  a error message is given.
$yvar COHERENCE $
$fit AGE+DELAY+GENDER+LOCATION+MATURITY $
$use tolvif $
$yvar DELAY$
$fit AGE+COHERENCE+GENDER+LOCATION+MATURITY $
$use tolvif $
$yvar MATURITY $
$fit AGE+COHERENCE+DELAY+GENDER+LOCATION $
$use tolvif $
!       Check the correlation between MATURITY and COHERENCE
$yvar COHERENCE $
$fit MATURITY $
$use rsq $
!       Fit the full main effects model and try removing MATURITY.
$yvar QUALITY $
$factor location 4 (3) $
$fit AGE+GENDER+LOCATION+COHERENC+DELAY+MATURITY $
$use rsq $
$dis e $
$use ttest $
$use rssdf1 $
$fit -MATURITY $
$use rssdf2 $
$use ftest $
$use rsq $
$dis e $
$use ttest $
!       Fit model of Table 1.22.
```

```
$factor LOC2 2 $
$calc LOC2=%eq(LOCATION,4)+1$
!       Refit model using LOC2.
$fit AGE+COHERENCE+LOC2 $
$use rsq $
$dis e $
$use ttest $
$use rssdf1 $
$fit -(AGE+COHERENCE+LOC2) $
$use rssdf2 $
$use ftest $
!       Refit model of interest.
$fit AGE+COHERENCE+LOC2 $
!       Predict fitted value for a new value of each explanatory
!       variable. First example.
$predict AGE=1 COHERENCE=3.2 LOC2=1 $
!       Call macro to display fitted value, confidence interval for
!       fitted Y, and the prediction interval, with a new value of
!       the explanatory variable.
$use cipinew $
!       Predict fitted value for a new value of each explanatory
!       variable. Second example.
$predict AGE=2 COHERENCE=1.3 LOC2=2 $
$use cipinew $
!       A simultaneous test of all two-way interaction terms.
$fit (AGE+GENDER+LOCATION+COHERENC+DELAY+MATURITY)**2$
$use rsq $
$use rssdf1$
$fit AGE+GENDER+LOCATION+COHERENC+DELAY+MATURITY $
$use rsq $
$use rssdf2$ $use ftest $
$return $
```

GLIM Macros for OLS Regression

```
!       Macros for use in OLS Regression Chapter.
$mac tolvif
$num tol vif $
!       Outputs tolerance and VIF values.
$use rsq $
$calc tol=1-%r $
$calc vif=1/tol $
$pr 'Tolerance ='tol' and VIF ='vif ;$
$endmac $
$mac ttest
!       Produces a table of t, P-values, and CI's
!       for the model parameters.
$num td95 $
$extract %pe %se $
$calc t=%pe/%se : pt=2*(1-%tp(%abs(t),%df)) $
$calc td95=%td(0.975,%df) $
$calc pelo=%pe-(td95*%se) $
$calc pehi=%pe+(td95*%se) $
```

```
$acc 4 3 $ $look %pe %se t pt pelo pehi $ $pr $ $acc $
$delete t pt pelo pehi $
$endmac $
$num rss1 rss2 df1 df2 df21 s2 ss21 ms21 f p $
$mac rssdf1 $assign rss1=%dv: df1=%df: s2=%sc $ $endmac $
$mac rssdf2 $assign rss2=%dv: df2=%df $ $endmac $
$mac ftest
!       Calculates F-test
$calc ss21=rss2-rss1: df21=df2-df1 $
$calc ms21=ss21/df21: f=ms21/s2 $
$calc p= 1-%fp(f, df21, df1) $
$pr 'F  ='f' with 'df21', 'df1' df and P ='p ;$
$endmac $
$mac cipiori
!       Calculate and tabulate fitted values, confidence intervals
!       for fitted Y, and the prediction intervals,
!       with the original values of the explanatory variables.
$calc td95=%td(.975,%df) $
$calc cifylo=%pfv-td95*%sqrt(%pvl): cifyhi=%pfv+td95*%sqrt(%pvl)$
$calc pilo=%pfv-td95*%sqrt(%sc+%pvl): pihi=%pfv+td95*%sqrt(%sc+%pvl)$
$acc 4 3$ $look %pfv cifylo cifyhi pilo pihi$ $pr $ $acc $
$endmac $
$mac cipinew
!       Calculate and display fitted value, confidence interval for
!       fitted Y, and the prediction interval, with a new value of
!       the explanatory variable.
$num cifynewlo cifynewhi pinewlo pinewhi  $
$calc td95=%td(.975,%df) $
$calc cifynewlo=%pfv-td95*%sqrt(%pvl): cifynewhi=%pfv+td95*%sqrt(%pvl)$
$calc pinewlo=%pfv-td95*%sqrt(%sc+%pvl): pinewhi=%pfv+td95*%sqrt(%sc+%pvl)$
$pr 'Fitted value = '%pfv' with CI for fitted Y of ('cifynewlo','cifynewhi')
and PI of ('pinewlo',' pinewhi')';$
$endmac $
$return $
```

[1] Although a continuous scale is assumed, OLS regression is often used to model ordered categorical data when there are a relatively large number of levels and an underlying continuous distribution can be assumed.

[2] Although we have so far dealt exclusively with continuous data, this example introduces the use of discontinuous data in regression. The variables ‘gender’ and ‘educational achievement’ are not continuous but can still be included in a regression model provided that they have been appropriately coded (see Section 3.3.7).

[3] It is also possible to test the significance of a partial regression coefficient using an *F* test with *k* and *n*-*k*-1 degrees-of-freedom (see Edwards, 1985, for a discussion of this).

[4] For this example w e will assume that all students applied to the same 10 colleges.

[5] 0.8 is an arbitrary figure and is used here because it is commonly quoted in a number of texts. It should be noted, however, that correlations smaller than 0.8 can also cause problems for the regression procedure.

[6] For reasons which will become clear in the chapter on logistic regression, it is not appropriate to model a binary response variable using OLS regression.

[7] Of course, someone who is designated ‘not female’ will be ‘male’ which allows a comparison to be made between male and female, even though we have, technically speaking, only coded for the presence of one of the categories.

[8] An analysis of the full data set is shown in Section *3.4*.

[9] Although this variable is not used in the present analysis, it is mentioned here as it is used in Chapter 4 to demonstrate logistic regression.

[10] The three location dummy variables have been kept in the model as a group (i.e., *j* – 1 variables) to enable comparison with the reference category.

http://dx.doi.org/10.4135/9780857028075.d49