

## Chapter 2

# Other Useful Topics in Regression

Once the user has grasped the fundamentals of multivariate linear regression, there are several related techniques that have application in business-orientated or socio-economic fields and which are described in this and the next chapter. Problems involving a *binary* (or *dichotomous*) response variable are common in any field where the researcher wishes to predict the presence or absence of a characteristic or outcome based upon a set of predictor variables. An example that is often cited is the presence or otherwise of coronary heart disease (binary, response variable) being dependent on smoking habits, diet, alcohol use and levels of exercise. Whether a person is accepted for a bank loan could depend on their income, employment history, monthly out-goings, amount of other loans etc. Problems involving a binary response variable may be examined by *binary logistic regression* as described in the first section of this chapter (2.1). The section that follows (2.2) briefly describes the method of *Multinomial Logistic Regression* which is similar to Logistic Regression, save that it is more general in that the response variable is not restricted to two categories.

The variables incorporated as predictors into our regression studies so far are *quantitative*; they have a well-defined scale of measurement. Sometimes, it is necessary to use *qualitative* or *categorical* variables as predictors in the regression. Examples of such categorical variables are employment status (employed or not), industrial shift worked (day, evening or night), sex (male or female) etc. Such variables have no natural scale of measurement. They are represented in statistical analyses by codes (e.g. male coded as '0' and female coded as '1') as discussed in the first volume of this guide. It is perfectly reasonable to postulate that a response variable might be related to a mix of quantitative and qualitative variables. For example, an employee's present salary might depend on age, gender, previous experience and level of education. Such problems are examined by *Dummy Regression* illustrated in the Sect. 2.3 while Sect. 2.4 discusses functional forms of regression models.

## 2.1 Binary Logistic Regression

The standard regression model assumes that the dependent variable,  $Y$ , is measured quantitatively. The independent (or regressor) variables,  $X_i$ , may be measured quantitatively or qualitatively. (A dummy regressor is an example of a variable that is measured qualitatively). Binary logistic models apply to situations where the dependent variable is dichotomous in nature, taking a 0 or 1 value. For example, the dependent variable,  $Y$ , could be whether or not a person is unemployed (“employed” = 1, “unemployed” = 0). The regressors could include  $X_1$  the average national wage rate,  $X_2$  the individual’s education,  $X_3$  the national unemployment rate,  $X_4$  family income etc. The question arises as to how we handle models involving dichotomous dependent variables.

### 2.1.1 The Linear Probability Model (LPM)

To fix ideas, consider the following simple model:

$$\hat{Y} = \beta_1 + \beta_2 X$$

where  $X$  is family income (£ 000’s) and  $Y$  is dichotomous, such that  $Y = 1$  if the family owns a house and  $Y = 0$  if the family does not own a house. Models such as the above which express the dichotomous  $Y$  as a linear function of the regressor variable(s)  $X$  are called *linear probability models*. However, there are problems with the assumptions that underpin regression when applying ordinary least squares to linear probability models.

(a) The residuals are not normally distributed. To see this:

$$\begin{aligned}\text{Residual} &= Y - \hat{Y} = Y - \beta_1 - \beta_2 X \\ \text{When } Y = 1, \text{Residual} &= 1 - \beta_1 - \beta_2 X \\ \text{When } Y = 0, \text{Residual} &= -\beta_1 - \beta_2 X.\end{aligned}$$

Consequently, the residuals cannot follow the normal distribution. (In fact, they are binomially distributed).

- (b) It can no longer be maintained that the residuals are homoscedastic. It can be shown that the variance of the residuals depends on the value taken by  $X$  and is thus not homoscedastic.
- (c) Consider the data in Fig. 2.1. Suppose a variable  $Y$  pertaining to home ownership is defined as above. When regression is applied to this LPM, we would obtain a result that:

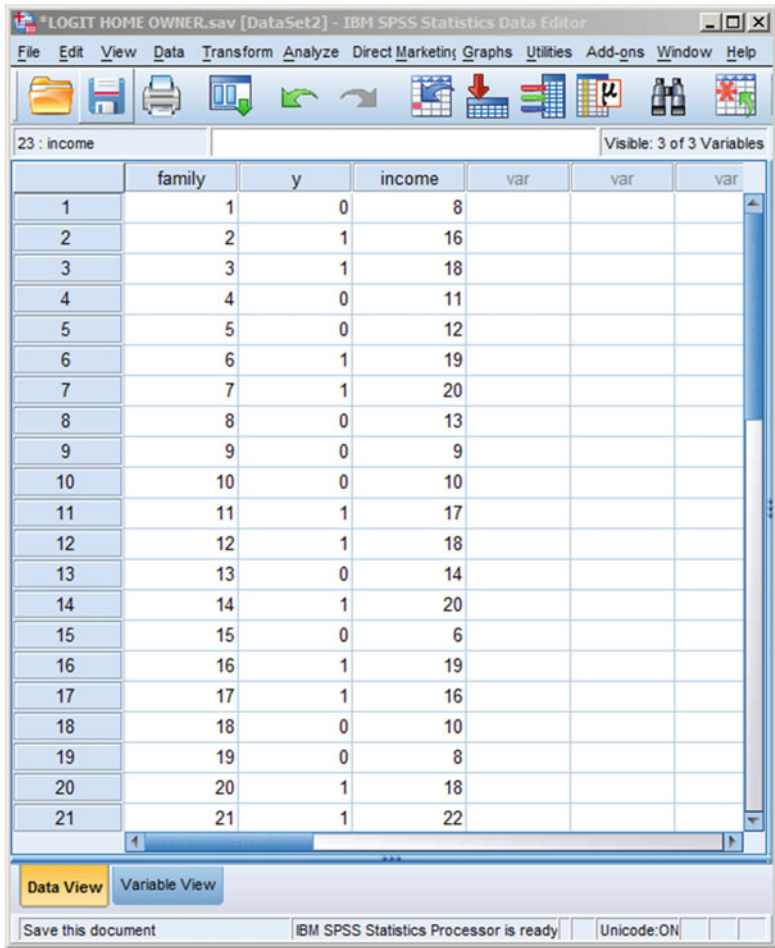
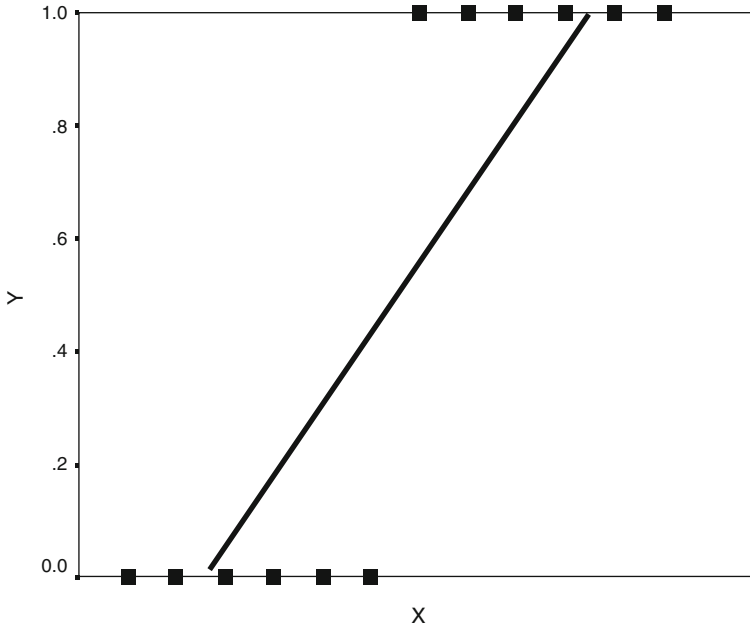


Fig. 2.1 Home ownership and income (£ 000's)

$$\hat{Y} = -0.874 + 0.098 (\text{INCOME})$$

LPM's want to treat  $\hat{Y}$  like a probability. For example if for a particular income level,  $\hat{Y} = 0.93$ , then we would guess that that family would be a home owner since the obtained result is closer to  $Y = 1$  than it is to  $Y = 0$ . However and continuing this theme, if a family had an income of £12,000 (i.e.  $X = 12$  in the LPM), then the predicted value of  $Y$  would be negative i.e. we would have negative probabilities. Indeed, it is possible to have an income level that coincides with a probability of home ownership in excess of 1. Consequently, the linear probability model is not recommended when the dependent variable is dichotomous.



**Fig. 2.2** Regression line when  $Y$  is dichotomous

- (d) The value of the coefficient of determination as a measure of goodness of fit becomes questionable. Corresponding to a given value of income ( $X$ ),  $Y$  is either 0 or 1. Therefore, all values of  $Y$  will either lie along the  $X$ -axis or along the line corresponding to  $Y = 1$  (see Fig. 2.2). Consequently, no linear probability model is expected to fit such a scatter well. The coefficient of determination is likely to be much lower than 100% for such models (even if the model is constrained to lie between  $Y = 0$  and  $Y = 1$ ).

There are ways to overcome some of the problems associated with the linear probability model. However, there remains a fundamental problem that is not very attractive because the model assumes that  $Y$  (or probability) increases linearly with  $X$ . This implies that the impact of  $X$  remains constant throughout. Thus, in the home ownership example, we find that as  $X$  increases by a unit (£1000), the probability of home ownership increases by 0.098. This is the case whether income is £8000, £80,000 or £800,000. This seems patently unrealistic. At a very low income, a family will not own a house. At a sufficiently high income say  $X^*$ , people will be most likely to own a house. Beyond  $X^*$ , income will have little effect on the probability of owning a home. Thus at both ends of the income distribution, the probability of owning a home will be virtually unaffected by a small increase in  $X$ . The probability of owning a home is nonlinearly related to income.

### 2.1.2 The Logit Model

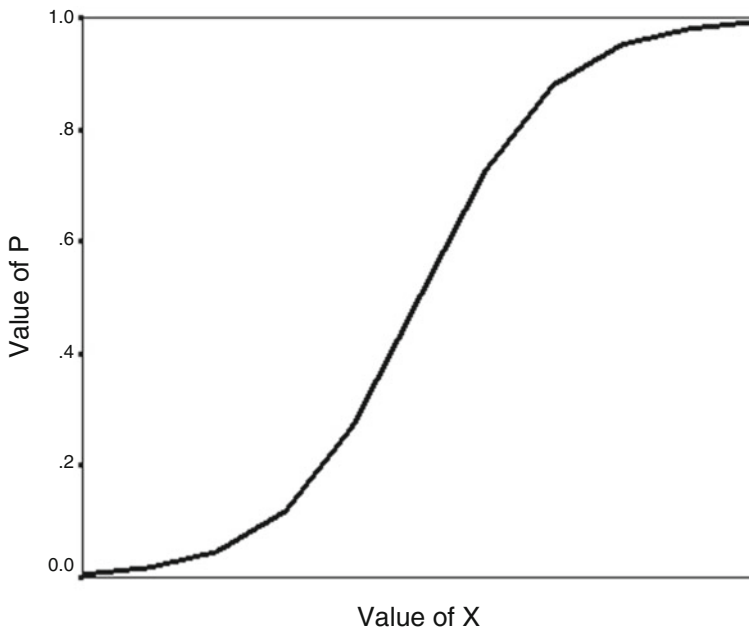
Now consider the following representation for home ownership, in which  $P$  represents the probability that a family owns a home i.e.  $P(Y = 1)$ :

$$P = \frac{1}{1 + \exp - (\beta_1 + \beta_2 X)} \dots \dots \quad (2.1)$$

in which  $\exp. - (X) = e^X$ . Equation (2.1) is called *the logistic distribution function*, which is plotted below.

As shown in Fig. 2.3, Eq. (2.1) permits  $P$  to range only between 0 and 1, thus solving one of the problems associated with the linear probability model. If  $P$  is the probability of owning a home, then  $(1 - P)$  is the probability of not owning a home and:

$$\begin{aligned} 1-P &= 1 - \frac{1}{1 + \exp - (\beta_1 + \beta_2 X)} = \frac{1 + \exp - (\beta_1 + \beta_2 X) - 1}{1 + \exp - (\beta_1 + \beta_2 X)} \\ &= \frac{\exp - (\beta_1 + \beta_2 X)}{1 + \exp - (\beta_1 + \beta_2 X)} = \frac{1/\exp(\beta_1 + \beta_2 X)}{1 + 1/\exp(\beta_1 + \beta_2 X)} = \frac{1/\exp(\beta_1 + \beta_2 X)}{\exp(\beta_1 + \beta_2 X) + 1/\exp(\beta_1 + \beta_2 X)} \\ &= \frac{1}{1 + \exp(\beta_1 + \beta_2 X)} \dots \dots \end{aligned} \quad (2.2)$$



**Fig. 2.3** A plot of the logistic distribution function

Therefore, using Eqs. (2.1) and (2.2), we can write:

$$\begin{aligned}\frac{P}{1-P} &= \frac{1}{1 + \exp - (\beta_1 + \beta_2 X)} \cdot [1 + \exp(\beta_1 + \beta_2 X)] \\ \frac{P}{1-P} &= \frac{1}{\frac{[\exp(\beta_1 + \beta_2 X) + 1]}{\exp(\beta_1 + \beta_2 X)}} \cdot [1 + \exp(\beta_1 + \beta_2 X)] \\ \frac{P}{1-P} &= \exp(\beta_1 + \beta_2 X)\end{aligned}$$

and taking natural logarithms (i.e. base e):

$$\begin{aligned}\ln\left(\frac{P}{1-P}\right) &= \ln[\exp(\beta_1 + \beta_2 X)] \\ \ln\left(\frac{P}{1-p}\right) &= \beta_1 + \beta_2 X \dots\dots\end{aligned}\tag{2.3}$$

because  $\ln(e^X) = X \ln e = X$ .

The left hand side of Eq. (2.3) is called the *logit* and the whole equation is called the *logit model*. The left hand side is the logarithm of the probability that a family owns a home against the probability that it does not. This is called the *logarithm of the odds ratio*. Naturally the logit model of Eq. (2.3) may be extended to the multivariate case:

$$\ln\left(\frac{P}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots\dots$$

### 2.1.3 Applying the Logit Model

The logit model of Eq. (2.3), where X is income (in £000's), was applied to the data in Fig. 2.1. (Computer packages use a method called “maximum likelihood” to generate the logit coefficients). The resultant model was:

$$\ln\left(\frac{\hat{P}}{1-\hat{P}}\right) = -1.6587 + 0.0792(\text{INCOME}) \dots\dots\tag{2.4}$$

The first family in Fig. 2.1 had an income of £8000 ( $X = 8$ ). Inserting this value of X into Eq. (2.4):

$$\ln \left( \frac{\hat{P}}{1 - \hat{P}} \right) = -1.0251, \text{ whereby } \frac{\hat{P}}{1 - \hat{P}} = e^{-1.0251} = 0.3588.$$

$$\text{Hence, } \hat{P} = 0.3588 - 0.3588\hat{P}$$

$$1.3588\hat{P} = 0.3588$$

$$\hat{P} = 0.2641.$$

The logit model estimates that there is a probability of 0.2641 that this family owns its home. It is possible to compute the change in probability of owning a home associated with a one unit (£1000) increase in income for this family who currently earn £8000. The change in probability is given by:

$$\hat{\beta}_2 \cdot \hat{P} (1 - \hat{P}) = (0.0792) * (0.2641) * (0.7359) = 0.0139.$$

If this family's income increases by £1000, there is an extra 1.39% chance that they will become a house owner. This extra probability is not constant, but varies with income level. The former was a disadvantage of the linear probability model.

### 2.1.4 The Logistic Model in IBM SPSS Statistics

An early, classic application of the logit model was in examining the choice of fertiliser used by Philippine farmers. The data are in the IBM SPSS Statistics data file called FERTILISER.SAV. The dependent variable to be explained is FERUSE – a binary variable equal to one if fertiliser is used and equal to zero otherwise. The explanatory variables are:

- CREDIT – the amount of credit (per hectare) held by the farmer,
- DMARKET – the distance of the farm to the nearest market,
- HOURMEET – no. of hours the farmer spent with an agricultural expert,
- IRSTAT – a dummy variable = 1 if irrigation is used, = 0 otherwise and
- OWNER – a dummy variable = 1 if the farmer owns the land, = 0 otherwise.

(There is an extra variable in this file called QFER, which records the amount of fertiliser used if the farmer indeed uses it). Four hundred and ninety one farms were examined. Binary logistic regression is accessed via:

```
Analyze
  Regression
    Binary logistic...
```

which generates the *Logistic Regression* dialogue box of Fig. 2.4.

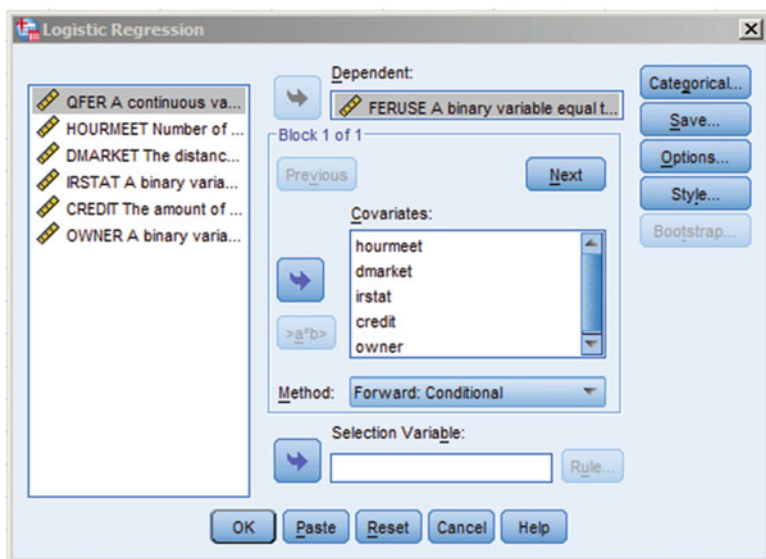


Fig. 2.4 The logistic regression dialogue box

The binary variable FERUSE is the Dependent variable and the five independent variables above are called Covariates in the context of logistic regression. Note that in the 'Method' box, the user can choose the Enter method (all independent variables entered simultaneously, Forward selection or Backward removal).

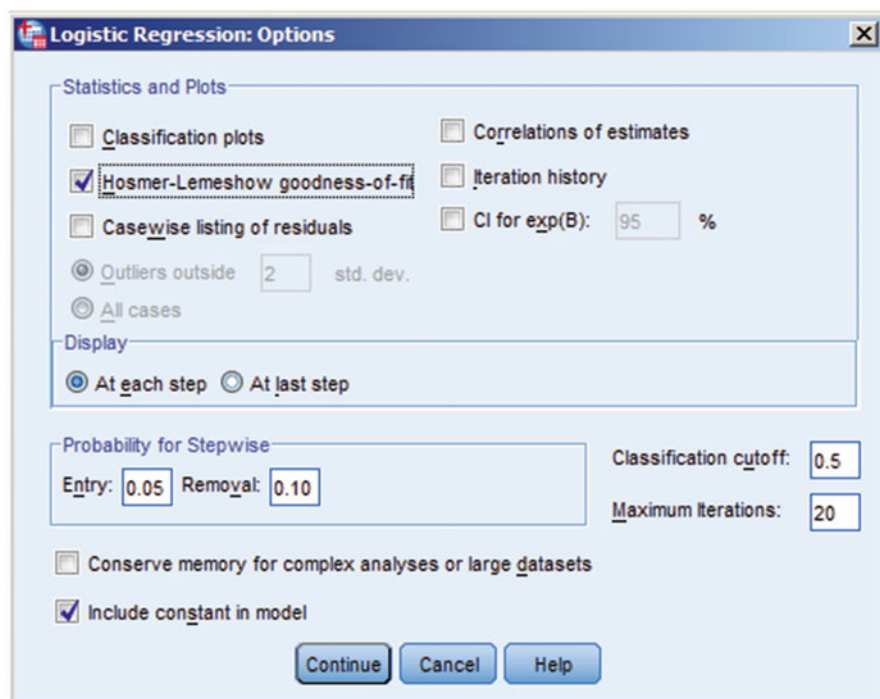
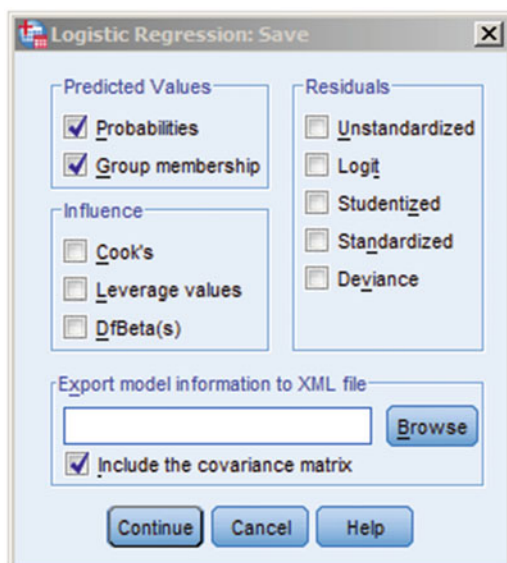
Clicking the Save... button generates the *Logistic Regression: Save* dialogue box of Fig. 2.5. This dialogue box permits the user to save the probabilities of group membership. If that probability is in excess of 0.5, the associated case is classified as being a member of the group that is coded as '1'; if that probability is less than 0.5, the case is deemed to be a member of the group coded as '0'. Standardized and unstandardized residuals can also be added to the active data file.

Clicking the Options... button in Fig. 2.4 produces the dialogue box of Fig. 2.6, where the Hosmer-Lemeshow test of model adequacy should be selected. This is discussed below. Note that a cut-off probability (Classification cutoff) of 0.5 is selected in Fig. 2.6: if the probability is above 0.5 group '1' membership is predicted and vice versa. Upon clicking the Continue and OK buttons, the results from the logistic regression are generated. Figure 2.7 presents the results for the first six farmers in the data file.

Based on the logistic analysis using the five covariates listed on page 33, the first farmer has a probability of 0.23519 of being in the group coded as '1' i.e. a fertilizer user. This probability appears under the heading PRE\_1. Since this probability is less than 0.5, the logistic model predicts that the farmer will not be a fertilizer user and therefore allocates him to the group coded as '0', shown under the heading PGR\_1. Examination of the FERUSE variable shows that he is indeed a group '0' member i.e. a non-fertilizer user. The second farmer has a probability of 0.22654 of



**Fig. 2.5** The logistic regression: save dialogue box



**Fig. 2.6** The logistic regression: options dialogue box

	qfer	hourmeet	dmarket	irstat	credit	owner	feruse	PRE_1	PGR_1	var
1	0	0	.5	0	0	0	0	23519	0	
2	0	0	1.5	0	0	0	0	22654	0	
3	167	472	1.7	0	460	1	1	1.00000	1	
4	0	40	6.0	1	0	1	0	.84800	1	
5	0	7	7.5	1	0	0	0	.53932	1	
6	0	3	7.4	0	167	0	0	.20245	0	

**Fig. 2.7** The first six cases in the active data file

being a fertilizer user and he is correctly classified. However, the fourth farmer has a probability of 0.84800 of being a fertilizer and he is incorrectly classified under the heading PGR\_1 as a group '1' member whereas he is a FREUSE = '0' member.

Figure 2.8 presents part of the output presented in the IBM SPSS STATISTICS Viewer. Via forward entry, the covariates are entered in the order of their importance. At step 1, the most important determinant of fertilizer usage is whether or not the farmer uses irrigation (IRSTAT). If the farmer was forward-thinking in using irrigation, he was also forward-thinking in applying chemical fertilizer. At step five, all the covariates are in the model and are significant (at  $p < 0.05$ ). Note that the frictional force of distance has a negative impact on fertilizer use as exemplified by the negative coefficient attached to the variable DMARKET.

From Fig. 3.7, the equation of the logistic model is:

$$\ln\left(\frac{P}{1-p}\right) = -1.155 + 0.557(OWNER) + 1.480(IRSTAT) + \dots + 0.0004(CREDIT),$$

from which the probabilities of group membership may be computed. For example, if for one particular farmer, HOURMEET = 30, DMARKET = 6, CREDIT = 200, IRSTAT = 1 and OWNER = 1, then from the above equation:

$$\begin{aligned}\ln\left(\frac{P}{1-p}\right) &= 1.507, \\ \frac{P}{1-P} &= e^{1.507} = 4.513 \\ P &= 4.513 - 4.513P \\ 5.513P &= 4.513 \\ \text{whereby } P &= 0.819.\end{aligned}$$

There is an over 80% chance that this particular farmer is a fertiliser user. Assessment of the logistic model's forecasting adequacy may be made by

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	irstat	1.601	.196	66.517	1	.000	4.957
	Constant	-.998	.144	48.002	1	.000	.369
Step 2 <sup>b</sup>	hourmeet	.034	.013	7.040	1	.008	1.035
	irstat	1.540	.199	59.715	1	.000	4.664
Step 3 <sup>c</sup>	Constant	-1.084	.148	53.360	1	.000	.338
	hourmeet	.031	.013	6.241	1	.012	1.032
	irstat	1.489	.201	54.733	1	.000	4.432
	credit	.000	.000	5.401	1	.020	1.000
	Constant	-1.158	.153	57.273	1	.000	.314
	hourmeet	.030	.013	5.683	1	.017	1.030
	irstat	1.521	.204	55.681	1	.000	4.579
	credit	.000	.000	6.314	1	.012	1.000
	owner	.522	.207	6.373	1	.012	1.686
	Constant	-1.392	.183	57.695	1	.000	.249
Step 5 <sup>e</sup>	hourmeet	.028	.012	5.362	1	.021	1.029
	dmarket	-.049	.022	4.695	1	.030	.952
	irstat	1.480	.205	52.067	1	.000	4.394
	credit	.000	.000	7.031	1	.008	1.000
	owner	.557	.209	7.109	1	.008	1.745
	Constant	-1.155	.208	30.897	1	.000	.315

a. Variable(s) entered on step 1: irstat.

b. Variable(s) entered on step 2: hourmeet.

c. Variable(s) entered on step 3: credit.

d. Variable(s) entered on step 4: owner.

e. Variable(s) entered on step 5: dmarket.

**Fig. 2.8** Variables in the final logistic model

examining the 'Classification Table' of Fig. 3.8 and which is part of the output in the IBM SPSS Statistics Viewer. At step 5 in the above table, 266 farmers are in the dependent variable = 0 category i.e. the FERUSE = 0 group – they do not use fertilizer. One hundred and eighty three of these farmers were predicted by the logistic model to have a probability of fertilizer use below the cut-off probability of 0.5. Hence, 183 (68.80%) of the farmers in the FERUSE = 0 group were predicted correctly. Therefore, 31.20% of the FERUSE = 0 group were incorrectly classified by the logistic model (Figs. 2.9 and 2.10).

Similarly, there were 225 farmers observed to be in the dependent variable = 1 category i.e. FERUSE = 1.159 (70.67%) farmers had probabilities above the cut-off point of 0.5 and were consequently correctly classified. Overall, 342 farmers (183 + 159) have been correctly classified into their FERUSE = 0 or FERUSE = 1 groups. This is an overall success rate of 342 out of 491 farmers or 69.7%.

Classification Table <sup>a</sup>					
Observed			Predicted		
			FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise		Percentage Correct
			0	1	
Step 1	FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise	0	179	87	67.3
		1	66	159	70.7
	Overall Percentage				68.8
Step 2	FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise	0	177	89	66.5
		1	63	162	72.0
	Overall Percentage				69.0
Step 3	FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise	0	176	90	66.2
		1	61	164	72.9
	Overall Percentage				69.2
Step 4	FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise	0	175	91	65.8
		1	61	164	72.9
	Overall Percentage				69.0
Step 5	FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise	0	183	83	68.8
		1	66	159	70.7
	Overall Percentage				69.7

a. The cut value is .500

Fig. 2.9 The classification table associated with logistic regression

Fig. 2.10 The Hosmer-Lemeshow test

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	.000	0	.
2	11.786	5	.038
3	27.492	6	.000
4	18.258	7	.011
5	7.156	8	.520

Beside the Classification Table, another assessment of the adequacy of the logit model is the Hosmer-Lemeshow (HL) goodness of fit test. The HL test has as its null  $H_0$ : the model adequately predicts group membership and the null is rejected if the associated level of significance is less than 5% or 0.05. In the above example, it is found that  $HL = 7.156$  with significance 0.520, so the null would not be rejected and the logit model deemed an adequate representation for the data.

### 2.1.5 A Financial Application of the Logistic Model

The logistic model was first in Finance to predict the probability that a given firm will be a merger target. The code ‘1’ is used if a firm was a merger target and ‘0’ if it was not. The subsequent logistic model can presented as follows:

$$\ln \left( \frac{P}{1-p} \right) = \beta_1 + \beta_2(\text{PAYOUT}) + \beta_3(\text{TURNOV}) + \beta_4(\text{SIZE}) + \beta_5(\text{LEV}) + \beta_6(\text{VOL})$$

where:

PAYOUT = payout ratio (dividend/earnings),

TURNOV = asset turnover (sales/total asset),

SIZE = market value of equity,

LEV = leverage ratio (long-term debt/total assets) and

VOL = trading volume in the year of acquisition.

$\beta_2$ ,  $\beta_4$  and  $\beta_5$  are expected to be negative and  $\beta_6$  to be positive while  $\beta_3$  to be positive or negative. Based on a sample of 24 merged firms (coded as ‘1’) and 43 non-merged firms (coded as ‘0’), the results shown in Table 2.1 were obtained:

The estimated coefficients had the expected signs and all but two were statistically significantly different from zero. The results, for example, illustrate that the higher the turnover and the larger the size, the lower are the odds of the firm being a takeover target. On the other hand, the higher the trading volume, the greater the odds of being a merger candidate, for high-volume firms may imply lower acquisition transaction costs due to marketability. Based on these analyses, we conclude that one of the important factors affecting the firm’s attractiveness is the inability of managers to generate sales per unit of assets. Moreover, low turnover must be accompanied by any one or a combination of low payout, low financial leverage, high trading volume and smallness in aggregate market value in order to produce a high probability of merger.

**Table 2.1** Logistic estimate results for the Dietrich and Sorenson study

Variable	Coefficient	Standard error	t-value
PAYOUT	−0.74	0.29	−2.51**
TURNOV	−11.64	3.86	−3.01**
SIZE	−5.74	2.39	−2.40**
LEV	−1.33	0.97	−1.37
VOL	2.55	1.58	1.62
Intercept	−10.84	3.40	−3.20**

\*\* Significant at  $p < 0.01$

## 2.2 Multinomial Logistic Regression

Now we consider situations in which the response variable has more than the two categories of the binary case. Variables with more than two categories are called *polychotomous* rather than dichotomous. Such situations are analysed by multinomial logistic regression which is very similar to the binary logistic regression of the previous section, save that it is more general since the dependent or response variable is not restricted to two categories. For example, a survey of opinions about a proposed road improvement scheme could produce responses (Y) of “against” (code ‘0’, say), “undecided” (code ‘1’) or “in favour” (code ‘2’). Multinomial logistic regression could be used to see if Y might depend on the resident’s proximity to the road ( $X_1$ ), the resident’s age ( $X_2$ ), whether or not the resident has children (a categorical variable  $X_3$ ) etc. As another example, in a Marketing scenario, a firm might be examining consumer attitudes towards several types of product packaging (the polychotomous response variable). Such attitudes may well depend on consumers’ income levels, their age, purchase purpose etc. I do not propose presenting an example on multinomial regression in IBM SPSS Statistics as it is so similar to the binary example, plus it does require a deeper statistical knowledge for the fullest use of the method.

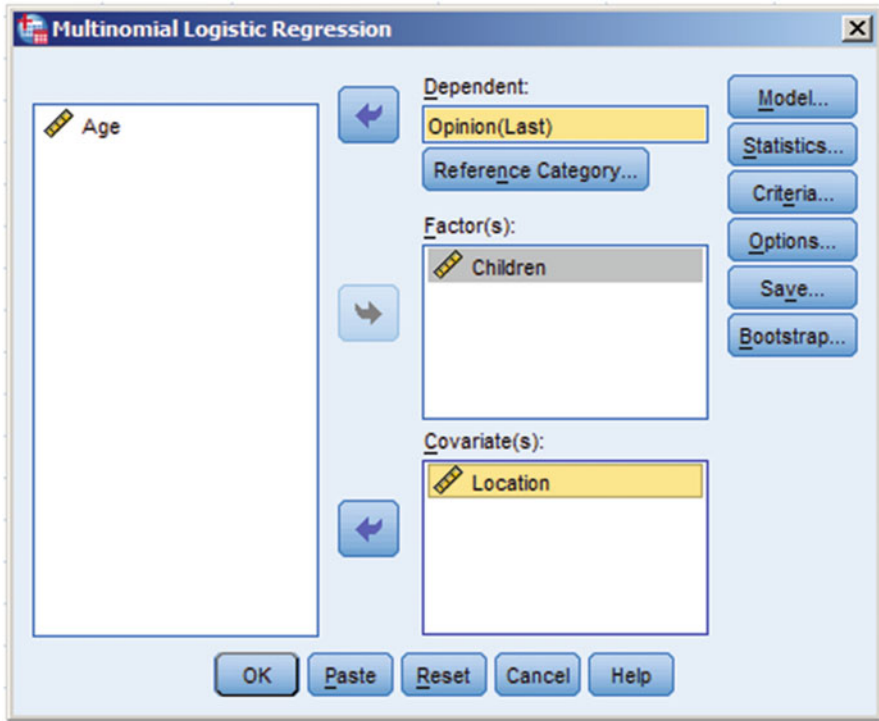
Suffice it to say, that multinomial logistic regression is accessed via:

```
Analyze  
  Regression  
    Multinomial logistic...
```

which gives rise to the *Multinomial Logistic Regression* dialogue box of Fig. 2.11. In the example of attitudes to the road scheme, OPINION is the dependent variable. The dialogue box of Fig. 2.11 requires ‘Factors’ and/or ‘Covariates’. Covariates are simply the quantitative variables (continuous measurement) in the analysis, such as LOCATION the distance of the resident from the road. Factors are categorical variables like CHILDREN – whether or not the resident has children. Such a categorical variable might be coded as ‘0’ and ‘1’. The analysis proceeds as in the binary case. Predicted probabilities, group membership, raw and standardized residuals and goodness of fit statistics may be generated and/or saved.

## 2.3 Dummy Regression

Dummy variables are most commonly used when a researcher wants to insert nominal scale categorical variables into a regression equation. A set of dummy variables is created by treating each category of a categorical variable as a separate variable and assigning arbitrary scores for all cases depending on their presence or absence in each category. Suppose that an engineer wishes to relate the effective



**Fig. 2.11** The multinomial logistic regression dialogue box

life time (Y) of a cutting tool used on a lathe to lathe speed in revolutions per minute ( $X_1$ ) and the type of cutting tool used ( $X_2$ ). This second variable is categorical and has two levels, tool types A and B. We may invoke a *dummy variable*  $X_2$  with the codes:

$X_2 = 0$  if the observation is from tool type A

$X_2 = 1$  if the observation is from tool type B

The choice of '0' and '1' to identify the levels of qualitative variable is arbitrary. We thus have a model of the form:

$$Y = b_0 + b_1X_1 + b_2X_2 + e$$

Where  $e$  represents the error term. To interpret the coefficients in the model, consider first tool type A, for which  $X_2 = 0$ . The regression model becomes:

$$Y = b_0 + b_1X_1 + e$$

The relationship between tool life and lathe speed for tool type A is a straight line with intercept  $b_0$  and gradient  $b_1$ . For tool type B,  $X_2 = 1$  and the regression model is:

$$Y = b_0 + b_1X_1 + b_2 + e$$
$$Y = (b_0 + b_2) + b_1X_1 + e$$

That is, for tool type B, the relationship between tool life and lathe speed is a straight line with gradient  $b_1$ , but with intercept  $(b_0 + b_2)$ . These two responses describe two parallel regression lines with different intercepts.

One may generalise this approach to qualitative factors with any number of levels. Suppose there were three tool types, then two dummy variables are necessary:

$X_2$	$X_3$	
0	0	If the observation is from tool type A
1	0	If the observation is from tool type B
0	1	If the observation is from tool type C

And the regression model is:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e$$

Generally, a qualitative categorical variable with  $L$  levels is represented by  $(L - 1)$  dummy variables, each taking on the values 0 and 1.

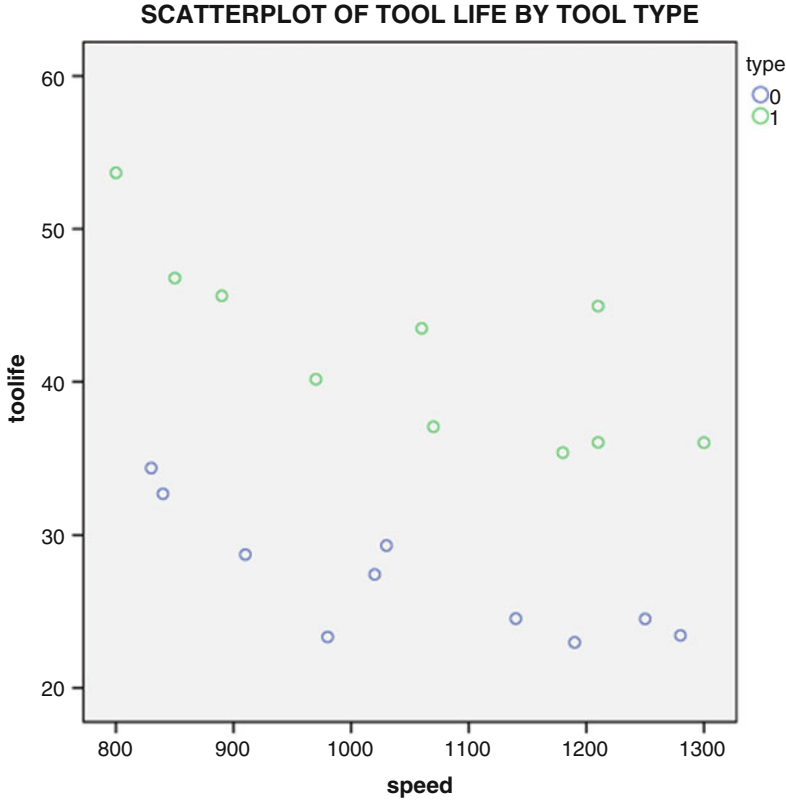
The file TOOLLIFE.SAV contains data about the lifetimes of the cutting tools (LIFETIME), the lathe (SPEED) and the type of cutting tool (TYPE). Figure 2.12 presents a scatterplot of the lifetimes of the two types of tool. This is created by selecting:

```
Graphs
  Legacy dialogs
    Scatter/Dot
      Simple Scatter
```

In the Scatterplot dialogue box, we want the two tool types plotted with different symbols, so under the heading ‘Set Markers By’ enter the variable TYPE. Inspection of the scatter diagram indicates that two different regression lines are required to model adequately these data, with the intercepts depending on the type of tool used. The dummy variable is coded as before,  $X_2 = 0$  if it is a type A tool and  $X_2 = 1$  if it is type B. The equation of regression is obtained in the usual manner:

```
Analyse
  Regression
    Linear
```





**Fig. 2.12** Scatterplot of tool life by tool type

The least squares fit using the Enter selection method is shown in Fig. 2.13 and is:

$$\widehat{TOOLLIFE} = 52.146 - 0.024 * SPEED + 14.957 * TYPE,$$

With  $r^2 = 0.900$ . The t statistics show that both regression coefficients are significantly different from zero. The negative coefficient associated with the variable SPEED makes sense, in that we expect TOOLLIFE to decrease as SPEED increases. The parameter (14.957) associated with TYPE is the change in mean TOOLLIFE resulting from a change from tool type A to tool type B. A 95% confidence interval could be selected for the equivalent population coefficient and it would be found to be:

$$11.879 < \beta_2 < 18.035$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.941 <sup>a</sup>	.885	.871	3.261

a. Predictors: (Constant), type, speed

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	52.416	4.917		10.659	.000
	speed	-.024	.005	-.433	-5.257	.000
	type	14.957	1.459	.845	10.254	.000

a. Dependent Variable: toolife

Fig. 2.13 Part of the output from dummy regression

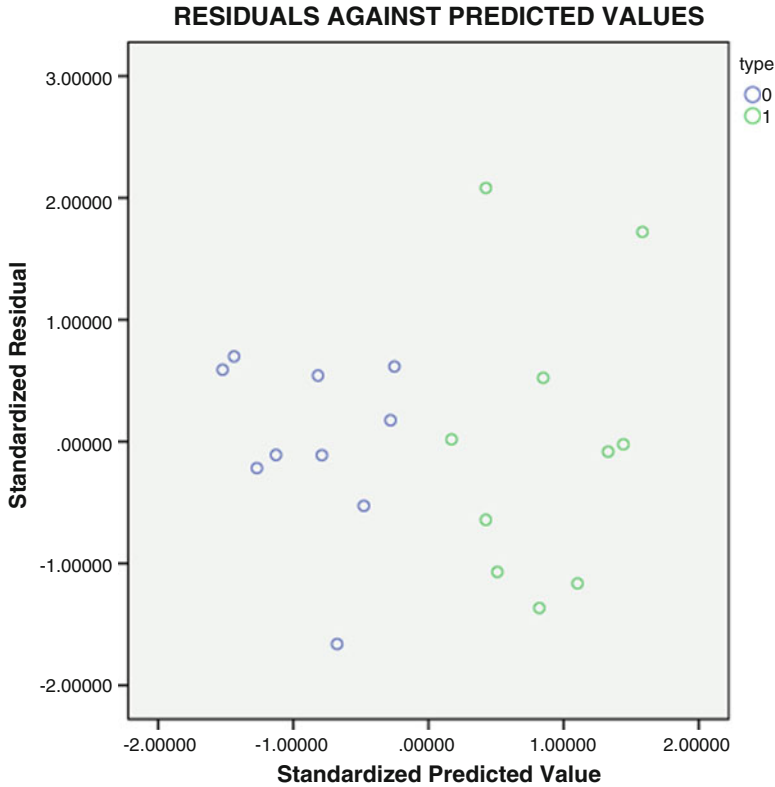
A plot of the residuals against the fitted values (both standardized) is shown in Fig. 2.14. These two variables were saved as part of the dummy regression procedure and the scatterplot constructed. The type B residuals in Fig. 2.14 exhibit slightly more scatter than those of type A, implying that there may be a mild inequality of variance problem. The normal probability plot revealed no model inadequacy in this respect.

Since two different regression lines are used to model the relationship between tool life and lathe speed, we could initially fit two separate straight line models instead of a single model with a dummy variable. However, the single-model approach is preferred because the analyst has only one equation to work with instead of two, a much simpler practical result. Furthermore, since both straight lines are assumed to have the same gradient, it makes sense to combine the data from both types to produce a single estimate of this common parameter. This approach also gives one estimate of the common residual variance,  $\sigma^2$ .

Suppose that we expect the regression lines relating tool life to lathe speed to differ in both intercept and gradient. It is possible to model this situation with a single regression equation by using dummy variables. The model is:

$$\text{TOOLLIFE} = b_0 + b_1 * \text{SPEED} + b_2 * \text{TYPE} + b_3 * \text{SPEED} * \text{TYPE}$$

We observe that a cross product term between lathe speed and the dummy variable (SPEED\*TYPE) has been added to the model. To interpret the coefficient  $b_3$  for this model, first consider a type A tool for which TYPE = 0, then the above model becomes:



**Fig. 2.14** A plot of residuals against predicted values

$$\text{TOOLLIFE} = b_0 + b_1 * \text{SPEED},$$

A line with intercept  $b_0$  and gradient  $b_1$ . For tool type B,  $\text{TYPE} = 1$  and our model becomes:

$$\text{TOOLLIFE} = (b_0 + b_2) + (b_1 + b_3) * \text{SPEED}$$

A line with intercept now of  $(b_0 + b_2)$  and gradient now of  $(b_1 + b_3)$ . Hence, the parameter  $b_2$  reflects the change in intercept associated with changing from tool type A to tool type B and  $b_3$  indicates the change in gradient associated with changing from tool type A to tool type B. Fitting this model is equivalent to fitting two separate regression equations. An advantage to the use of the dummy variable is that tests of hypotheses may be performed directly. For example, to test whether two regression lines have the same intercept but possibly different gradients, then by reference to the above equation, we should examine:

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

To test that the two regression lines have a common gradient, but possibly different intercepts, the hypotheses are:

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0$$

It is simple matter to compute the cross product term (variable name CROSSPRO, say) = SPEED.TYPE via:

Transform  
Compute

Which generate the Compute Variable dialogue box of Fig. 2.15. Operationalising, the new variable CROSSPRO is added to the working file as shown in Fig. 2.16. Part of the results of running the dummy regression with the cross product term are shown in Fig. 2.17.

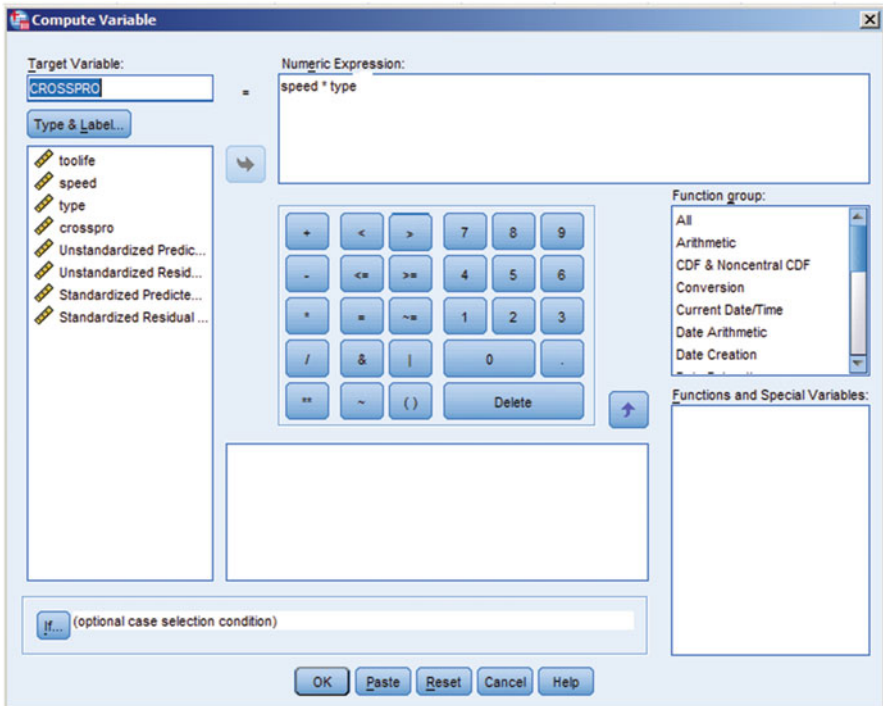


Fig. 2.15 Computation pf the cross-product term

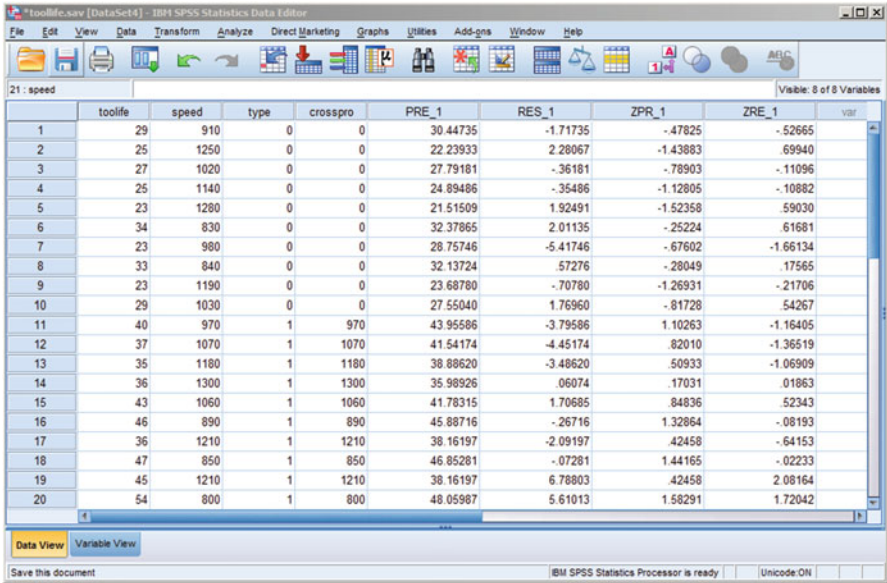


Fig. 2.16 The new data file

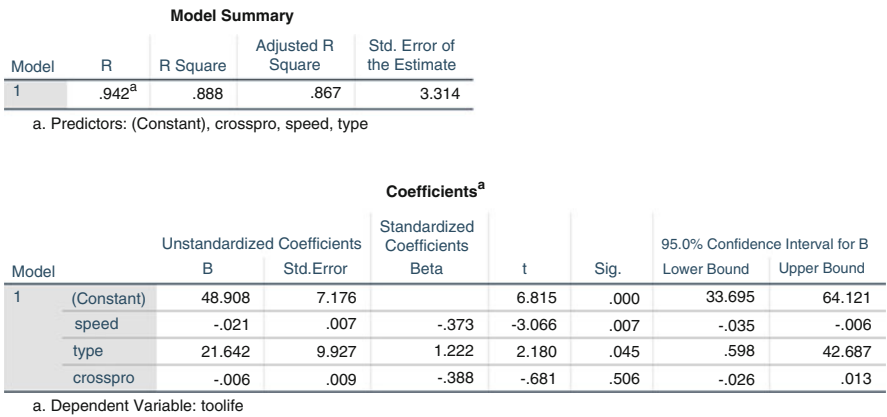


Fig. 2.17 Part of the output for dummy regression with a cross product term

## 2.4 Functional Forms of Regression Models

Several classes of model occur in finance, business and economics that are not linear in form. Such models can, however, be transformed into linear ones and ordinary least squares (OLS) applied. To understand these, it is necessary to remind ourselves of the following laws of logarithms that are valid regardless of the base used:

	y	x	YEAR_	DATE_
1	3.57	1.77	2006	2006
2	3.50	1.74	2007	2007
3	3.35	1.72	2008	2008
4	3.30	1.73	2009	2009
5	3.25	1.76	2010	2010
6	3.20	1.75	2011	2011
7	3.11	2.08	2012	2012
8	2.94	2.81	2013	2013
9	2.97	2.39	2014	2014
10	3.06	2.20	2015	2015
11	3.02	2.17	2016	2016

**Fig. 2.18** Raw data

- (a)  $\log(XY) = \log X + \log Y$
- (b)  $\log(X/Y) = \log X - \log Y$
- (c)  $\log(X)^n = n \cdot \log X$

In the above,  $X$  and  $Y$  are assumed to be positive and  $n$  is some constant.

Consider the data in Fig. 2.18 below which presents average annual coffee consumption ( $Y$ ; cups per day) in the London area in relation to average annual retail price ( $X$ ; £ per lb.).

The data are in the file COFFEE.SAV. These data were subjected to bivariate regression analysis and the results obtained are presented below in Fig. 2.19.

The two variable regression model is thus of the form  $\hat{Y} = 4.171 - 0.48x$  which indicates that if the average retail price increases by a pound, then the average consumption of coffee would reduce by nearly half a cup. The coefficient of determination indicated that about 66% of the variation in average daily coffee consumption is explained by a linear relationship with the retail price of coffee.

Coefficients <sup>a</sup>					
Model		Unstandardized Coefficients		Standardized Coefficients	t
		B	Std. Error	Beta	
1	(Constant)	4.171	.233		17.935
	x	-.480	.114	-.814	-4.206

a. Dependent Variable: y

Fig. 2.19 Bivariate regression results

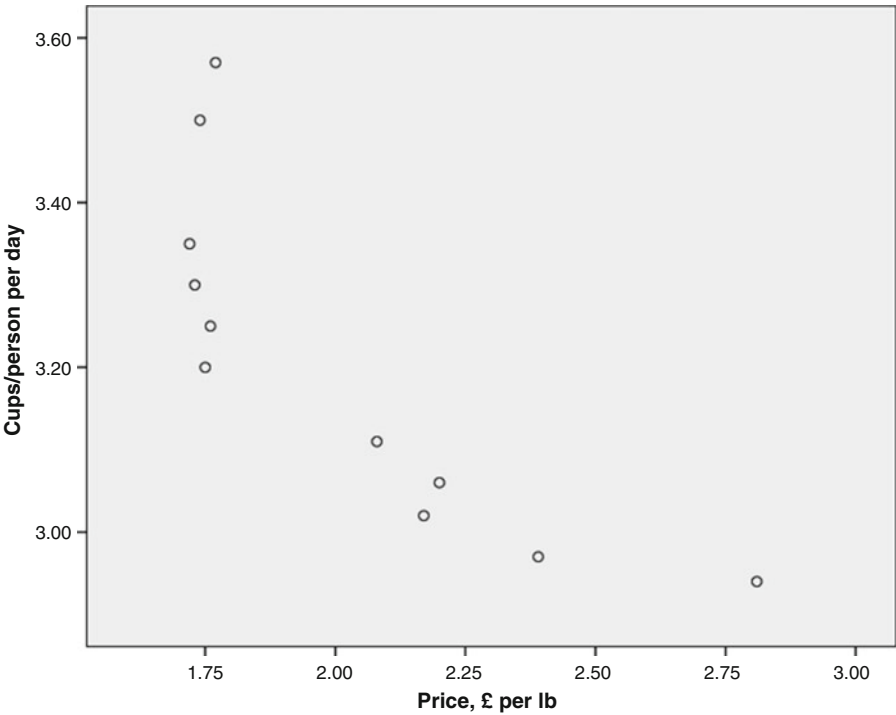


Fig. 2.20 A plot of average annual coffee consumption against average price

2.4.1 The Power Model

Figure 2.20 casts grave doubt as to whether the relationship between these two variables is a linear one. In fact, a more suitable model for the raw data in Fig. 2.18

$$\hat{Y} = \beta_1 X^{\beta_2} \dots \dots$$

(2.5)

in which  $\beta_1$  and  $\beta_2$  are parameters to be estimated. Eq. (2.5) is referred to as a **power model**. Eq. (2.5) may be expressed in a linear form by the simple expedient of taking natural logarithms (i.e. base e) to derive:

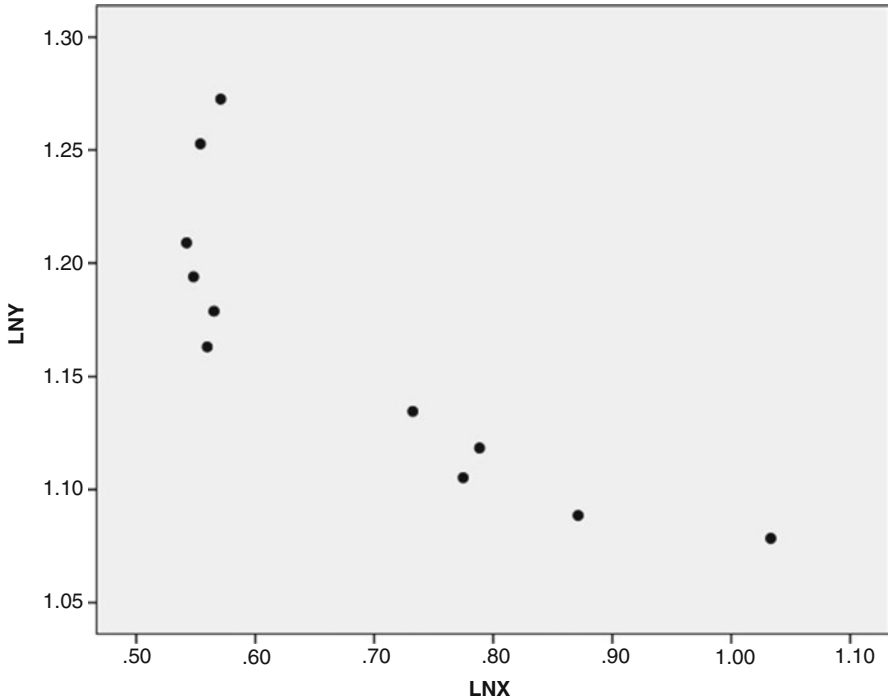
$$\ln Y = \ln \beta_1 + \beta_2 \ln X \dots\dots (2.6)$$

The fact that Eq. (2.6) is in linear form may be shown by letting  $Y^* = \ln Y$ ,  $\alpha = \ln \beta_1$  and  $X^* = \ln X$  to establish that:

$$Y^* = \alpha + \beta_2 X^* \dots\dots (2.7)$$

Ordinary least squares may now be used to estimate  $\alpha$  and  $\beta_2$  in Eq. (2.7). Knowing the value of  $\alpha$ , we can readily find  $\beta_1$  because  $\alpha = \ln \beta_1 \Rightarrow \beta_1 = e^\alpha$ . The linear Eq. (2.6) is known as a **log-log model**, a **double-log model** or perhaps most commonly, a **log-linear model**.

How can we assess if Eq. (2.5) is an adequate representation of the data in Fig. 2.18? If it is and from Eq. (2.6), a plot of  $\ln Y$  against  $\ln X$  should be linear or closely so. Figure 2.21 presents this plot, from which it is seen that the plot is not perfectly linear, but it seems to be more linear than the plot in Fig. 2.20.



**Fig. 2.21** A plot of  $\ln Y$  against  $\ln X$



Coefficients <sup>a</sup>						
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1	(Constant)	1.390	.049	28.489	.000	
	LN <sub>X</sub>	-.331	.069	-.847	-.4.771	.001

a. Dependent Variable: LN<sub>Y</sub>

**Fig. 2.22** Results of regressing lnY against lnX

Regressing lnY against lnX by means of OLS, the results in Fig. 2.22 are obtained:

The coefficient of determination of the model in Fig. 2.22 is 71.7%. *However, this is not directly comparable with the coefficient of determination obtained for the bivariate linear model in Fig. 2.19, since the models are different.* The power model of Eq. (2.5) may well be a more adequate representation of the raw data than is the purely linear model presented in Fig. 2.19.

A very attractive feature of the log-linear model is that the slope coefficient (or gradient)  $\beta_2$  in Eq. (2.6) measures **the elasticity** of the variable Y with respect to variable X. The elasticity of Y with respect to X is defined as the percentage change in Y for a given (small) percentage change in X. If Y represents the quantity of a commodity demanded and X is its unit price, then  $\beta_2$  measures what is called **the price elasticity of demand**:

$$\text{Price elasticity of demand} = \frac{\% \text{change in quantity demanded}}{\% \text{change in price}}$$

From Fig. 2.22, the results corresponding to Eq. (2.6) are:

$$\text{Ln}Y = 1.39 - 0.331 \text{ Ln}X \dots \dots \quad (2.8)$$

whereby the price elasticity coefficient is about  $-0.33$ . This implies that in the face of a 1% increase in the price of coffee, demand for coffee (as measured by cups of coffee consumed) decreases on average by 0.33%. When the price elasticity is less than 1 in absolute terms, we say that the demand for coffee is price inelastic; if the price elasticity exceeds 1, we say that the demand for coffee is price elastic.

The results in Eq. (2.8) may be used to compute the parameters of the power model of Eq. (2.5). Comparing Eq. (2.8) with Eq. (2.6), we find that  $\ln\beta_1 = 1.39$  which implies that  $\beta_1 = e^{1.39} = 4.014$  and that  $\beta_2 = -0.331$ . Inserting these values back into the original power model of Eq. (2.5), we establish that:

$$\hat{Y} = 4.014X^{-0.331}$$

This numerical result may be obtained directly in IBM SPSS Statistics by using the ‘Curve Estimation’ procedure, which is part of that package’s Regression routine.

The idea of taking a logarithmic transformation may be extended. Consider the compound interest formula and for simplicity let us suppose annual compounding for  $t$  years:

$$FV = PV(1 + r)^t \dots\dots \quad (2.9)$$

in which  $FV$  and  $PV$  are respectively future and present values of an investment and  $r$  is the annual interest rate. Taking logarithms:

$$\ln FV = \ln PV + t \ln (1 + r),$$

which is in linear form. This is called a **semilog model**, because only one variable ( $FV$ ) appears in logarithmic form. Again, this model has an important property. Here, the slope coefficient (or gradient) represents **the rate of growth of  $Y$** . If the gradient is negative, we have a **rate of decay**. Research has shown that the U.S GDP (in billions of dollars, constant prices) between 1998 and 2015 inclusive is well approximated by a semilog model. It has been established that:

$$\ln GDP = 9.164 + 0.039t$$

which suggests that GDP grew at a rate of 3.9% over this period. Note that in 1998,  $t = 0$ , so we estimate that  $\ln GDP = 9.164 \Rightarrow GDP = e^{9.164} = 9547$  billion dollars.

### 2.4.2 The Reciprocal Model

Models of the following type are known as **reciprocal models**:

$$Y = \beta_1 + \beta_2 \left( \frac{1}{X} \right) \dots\dots \quad (2.10)$$

in which  $\beta_1$  and  $\beta_2$  are parameters to be estimated. The model has the feature that as  $X$  increases indefinitely, the term  $\beta_2 \left( \frac{1}{X} \right)$  approaches zero, and  $Y$  therefore approaches the limit (or **asymptotic value**)  $\beta_1$ . Reciprocal models have built in them an asymptote or limit value that the dependent variable will take when the value of the  $X$  variable increases indefinitely.

The shape of the reciprocal model is shown in Fig. 2.23 in which the asymptote is represented by the horizontal line. In Fig. 2.23,  $\beta_1$  and  $\beta_2$  are both positive, non-zero. One application of the reciprocal model is the average fixed cost of production ( $Y$ ) against levels of output ( $X$ ). As  $X$  increases,  $Y$  decreases to a finite limit.

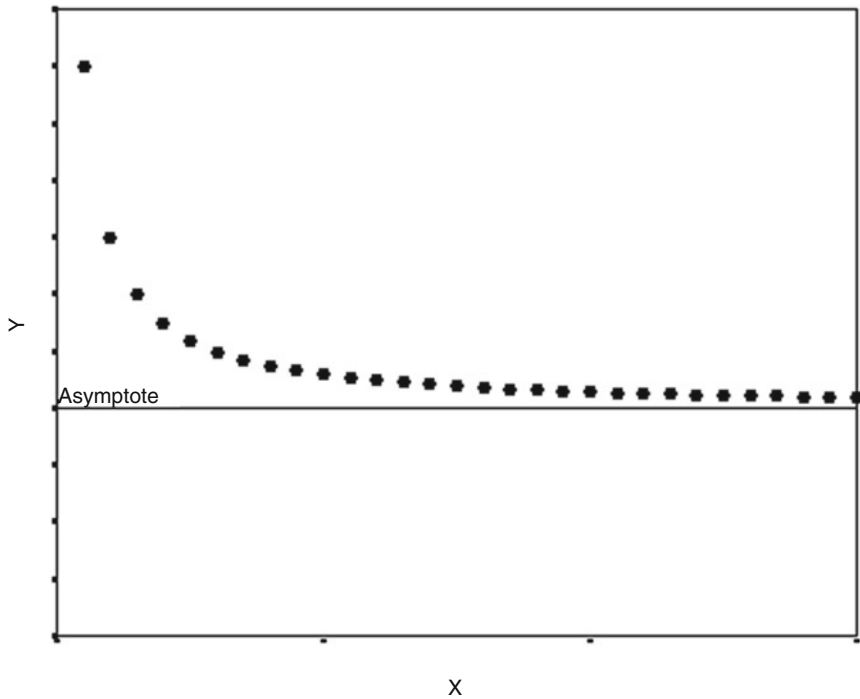


Fig. 2.23 The reciprocal model with asymptote

An important application of the reciprocal model is the **Phillips curve**. This is based on empirical observation by the economist A.W. Phillips of the relationship between the level of unemployment and the year to year increase or rate of change in wage rates. (Note that by inference, the rate of change in wages will impact on the rate of change in commodity prices or inflation). Figure 2.24 presents U.K. annual increases in wage rates (Y%) against unemployment (X%) from 2000 to 2016 inclusive. The data are available on the book webpage under the file PHILLIPS.SAV.

Upon regressing Y against  $1/X$ , the results below are obtained (Fig. 2.25):

The reciprocal model obtained is thus:

$$\hat{Y} = 1.232 + 5.63 (1/X) \dots\dots \quad (2.11)$$

Inherent in Eq. (2.11) is that as the unemployment rate increases, the % increase in wages declines. Like any equation, Eq. (2.11) cuts the X-axis when  $Y = 0$ , i.e.  $1.232/5.63 = 1/X = 0.2188 \Rightarrow X = 4.6\%$ . This is the rate of unemployment consistent with no change in wage rates, which theoretically means stable prices. This point is called **the non-accelerating inflation rate of unemployment (NAIRU)** or **the natural rate of unemployment**. The fact that the  $\beta_1$  is positive



2.4.3 The Linear Trend Model

Instead of the semilog model where  $\ln Y$  is regressed against time  $t$ , some researchers have advocated the linear trend model wherein  $Y$  is regressed against time  $t$ :

$$Y = \beta_1 + \beta_2 t \dots \dots \quad (2.12)$$

By **trend**, we mean any sustained upward or downward movement in the behaviour of a variable. If the slope coefficient or gradient  $\beta_2$  is positive, there is an upward trend in  $Y$ , whereas if  $\beta_2$  is negative there is a downward trend in  $Y$ . The data in Fig. 2.26 present the GDP of the United States in current billions of dollars, between 1972 and 1991 inclusive. The source is *The Economic Report of the President*, January 1993 and the figures are in the file GDPUSA.SAV on the file server.

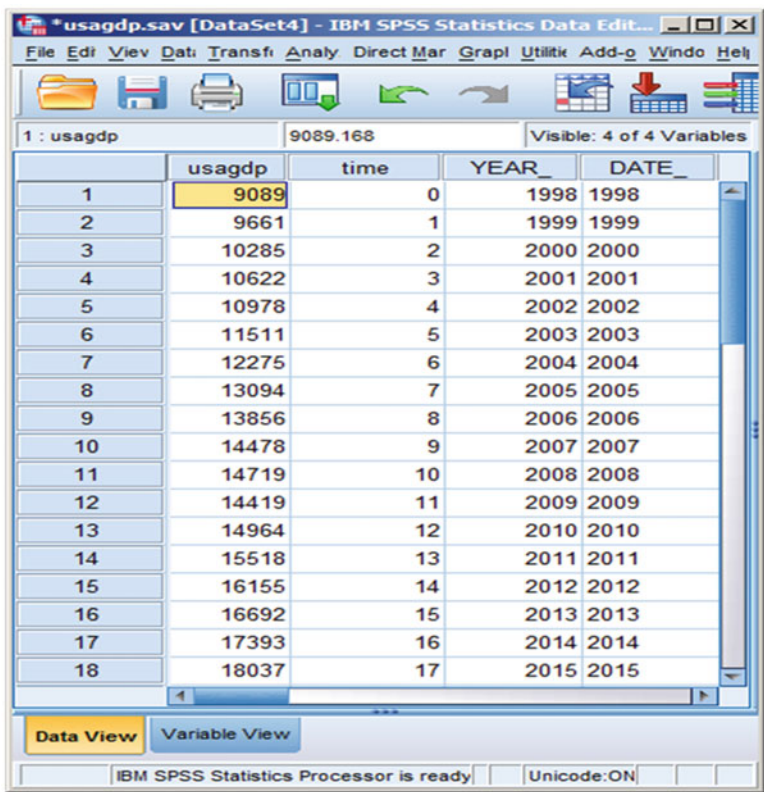


Fig. 2.26 United States GDP, 1972–1991

(The U.S. GDP data of the semilog section were at constant prices). Figure 2.27 presents a plot of these data over time. Note that the time variable,  $t$ , in Eq. (2.12) is given values from 0 to 17 inclusive. Figure 2.27 suggests that the data exhibit a reasonable trend with a positive gradient. The results of regressing GDP against time were:

so the derived linear trend model is (Fig. 2.28):

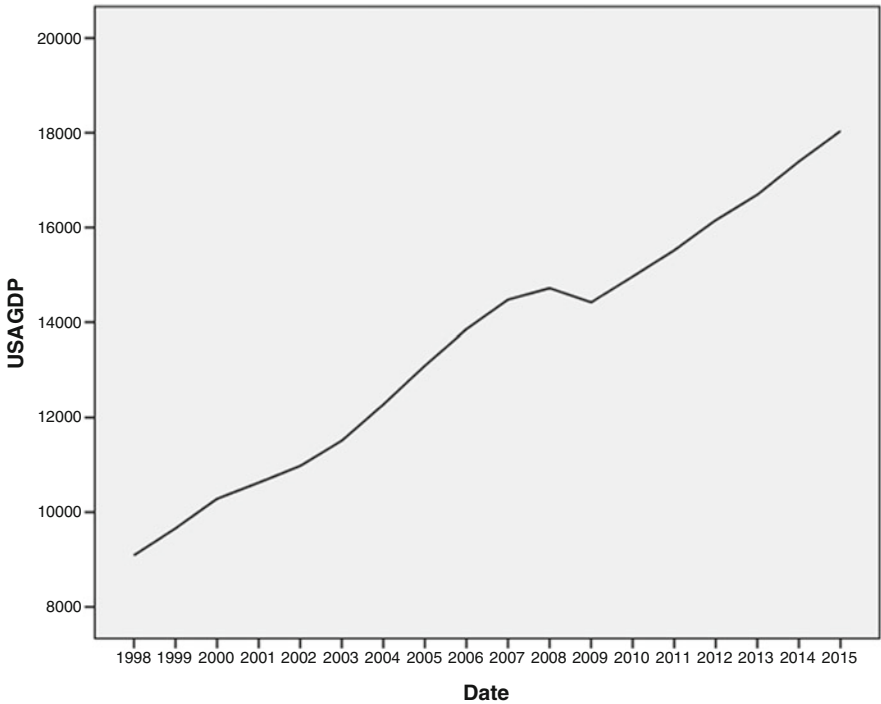


Fig. 2.27 A plot of U.S.A. GDP over time

Coefficients <sup>a</sup>					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	9212.284	147.713	62.366	.000
	time	509.293	14.833	.993	.000

a. Dependent Variable: usagdp

Fig. 2.28 Regression results for GDP against  $t$

$$\hat{Y} = 9212.284 + 509.293t$$

Note that an important point is choice between the semilog and linear trend model depends on whether one is interested in the rate of growth (semilog) or absolute growth (linear trend model). Again recall that it is not possible to compare the coefficients of determination between such competing models.

Multivariate Methods and Forecasting with IBM® SPSS®  
Statistics

Aljandali, A.

2017, XVII, 178 p. 133 illus., 80 illus. in color.,

Hardcover

ISBN: 978-3-319-56480-7