

11.6 Dummy Variables In Regression

Regression analysis requires all variables to be quantitative. From time to time it may be necessary to include qualitative variables. For example, we might be interested in knowing if gender plays an important role in the outcome of some event. Another example might be weather conditions; wet or dry, hot or cold. Mass valuation models typically include property attributes that can not be directly quantified, these include type of construction, views, central heating, etc.

The introduction of dummy variables enables the quantification of these qualitative variables which may then be included in the regression model. Dummy variables may be used for either cross-sectional studies or time series data.

Dummy variables are variables that have a value of 0 or 1

Consider the example of annual savings by people who own their own homes and those who rent them. There is a view that those who own their own homes have a different savings function to those who rent. To test this hypothesis the theoretical model given by equation (11.26) explicitly accounts for these two groups.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad (11.26)$$

where Y = a family's annual saving rate

X_1 = the family's income

X_2 = the family owns its home or rents

X_2 is a qualitative variable that distinguishes between owning and renting.

$X_2 = 1$ if family owns home, $X_2 = 0$ if family rents.

The following sample data, Table 11.5, for 20 families has been collected, annual Savings and Annual Income of 10 Home-Owning and 14 Home-Renting Families.

Name	Annual savings (\$'000) Y	Annual income (\$'000) X ₁	Owns/ rents X ₂	X ₂
Jones	1.0	20	Rents	0
Smith	1.3	24	Rents	0
Kargill	0.7	12	Rents	0
Mennon	0.8	16	Rents	0
Chou	1.8	27	Owns	1
Billings	0.5	11	Rents	0
Stratahan	2.4	32	Owns	1
Cohen	0.3	10	Rents	0
Lamb	3.2	40	Owns	1
Schmidt	2.8	32	Owns	1
Palucci	0.0	7	Rents	0
Chichester	0.3	9	Rents	0
O'Neill	3.8	36	Owns	1
Dwark	2.1	23	Rents	1
LaRue	0.0	6	Rents	0
Liu	1.0	18	Rents	0
Armour	2.0	20	Owns	1
Christenson	0.4	12	Rents	0
Howe	0.7	14	Rents	0
Pitt	1.5	15	Owns	1
Drummond	1.6	16	Owns	1
Tracy	0.6	15	Rents	0
Ming	2.2	25	Owns	1
Holland	0.6	14	Rents	0

Table 11.5: Savings for Home Owners and Home Renters

The coefficients for both X_1 and X_2 are expected to be positive since both groups are expected to save. Using least-squares to estimate the parameters of the theoretical model produces,

	Coefficients	Standard Error	t Stat	P-value	R ²	
Intercept	-0.4293	0.1209	-3.5502	0.0019	Adj. R ²	0.947
X ₁	0.0756	0.0077	9.8560	0.0000	S _e	0.237
X ₂	0.7587	0.1409	5.3867	0.0000	F	208.247

Interpreting the Coefficients

The estimated equation satisfies the theoretical considerations, the coefficients for both X_1 and X_2 are positive. Both variables are significant, their respective t-ratios are greater than the critical value from the t distribution tables, $t_{0.01,21} = 2.831$. The constant term indicates that a family has to go into deficit if there is no income, however, those who own do not need to borrow. The F statistic is highly significant (critical $F_{0.01,2,21} = 8.02$) and the adjusted R² is close to 1.

Two regression equations have been estimated, one for families owning their own home and one for those renting. This is illustrated in Figure 11.4.

$$\begin{aligned} \hat{Y}_{own} &= -0.4293 + 0.0756X_1 + 0.7587(1) \\ &= 0.3294 + 0.0756X_1 \\ \hat{Y}_{rent} &= -0.4293 + 0.0756X_1 + 0.7587(0) \\ &= -0.4293 + 0.0756X_1 \end{aligned}$$

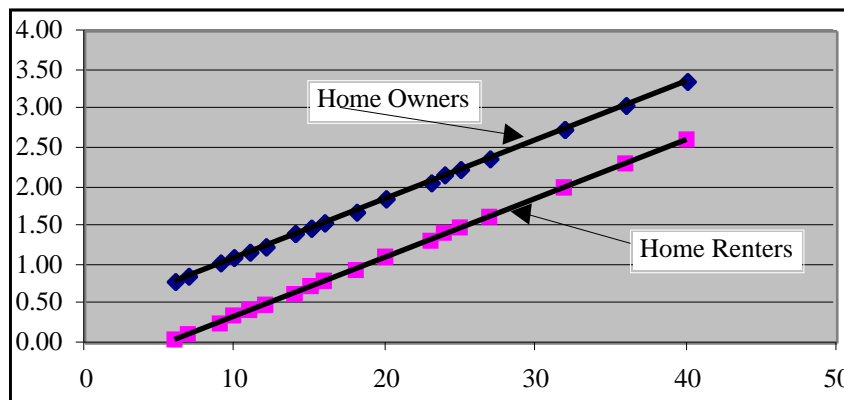


Figure 11.4 Regression Equations For Owners and Renters

If the dummy variable, X_2 , is omitted and only X_1 is used in the model, the result would be a biased estimate of β_1 . All the data points would be combined to find the best fitting line, therefore, the slope would be steeper.

When this dummy variable technique is used, it is assumed that the values of the other regression coefficients in the equation are not affected by the value of the dummy variable. It is assumed that β_1 is the same regardless of whether $X_2 = 0$ or $X_2 = 1$, that is the slopes of both lines are identical. This assumption may be true or untrue. To check two separate regressions may be estimated, one for home owners and the other for renters, and the results compared.

Regression models allow the use of dummy dependent variables as well as dummy independent variables. Dummy dependent variables are briefly discussed in the next

chapter, students who are interested in a more formal coverage of this topic should consult some of the references provided at the end of the chapter.⁵

Dummy Variables to Account For Seasonality and Cyclicity

Dummy variables are sometimes used in a regression model to isolate seasonality. A firm's sales may vary with the time of year or across different seasons. In formulating a model the supporting theory and a knowledge of the environment would suggest the nature of seasonality. The length of a season is usually regarded as some time period not exceeding twelve months and is typically monthly or quarterly. However, its length may be any time period. Periods greater than one or two years are referred to as cycles. Cyclicity is seasonality over a longer time period and may be analysed in a similar manner.

For illustrative purposes suppose a particular firm believes that its sales fluctuate quarterly according to some underlying pattern that is repeated annually.

A regression model is used to **test** for the existence of this pattern and **quantify** the pattern if it exists. To illustrate seasonality assume the firm has *quarterly data* on sales, Y , price, X_1 , and advertising, X_2 , the model is given by equation (11.27).

$$\text{Sales} = Y = f(X_1, X_2, X_3, X_4, X_5) \quad (11.27)$$

Where X_1 = price per unit
 X_2 = advertising expenditure
 X_3, X_4 and X_5 are dummy variables with the following values

	X_3	X_4	X_5
Mar - 1st quarter	0	0	0
Jun - 2nd quarter	1	0	0
Sept - 3rd quarter	0	1	0
Dec - 4th quarter	0	0	1

When the model has been estimated and verified it may be used to explain, or perhaps predict, sales for any quarter. Using the data, from Table 11.6, dummy variables are used to capture seasonality (all money variables are assumed to be in constant dollars).

Notice the dummy variables that are introduced, the values for the March quarter are all 0's while each of the other quarters has 1 and two 0's appropriately placed. The principle reason for this coding is to avoid the **dummy variable trap**. If a variable is defined for each quarter then we would have four variables, coded in such a way as to allow four values for each quarter: these values are 1 and three 0's, coded so that each quarter would be uniquely defined. If this form of coding was adopted it would give rise to multicollinearity among the X s, since the first column of the X matrix always contains 1s. A linear combination of the dummy variables would therefore be perfectly correlated with the first column of the X matrix, causing the $(X'X)$ to have a determinant of zero and hence it does not possess an inverse. To avoid this problem of perfect multicollinearity one fewer dummy variable than the number of qualitative attributes is used. Multicollinearity is discussed in the next chapter.

To avoid the dummy variable trap variables are coded such that there is one less variable than the number of attributes. In the current example there are four attributes (quarters) and three variables (X_3, X_4 and X_5). The March quarter is absorbed in the regression constant, hence, this becomes the benchmark or reference quarter for each of the others. The coefficients of X_3 (June qtr), X_4 (Sept qtr) and X_5 (Dec qtr) are compared in sign and magnitude to the constant. From the regression output, provided

⁵Gujarati and Ramanathan provide an introduction to this modelling approach.

in Table 11.7, it is apparent that, relative to the March quarter, sales are lower for both the June and September quarters but they are higher for the December quarter.

Quarter	Sales	Per Unit	Advert.	DUMMY VARIABLES		
	\$'000 Y	Price \$ X ₁	Expn. \$'000 X ₂	X ₃	X ₄	X ₅
1978.1	130.2	7.4	5.4	0.0	0.0	0.0
.2	113.3	7.7	7.0	1.0	0.0	0.0
.3	95.0	7.3	6.0	0.0	1.0	0.0
.4	162.8	8.2	7.2	0.0	0.0	1.0
1979.1	109.3	7.2	7.8	0.0	0.0	0.0
.2	95.1	7.3	5.8	1.0	0.0	0.0
.3	79.8	7.3	5.0	0.0	1.0	0.0
.4	136.6	7.9	6.1	0.0	0.0	1.0
1980.1	127.8	6.4	7.5	0.0	0.0	0.0
.2	111.2	6.6	6.8	1.0	0.0	0.0
.3	93.3	6.7	5.9	0.0	1.0	0.0
.4	159.8	7.3	7.1	0.0	0.0	1.0
1981.1	122.2	6.6	8.1	0.0	0.0	0.0
.2	106.3	6.3	6.5	1.0	0.0	0.0
.3	89.2	5.8	5.6	0.0	1.0	0.0
.4	152.8	6.7	6.8	0.0	0.0	1.0
1982.1	122.2	6.0	7.9	0.0	0.0	0.0
.2	106.3	6.2	6.5	1.0	0.0	0.0
.3	89.2	6.5	5.6	0.0	1.0	0.0
.4	152.8	6.7	6.8	0.0	0.0	1.0
1983.1	120.6	5.7	7.8	0.0	0.0	0.0
.2	104.9	5.9	6.5	1.0	0.0	0.0
.3	88.0	6.0	5.5	0.0	1.0	0.0
.4	150.8	6.5	6.7	0.0	0.0	1.0
1984.1	121.7	6.1	7.8	0.0	0.0	0.0
.2	105.9	5.7	6.5	1.0	0.0	0.0
.3	88.8	5.3	5.6	0.0	1.0	0.0
.4	152.1	5.9	6.8	0.0	0.0	1.0
1985.1	122.3	5.3	7.9	0.0	0.0	0.0
.2	106.4	5.7	6.5	1.0	0.0	0.0
.3	89.3	5.3	5.6	0.0	1.0	0.0
.4	152.9	5.5	6.8	0.0	0.0	1.0
1986.1	122.7	4.9	7.9	0.0	0.0	0.0
.2	106.7	5.3	6.6	1.0	0.0	0.0
.3	89.6	5.3	5.6	0.0	1.0	0.0
.4	153.4	5.8	6.8	0.0	0.0	1.0

Table 11.6 Seasonal Data for Sales Price and Advertising

Based on the t-ratios the variables Unit Price (X_1) and Advertising (X_2) are not significant in explaining sales (Y). All four quarters are significant, the intercept contains the first quarter, with the December quarter being the most significant. Based on this analysis Sales are affected a good deal more by seasonality than by price or expenditure on advertising.

The estimated equation for this data is:

$$\hat{Y} = 101.658 + 0.209X_1 + 2.533X_2 - 13.258X_3 - 27.995X_4 + 32.412X_5 \quad (11.28)$$

Assuming a per unit price of \$6.0 and advertising expenditure of \$8,000 the firm could predict sales revenue for the second and fourth quarters as follows:

$$\begin{aligned} \hat{Y}_{2\text{nd}} &= 101.658 + 0.209(6) + 2.533(8) - 13.258(1) - 27.995(0) + 32.412(0) \\ &= 109.918 \quad (\text{or } \$109,918) \end{aligned}$$

$$\begin{aligned}\hat{Y}_{4\text{th}} &= 101.658 + 0.209(6) + 2.533(8) - 13.258(0) - 27.995(0) + 32.412(1) \\ &= 155.588 \text{ (or \$155,588)}\end{aligned}$$

	Coefficients	Standard Error	t Stat	P-value		
Intercept	101.658	19.447	5.227	0.00001	Multiple R	0.976
X ₁	0.209	1.233	0.169	0.86667	R Square	0.953
X ₂	2.533	2.120	1.195	0.24159	Adj R Square	0.945
X ₃	-13.258	3.462	-3.829	0.00061	Std Error	5.702
X ₄	-27.995	4.963	-5.640	0.00000	F	121.525
X ₅	32.412	3.140	10.324	0.00000		

Table 11.7: Regression Output for Sales Data

It is obvious from the sales data that seasonality plays an important part in determining the company's sales. The coefficients of X₃ and X₄, corresponding to the 2nd and 3rd quarters respectively, are negative indicating that sales in these quarters are always less than sales in the other two quarters. Since the coefficient of X₄ (Sept qtr) is larger in magnitude (-27.995) than sales for this quarter are less than in any other quarter. In the 4th quarter we see that the coefficient of X₅ is +32.412 indicating that sales are highest in this quarter. The use of dummy variables in the regression equation has proven to be effective in capturing seasonality in this data.

Slope Dummies – Testing for Structural Change

The discussion on dummy variables thus far examined situations where the regression constant shifted up or down. The coefficients of the X's were assumed to be unchanging. Explanatory variables may themselves take on different characteristics due to a major change in the market or perhaps due to government regulation, a *regime shift*. Suppose we wish to examine properties across two time periods and a change in zoning regulations occurred during the second half of the period under review.

For ease of exposition consider the simple model where the price of land, Y, is determined by population density, X. Population density is affected by zoning regulations; in the second half of the period (Period II) the government imposed a restriction that there should be no more than one dwelling to the acre, whereas prior to this up to four dwellings to the acre were allowed. This restriction would be expected to flatten out the effect of population density on land prices, this is illustrated in Figure 11.5.

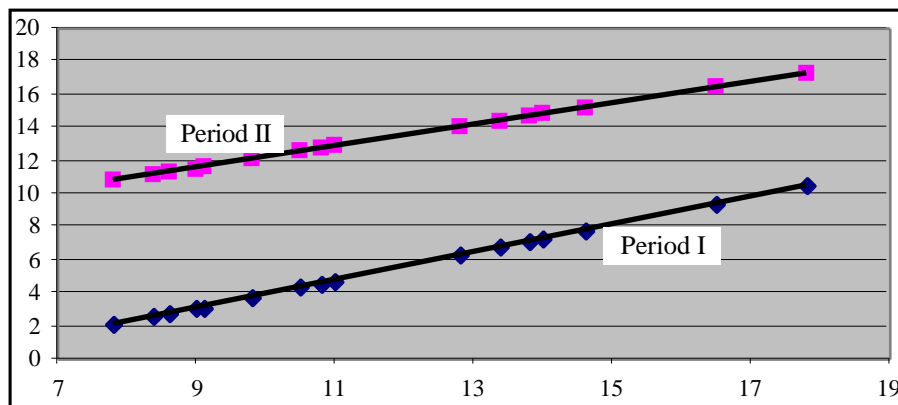


Figure 11.5: Slope and Intercept Changes

The data for both periods is pooled and equation (11.29) is estimated.

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i X_i) + \epsilon_i \quad (11.29)$$

where Y_i is land price, X_i is population density, D_i is a dummy variable with a value of 0 prior to the introduction of the restrictive zoning and a value of 1 for the new zoning period.

Pre $Y_i = \beta_0 + \beta_2 X_i + \epsilon_i$ for $D_i = 0$ (11.29)'

Post $Y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_i + \epsilon_i$ for $D_i = 1$ (11.29)''

Land Price \$'000 Y	Popn Density ('000) X	Zone Change $D_i = 0$ prior D
6.20	12.80	0
6.60	13.40	0
6.70	13.80	0
7.60	14.00	0
8.00	14.60	0
9.60	16.50	0
10.20	17.80	0
10.50	7.80	1
10.80	8.40	1
11.10	8.60	1
12.40	9.00	1
11.80	9.10	1
12.00	9.80	1
12.40	10.50	1
12.60	10.80	1
12.75	11.00	1

Table 11.8 Hypothetical Data for Land Prices Vs Population Density

The computer generated output for the estimated regression model is provided in table 11.9. The t-ratios, F-statistic and R^2 all indicate that this is an acceptable model. The dummy variable used to shift the constant has a positive sign indicating that, in the period after the new zoning regulation is introduced, land prices begin from a new base level. The slope dummy is negative indicating that increase in population in the post regulation period have a downward pressure on land prices.

	Coefficients	Standard Error	t Stat	P-value		
Intercept	-4.54	1.19	-3.82	0.00245	Std Error	0.355
X_i	0.84	0.08	10.48	0.00000	R Square	0.982
D_i	10.16	1.59	6.41	0.00003	Adj R Square	0.977
$D_i X_i$	-0.19	0.14	-1.36	0.19794	F	213.300

Table 11.9 Estimated Model for Land Value after Zoning Change

Equation (11.29)' is $\hat{Y}_i = -4.54 + 0.84X_i$

Equation (11.29)'' is $\hat{Y}_i = (-4.54 + 10.16) + (0.84 - 0.19)X_i$
 $= 5.62 + 0.65X_i$

As indicated previously, the slope of the model in the post era is lower than in the period before the new zoning regulation was introduced, the effect of population density has been flattened out.

11.7 Comparing Two Regressions

To test several coefficients jointly an F test is constructed from two regressions, the restricted and unrestricted models. Consider the two models:

Unrestricted: $Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_m X_{im} + \beta_{m+1} X_{i,m+1} + \dots + \beta_k X_{ik} + \epsilon_i$

Restricted: $Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_m X_{im} + \epsilon_i$

The unrestricted model contains k unknown regression coefficients and the restricted model m unknown coefficients. The restricted model has k-m restrictions and if this is the correct model the null hypothesis is

$$H_0: \beta_{m+1} = \beta_{m+2} = \dots = \beta_k = 0$$

The question to address is whether the k-m excluded variables have a significant **joint** effect on Y. If these excluded variables have no effect on Y then the error sum of squares for the restricted model, SSE_R, will not be vary significantly from the error sum of squares for the unrestricted model, SSE_{UR} – in other words their difference, SSE_R – SSE_{UR}, will be small.

The difference between the two sums of squares is compared to the error sum of squares for the unrestricted model to form the F statistic given by equation (11.30).

$$F_c = \frac{(SSE_R - SSE_{UR}) / (df_R - df_{UR})}{SSE_{UR} / df_{UR}} \tag{11.30}$$

Consider the unrestricted and restricted models estimated for the mass valuation data set employed in Section 11.5. The regression results are presented in Table 11.10.

Unrestricted: Price = $\beta_0 + \beta_1 \text{SQRTAREA} + \beta_2 \text{CPKS} + \beta_3 \text{BDRMS} + \beta_4 \text{INTQUAL} + \beta_5 \text{FRONT} + \beta_6 \text{GNDFL} + \beta_7 \text{TPLUS} + \epsilon_i$

Restricted: Price = $\beta_0 + \beta_1 \text{SQRTAREA} + \beta_2 \text{CPKS} + \beta_3 \text{BDRMS} + \beta_4 \text{INTQUAL} + \epsilon_i$

Unrestricted Model					Standard	
	Coefficients	Error	t Stat	P-value		
Intercept	-105664.80	9462.13	-11.17	0.00000	SSE	129,701,523,032
SQRTAREA	21937.70	1611.82	13.61	0.00000	Std Error	21999.112
CPKS	10581.55	3513.02	3.01	0.00284	R Square	0.816
BDRMS	733.09	3893.90	0.19	0.85081	Adj R Square	0.811
INTQUAL	22069.34	2285.07	9.66	0.00000	F	169.401
FRONT	-2298.97	2749.09	-0.84	0.40375		
GNDFL	-2082.20	3658.91	-0.57	0.56978		
TPLUS	-1545.16	10192.84	-0.15	0.87962		

Restricted Model					Standard	
	Coefficients	Error	t Stat	P-value		
Intercept	-104542.04	9188.20	-11.38	0.00000	SSE	130,255,180,495
SQRTAREA	21677.19	1550.43	13.98	0.00000	Std Error	21923.650
CPKS	10530.18	3478.20	3.03	0.00270	R Square	0.815
BDRMS	971.83	3797.12	0.26	0.79819	Adj R Square	0.812
INTQUAL	21780.52	2259.51	9.64	0.00000	F	298.209

Table 11.10 Regression Results for Unrestricted and Restricted Models

The joint hypothesis to test is:

$$H_0: \beta_5 = \beta_6 = \beta_7 = 0$$

$$H_1: \text{At least one of } \beta_5 \text{ or } \beta_6 \text{ or } \beta_7 \text{ not zero}$$

$$F_c = \frac{(130,255,180,495 - 129,701,523,032)/3}{129,701,523,032/268} = 0.572$$

The critical value from the F-distribution table for $F_{0.01,3,271} = 3.78$

Since the computed F statistic is less than the critical value from the tables accept H_0 and conclude that the extra sums of squares, associated with the unrestricted model, are not significant in explaining the variation in price – the excluded variables have an insignificant joint effect on Y.

The result obtained from the joint F-test is hardly surprising, the variables excluded; FRONT, GNDFL and TPLUS, all have t-ratios that are well below the critical t-value for these coefficients. The F-test merely confirms what is apparent from the results for the unrestricted model.

The variable INTQUAL is a highly significant variable in both the unrestricted and restricted models, an F-test to exclude this variable should indicate that it should remain in the model. The unrestricted model remains the same but this time the variable INTQUAL is also excluded from the restricted model. The test using the new SSE_R may be obtained from:

$$F_c = \frac{(174,916,682,103 - 129,701,523,032)/4}{129,701,523,032/268} = 23.357$$

On this occasion the F-statistic is significant, indicating that the extra sums of squares associated with the variable INTQUAL contributes to the explanation of the variance in price. It will be observed from Table 11.10 that the variable BDRMS is not significant, it may be beneficial to repeat the test excluding this variable only, this test is left as an exercise.

Testing A Linear Combination of Coefficients

Frequently situations are encountered where hypothesis may be stated as linear combinations of coefficients. For example, the valuation model, with price (Y) as a function of bedrooms (X_1), total rooms (X_2), area (X_3) and bathrooms (X_4).

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon_i \quad (11.31)$$

The marginal contribution to price by bedrooms is β_1 and for bathrooms it is β_4 . Suppose we wish to test the hypothesis that the marginal contribution of bathrooms is twice that of bedrooms, that is, $\beta_4 = 2\beta_1$. This hypothesis may be tested by explicitly including it in a restructured model.

The unrestricted model is given by (11.31) and the restricted model by (11.32).

$$\begin{aligned} Y_i &= \beta_0 + (\beta_1 X_1 + \beta_4 X_4) + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i \\ &= \beta_0 + (\beta_1 X_1 + 2\beta_1 X_4) + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i \\ &= \beta_0 + \beta_1 (X_1 + 2X_4) + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i \\ Y_i &= \beta_0 + \beta_1 X_1^* + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i \end{aligned} \quad (11.32)$$

where $X_1^* = X_1 + 2X_4$

The parameters of equation (11.31), the unrestricted model, is estimated and used to form the restricted model, given by (11.32). Next (11.32) is estimated and the error

sum of squares is obtained for the unrestricted and restricted models to carry out an F test of the form.

$$F_c = \frac{(SSE_R - SSE_{UR})/(\text{Number of Restrictions})}{SSE_{UR}/df_{UR}} \quad (11.33)$$

The hypothesis to test is: $H_0: \beta_1 = 2\beta_4$
 $H_0: \beta_1 - 2\beta_4 = 0$

If the computed F statistic, F_c , is greater than the critical F, from the F-distribution tables, at the chosen level of significance, reject H_0 . In the context of the example discussed rejection of H_0 would lead us to conclude that the marginal contribution of bathrooms to price is not twice the marginal contribution of bedrooms, that is, the coefficients are significantly different at the chosen level of significance.

11.8 Non Linear Regression Models

Up to now the discussion has focused on models of the form,

$$Y_i = b_0 + b_1X_{i1} + b_2X_{i2} + \dots + b_kX_{ik} + \epsilon_i$$

This is referred to as the multiple **linear** regression model. The term linear refers to the role of the parameters in the model. A linear model does not contain terms such as $\frac{2}{1}$, or X_1^2 , or $\log(X_1)$. The model does not exclude the possibility of the variables, dependent or explanatory, being transformed in a non-linear fashion. Thus the methods discussed thus far can equally be applied to models that are non linear in variables. The relationship described by the *bid rent curve* for property suggests that as the distance from the central location becomes greater the price of property decreases. Thus, price is some function of distance,

$$\text{Price} = f(\text{distance})$$

Observation of actual market data would suggest the notion that the relationship is not a linear function, as distance increases price decreases but at a decreasing rate. This belief may be tested by transforming the variable distance to reflect the *correct* relationship. It may be appropriate to consider a model that represents price as a function of the log of distance.

$$\text{Price} = f[\log(\text{distance})]$$

In other words, the variable distance is transformed into a new variable that more accurately reflects its relationship with price. The bid rent curve is described by a negative exponential relationship. The model is still linear in parameters, that is, it may be estimated by a linear estimating technique such as least squares. For example, the following non-linear relationships may be transformed to create new variables so that the parameters may be estimated using OLS.

$$(i) \quad Y_i = \beta_0 + \beta_1 X_{i1}^2 + \beta_2 X_{i2}^3 + \epsilon_i$$

$$(ii) \quad \log(Y_i) = \beta_0 + \beta_1 \log(X_{i1}) + \beta_2 X_{i2} + \epsilon_i$$

The models are estimated simply by prior transformation of the data values and by calculating the regression equation using the transformed values. Computer programs normally allow for such transformations of the data to be made before proceeding with the regression analysis. This task is straightforward in a spreadsheet.

Different transformations of the same explanatory variable can enter the same model. For example, the model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \quad (11.34)$$

is based on a two variable data matrix, but an additional explanatory variable has been generated by a square transformation. The right-hand side is a quadratic functional form of the variable X and the model would be appropriate if there was a quadratic relationship between the two variables. Models where basic variables enter as non-linear transformations are known as **curvilinear regression** models.

Sometimes the systematic component of a model may be formulated in such a way that the parameters are non linear, but a prior transformation of the model itself will yield an expression linear in the parameters which may in turn be estimated by the standard regression method. An example is the widely used constant elasticity function in economics. With two explanatory variables, the function takes the form

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \quad (11.35)$$

implying that the changes in Y are in constant proportion to changes in X_1 or X_2 . In this theoretical model β_0 , β_1 , and β_2 are the parameters to be estimated and Y , X_1 and X_2 are economic variables.

Equation (11.35) may be transformed by taking logarithms of both sides of the equation

$$\log(Y) = \log(\beta_0) + \beta_1 \log(X_1) + \beta_2 \log(X_2) \quad (11.35)'$$

The equation is now linear in parameters, with the exception that the constant term is expressed as a logarithm. The addition of a random component leads to a regression model that is estimable.

Sometimes the background theory may suggest the kind of transformations to be applied to some or all of the variables. Our previous example of a production function with inputs of labour, X_1 and capital, X_2 is indicative of the type of functional form suggested by theory and hence the transformation required to produce a least squares model.

More often, while the theory or background knowledge may suggest the existence of a statistical relationship, it provides no guidance as to the functional form: whether it is linear, logarithmic, quadratic, and so on. One way out of the difficulty may be to fit a small number of models of different functional form and choose between them on the basis of either formal significance tests or descriptive procedures such as residual plots. For example, a quadratic form may be compared to a linear form by testing the significance of adding the term X^2 .

The competition is often between the normal linear model and the log-linear model in which some or all of the variables have been transformed logarithmically. The log-linear model is sometimes preferred, partly because of its proportionate effect interpretation and partly because it tends to remove skewness. Some common variable transformations are presented in Table 11.11.

The appropriate functional form to use may not always be obvious, trial and error is sometimes necessary. During an earlier example, discussed in section 11.5, we examined the relationship between price of residential units and the attributes of the property – AREA, VIEWS, CARPARKS, etc. Most of these variables were expected to have a positive relationship with price. The variable AREA, the most important variable in the model, contributed to price in a non-linear form. Price is positively affected by AREA but as AREA increases price increases at a decreasing rate. The appropriate

transformation of the variable AREA was to take its square root⁶, a linear relationship exists between price and the square root of AREA.

When modelling property prices the log-linear transformation is often considered, i.e., a linear relationship between the log of price and the attributes of the property exists. The Carbone-Longini Feedback system, used for mass valuation, is a general form of a non-linear model.⁷ This model, however, cannot be linearised and must be estimated using a non-linear algorithm.

Functional Form	Linear Transformation
	* indicates logged variable
General Linear Model	$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$
Log-linear $Y = e^{\beta_0 + \beta_1 X_1}$	$Y^* = \beta_0 + \beta_1 X_1$
Linear-log $Y = \beta_0 + \beta_1 \ln(X_1)$	$Y = \beta_0 + \beta_1 X_1^*$
Double Log $Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2}$ $\ln(Y) = \ln(\beta_0) + \beta_1 \ln(X_1) + \beta_2 \ln(X_2)$	$Y^* = \beta_0^* + \beta_1 X_1^* + \beta_2 X_2^*$
Quadratic $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ $X_2 = X_1^2$
Square root $Y = \beta_0 + \beta_1 \sqrt{X_1}$	$Y = \beta_0 + \beta_1 X_2$ $X_2 = \sqrt{X_1}$
Reciprocal $Y = \beta_0 + \beta_1 \frac{1}{X_1}$	$Y = \beta_0 + \beta_1 X_2$ $X_2 = \frac{1}{X_1}$
Interaction $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ $X_3 = X_1 X_2$
Trans-log $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$ $X_3 = X_1^2, X_4 = X_2^2, X_5 = X_1 X_2$
Logit $P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1)}}$ $Y = \ln \frac{P_i}{1 - P_i} = \ln[e^{\beta_0 + \beta_1 X_1}]$	$Y = \beta_0 + \beta_1 X_1$

Table 11.11 Transformation of Non-linear Relationships to Linear Relationships

Interaction Effects

⁶Other transformations were considered, for example $\ln(\text{AREA})$, but the result was not as good.

⁷The Carbone-Longini model is described in J.K. Eckert (Ed) 1990, **Property Appraisal and Assessment Administration**, The International Association of Assessing Officers. See also R. Carbone and R. Longini, 1977, "A Feedback Model for Automated Real Estate Assessment," **Management Science**, Vol 24, No 3, November.

Given the relationship $Y = f(X_1, X_2)$ prior knowledge may suggest that while the variables X_1 and X_2 contribute to explaining the variation in Y individually their interaction, X_1X_2 , also affects Y .

Two variables X_i and X_j are said to interact in the change in Y for a unit change in X_i – when X_j is held fixed – is dependent on the value of X_j .

To test for the interaction effect the following model may be specified

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

and a test of significance for the coefficient β_3 is carried out.

$$\begin{array}{ll} \text{Null Hypothesis} & H_0: \beta_3 = 0 \text{ (interaction not effective)} \\ \text{Alternative Hypothesis} & H_1: \beta_3 \neq 0 \text{ (interaction is effective)} \end{array}$$

If β_3 is significant then the interaction effect should be included in the equation. In the business environment there is a high degree of interaction between economic variables so it is generally a good idea to consider including an interaction term in the model. In some cases the variables in the model may affect the dependent variable in their linear form, in a nonlinear form, and also have an interaction effect. The following second order model is an example of this type of relationship

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$$

This model specification is referred to as a **Trans-log** model. Second order models of this type dramatically increase the number of parameters to be estimated, with only two explanatory variables there are five parameters to be estimated. If three independent variables are included this specification requires that ten parameters are estimated. There is, therefore, a significant loss in degrees of freedom from the model and possibly problems of multicollinearity which could arise. Trans-log models also give rise to multicollinearity problems and should be used with a good deal of caution.

Regression Analysis of Shopping Centre Rents

Gerbish (1998) examined a sample of 293 New Zealand shopping centre leases to empirically test some widely held beliefs regarding tenancy mix. The model proposed by Gerbish is presented here as equation (11.36).

$$R_i = b_0 + b_1 S_i + b_2 G_i + b_3 T_i + b_4 Q_1 + b_5 Q_2 + b_6 S_i Q_1 + b_7 S_i Q_2 + e_i \quad (11.36)$$

A description of the variables, their expected signs and estimated coefficients are provided in Table 11.12.

The regression constant is the *Mall Stores* base scalar comparison category for R_i , relative to this base the *Anchor* and *Food Court* dummies will be positive or negative. The results indicate that the rent for *Anchor* tenants is less than this base while that for the *Food Court* is higher.

The interaction term, to take account of the possibility of slope changes in the *Anchor* variable due to *Size*, is insignificant while the interaction between *Food Court* and *Size* is just significant at the 5% level. "This indicates that *Food Court* base rentals are more sensitive to size differences than either *All Mall Stores* or an *Anchor*. *Food Court* tenants would appear to suffer significant diseconomies from operating in larger premises, resulting in lower rental rates as the size of tenancy increases." [Gerbish, page 290]

Variable	Description	Expected sign	Coefficient	t-stat
----------	-------------	---------------	-------------	--------

R _i (Dep. var)	ln of base rent /m ²		R ² = 0.49	
	Constant (<i>All Mall Stores</i>)		2.83	4.4
S _i (net let. area)	ln of Size in m ²	negative	-0.20	-7.1
G _i	ln of Occupancy cost/m ²	negative	-0.05	-0.5
T _i	ln of <i>Centre Turnover</i>	positive	0.50	5.7
Q ₁	Anchor dummy	negative	-1.66	-2.3
Q ₂	Food Court dummy	positive	1.15	2.9
S _i Q ₁	Interaction – Anchor and Size	unspecified	0.16	-1.6
S _i Q ₂	Interaction – Food Court and Size	unspecified	-0.05	-2.1

Table 11.12 Hypothesised and Estimated Model for Shopping Centre Rents

Appendix 11.1 The Covariance Matrix

Given the general linear model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i$

Expressed in matrix form it is: $Y = X\beta + \epsilon$ (A11.1)

where β is a (k+1) x 1 vector of parameter estimates

X is the n x (k+1) matrix of observations of the independent variables

and ϵ is the n x 1 vector of random errors

Proof that the Covariance Matrix = $S_e^2(X'X)^{-1}$

The normal equations are: $X'Y = (X'X)\beta$ (A11.2)

Solving for β : $\beta = (X'X)^{-1}X'Y$ (A11.3)

$$= (X'X)^{-1}X'(X\beta + \epsilon)$$

$$= X\beta + (X'X)^{-1}X'\epsilon \quad \text{since } (X'X)^{-1}X'X = I_n$$

rearranging gives $\beta - \beta = (X'X)^{-1}X'\epsilon$ (A11.4)

If X is nonstochastic then $E(X'\epsilon) = 0$ and $E(\beta) = \beta$

$E(\beta) = \beta$ where β represents the vector of parameter estimates for β

$$\begin{aligned} \text{VAR}(\beta) &= E[(\beta - E(\beta))(\beta - E(\beta))'] \\ &= E[(\beta - \beta)(\beta - \beta)'] \end{aligned} \quad \text{(A11.5)}$$

Substituting from (A11.4)

$$\begin{aligned} \text{VAR}(\beta) &= E[(X'X)^{-1}X'\epsilon][(X'X)^{-1}X'\epsilon]'] \\ &= E[(X'X)^{-1}X'\epsilon][\epsilon'X(X'X)^{-1}] \end{aligned} \quad \text{(ABC)' = C'B'A'}$$

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

$$= (X'X)^{-1}X'S_e^2I_nX(X'X)^{-1} \quad \text{since } E[\epsilon\epsilon'] = S_e^2I_n$$

$$= S_e^2I_n(X'X)^{-1}X'X(X'X)^{-1}$$

$$= S_e^2(X'X)^{-1} \quad \text{(A11.6)}$$

The Projection, or Hat, Matrix

The least squares fit: $\hat{Y} = X\hat{\beta}$

$$\begin{aligned}
 &= X(X'X)^{-1}X'Y \\
 &= HY \qquad \qquad \qquad (A11.7)
 \end{aligned}$$

The H matrix is known as the **hat matrix** or **projection matrix**, when it is multiplied by the vector Y the result is the fitted values in the least squares regression of Y on X. H is symmetric, $H = H'$ and idempotent $H = H^2$.

$$\begin{aligned}
 \text{H is symmetric: } X(X'X)^{-1}X' &= (X(X'X)^{-1}X')' \\
 &= (X')'[(X'X)^{-1}]'X' && \boxed{(ABC)' = C'B'A'} \\
 &= X(X'X)^{-1}X' && \boxed{(X'X)^{-1} \text{ is symmetric}}
 \end{aligned}$$

$$\begin{aligned}
 \text{H is idempotent: } X(X'X)^{-1}X' &= [X(X'X)^{-1}X'][X(X'X)^{-1}X'] \\
 &= XI(X'X)^{-1}X' && \boxed{(X'X)^{-1}X'X = I} \\
 &= X(X'X)^{-1}X'
 \end{aligned}$$

The vector of least squares residuals is:

$$\begin{aligned}
 \underline{e} &= Y - \hat{Y} = Y - X\underline{b} && \boxed{\hat{Y} = X\underline{b}} \\
 &= Y - X(X'X)^{-1}X'Y && \boxed{\text{Replacing } \underline{b} \text{ with (A11.3)}} \\
 &= (I - X(X'X)^{-1}X')Y \\
 &= MY \qquad \qquad \qquad (A11.8)
 \end{aligned}$$

The matrix M is symmetric ($M = M'$) and idempotent ($M = M^2$).

$$\begin{aligned}
 M &= [I - X(X'X)^{-1}X'] = M' \\
 M^2 &= [I - X(X'X)^{-1}X'] [I - X(X'X)^{-1}X'] \\
 &= I - 2IX(X'X)^{-1}X' + X(X'X)^{-1}X'X(X'X)^{-1}X' \\
 &= I - 2X(X'X)^{-1}X' + X(X'X)^{-1}X' \\
 &= I - X(X'X)^{-1}X' \\
 &= M
 \end{aligned}$$

M may be interpreted as the matrix that, when it pre-multiplies any vector Y, produces the vector of least squares residuals in the regression of Y on X. It follows that $MX = 0$, which states the regression of X on X produces a perfect fit and the residuals will be zero.

Estimating the Variance – 2

$$\underline{e}'\underline{e} = \underline{e}'M'\underline{e} = \underline{e}'\underline{e}$$

$$E(\underline{e}'\underline{e}) = E(\underline{e}'[I - X(X'X)^{-1}X']\underline{e}) = E(\underline{e}'\underline{e})$$

$$\begin{aligned}
 &= E[\text{tr}(\underline{e}'\underline{e})] = E[\text{tr}(M\underline{e}\underline{e}') && \boxed{\text{The trace of a matrix is discussed in chapter 3, Section 3.6}} \\
 &= \text{tr}[E(M\underline{e}\underline{e}')] = \text{tr}[ME(\underline{e}\underline{e}')] \\
 &= \text{tr}[M] \quad 2 \\
 &= \text{tr}[I - X(X'X)^{-1}X'] \quad 2 \\
 &= 2[\text{tr}(I) - \text{tr}(X(X'X)^{-1}X')] \\
 &= 2[\text{tr}(I) - \text{tr}(H)] \\
 &= 2[n - (k+1)]
 \end{aligned}$$

$$\text{Thus,} \quad 2 \quad = \frac{E(\underline{e}'\underline{e})}{n - (k+1)} = \frac{1}{n - (k+1)} [Y'Y - \underline{b}'X'Y]$$

References

- Berk, K.N., & Carey, P., 1998, *Data Analysis with Microsoft Excel*, Duxbury Press. Chapter 9.
- Gerbish, M., 1998, "Shopping Center Rentals: An Empirical Analysis of the Retail Tenant Mix," in *Journal Of Real Estate Research*, American Real Estate Society, Volume 15, Number 3.
- Gujarati, D.M., 1996, *Basic Econometrics*, Third edition, McGraw-Hill International Editions. Chapters 7, 8, 9 and 15.
- Lee, C.F., 1993, *Statistics for Business and Financial Economics*, D.C. Heath and Company. Chapters 15.
- Levine, D.M., Berenson, M.L. & Stephan, D., 1997, *Statistics for Managers Using Microsoft Excel*, Prentice -Hall, Inc. Chapter 12.
- Middleton, M.R., 1997, *Data Analysis Using Microsoft Excel*, Duxbury Press. Chapters 15 and 16.
- Neter, J., Kutner, M.H., Nachtsheim, C.J, & Wasserman, W, 1996, *Applied Linear Statistical Models*, Fourth Edition, Irwin. Chapters 6, 7 and 8.
- Ramanathan,R., 1995, *Introductory Econometrics With Applications*, Harcourt Brace Jovanovich International Edition. Chapters 4 and 5.
- Thomas, R.L., 1997, *Modern Econometrics an Introduction*, Addison-Wesley. Chapter 7.

Exercises

1. What are the advantages of multiple regression over simple regression?
2. Measurement vector Y represents petrol consumption taken at six monthly intervals (June and December) over five years in a certain growing town. Assuming that the population of cars in the town increases by the same amount in January each year and that we expect seasonality to be additive (in each year Y in December is above June by the same amount, b_2), what are the least squares estimators of the population parameters of a model which explains the petrol consumption figures?

The model is: $\hat{Y} = b_0X_0 + b_1X_1 + b_2X_2$

where X_0 is a unit column vector, X_1 starts at 1 and increases by 1 each year, and X_2 is 0 in June and 1 in December.

Matrix X is

and the vector Y is

1	1	0	4
1	1	1	6
1	2	0	4
1	2	1	8
1	3	0	6
1	3	1	10
1	4	0	10
1	4	1	12
1	5	0	11
1	5	1	14

X is the matrix of independent (exogenous) variables and Y is the vector of dependent (endogenous) variables and is interpreted as the vector of raw data. Multiple regression is an algorithm which will minimise the sum of squares of the error terms of $(Y_i - \hat{Y}_i)^2$ for the above linear model and yield estimates of \underline{b} .

Thus, for example, calculate -

- (i) $X'X$ (ii) $\det.(X'X)$
 (iii) $(X'X)^{-1}$ (iv) $(X'X)^{-1}X'Y$
 (v) Check your answer by showing: $(X'X)\underline{b} = X'Y$ (vi) $Y'Y$
 (vii) $\underline{b}'X'Y$ (viii) $R^2 = \frac{\underline{b}'X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}$
 (ix) $S_e^2 = \frac{1}{n-(k+1)}$ (x) $H = X(X'X)^{-1}X'$
 $= \frac{1}{n-(k+1)} [Y'Y - \underline{b}'X'Y]$ (xi) $M = M' = M^2 = I - P$
 (xii) $X\underline{b} = PY$ (xiii) $HY = \hat{Y}$
 (xiv) Find the product of $S_e\sqrt{(X'X)^{-1}}$ for the diagonal elements of $(X'X)^{-1}$ and calculate the t-ratios.
 (xv) \hat{Y} , being the predicted value of Y , for each of the ten half years.
 (xvi) $F = \frac{(\underline{b}'X'Y - n\bar{Y}^2)/k}{(Y'Y - \underline{b}'X'Y)/(n-k-1)}$
 (xvi) Project \hat{Y} for December half year in year 6.

3. Application of regression in evaluating a Real Estate problem

	Sale Price \$1'000 Y	House Size 100 Sq Ft X_1	Condition Rating (1 to 10) X_2
1	60.0	23	5
2	32.7	11	2
3	57.7	20	9
4	45.5	17	3
5	47.0	15	8
6	55.3	21	4
7	64.5	24	7
8	42.6	13	6
9	54.5	19	7

10 57.5 25 2

Source: R.L. Andrews and J.T. Ferguson, "Integrating Judgement with a Regression Appraisal," *The Real Estate Appraiser and Analyst*, Vol 52, No.2, Spring 1986, Table 1.

- (i) Fit the models: $\hat{Y} = b_0 + b_1X_1$ and $\hat{Y} = b_0 + b_1X_1 + b_2X_2$
 - (ii) Which model is best ? Are your conclusions supported by the model selection criteria from Table 11.4.
4. Miles & Co. Inc. placed 12 graduates in its six month management training program. At the beginning, each trainee was given two different aptitude tests. At the end of the training program, each trainee was rated by a committee consisting of three senior managers. The relevant data are given below.

- (i) Estimate the linear multiple regression equation, using the method of ordinary least squares for the model $\hat{Y} = b_0 + b_1X_1 + b_2X_2$.
- (ii) What relationship is described by the regression equation?
- (iii) On the average, what is the mean rating of trainees who receive scores 600 and 80 on aptitude tests A and B, respectively? Use a point estimate.
- (iv) Calculate the standard error of the estimate. Is it relatively large or small?
- (v) Calculate the coefficient of multiple determination. What is its meaning here?
- (vi) Calculate and interpret the coefficient of multiple correlation.
- (vii) Include X_3 in the model, i.e., estimate $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$ Is X_3 significant? What is your interpretation of this coefficient?

Trainee	Test A Score X_1	Test B Score X_2	Final Rating Y	Test Score A 700 or higher X_3
Squinton	694	82	87	0
Webston	512	84	75	0
Shinten	723	90	96	1
McCarthen	660	85	90	0
Matthews	623	82	88	0
Lambeton	778	94	98	1
Warrington	594	75	78	0
Lenton	588	73	75	0
Frampton	724	91	96	1
Lau	638	81	88	0
Hrekpth	659	82	87	0
Jptnap	715	87	92	1

5. A real estate appraiser wishes to use the sample data below to establish a predictive relationship between the selling price of a home and the building and lot size.
- (i) Estimate the equation: $\hat{Y} = b_0 + b_1X_1 + b_2X_2$
 - (ii) Calculate S_e , R^2 , Adj. R^2 , F and the t-ratios.
 - (iii) Test the hypothesis that b_1 and b_2 are not significant at the 1% level. What is your conclusion about the equation?
 - (iv) What practical applications could an equation such as this be used for? What, if any, are its limitations?

Price Building Size Lot Size

\$'000	100 sq. ft.	1000 sq. ft.
Y	X ₁	X ₂
45	21	21
37	16	23
26	17	7
32	14	9
34	19	11
49	18	45
53	23	12
65	22	10
71	24	10
88	26	22

6. Consider the problem of predicting profit Y (in thousands of dollars) for supermarkets in a large metropolitan area. As independent variables we use the total sales (in tens of thousands of dollars) X_1 of foods and X_2 of non-foods. One reason for splitting sales into food and non-food categories is that stores will differ significantly from each other in their offerings of non-food items. The variable X_3 for store size, is included to improve profit prediction.

S'market Number	Food Sales \$10,000 X ₁	Non-food Sales \$10,000 X ₂	Store Size sq. ft. 1000 X ₃	Profit \$1,000 Y
1	305	35	35	20
2	130	98	22	15
3	189	83	27	17
4	175	76	16	9
5	101	93	28	16
6	269	77	46	27
7	421	44	26	35
8	195	57	12	7
9	282	31	40	22
10	203	92	32	23

(i) Estimate: $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$$

$$\hat{Y} = b_0X_1^{b_1}X_2^{b_2}X_3^{b_3} \quad \text{and} \quad \hat{Y} = b_0X_1^{b_1}X_2^{b_2}$$

- (ii) For each equation calculate S_e , R^2 , Adj. R^2 , F and t -ratios and test the hypothesis that the regression coefficients are not significant at the 1% level.
- (iii) What do you conclude about the addition of the interaction term, X_1X_2 , and X_3 to the equation and the functional form of the equation?
- (iv) How should the choice of functional form be made? What is the preferred choice of model using the model selection criteria in Table 10.2?
7. Use the Residential Units data set to carry out the analysis (available at text Web site)
- (i) Estimate the regression model that best fits the data.
- (ii) Discuss the features of the data from a theoretical and statistical perspective. Use the model selection criteria available in the text, Table 11.4, to assist in the process of model selection.

8. RENT BREAKDOWN: We are attempting to set rentals in a project consisting of small warehouse and unserviced office units. Rentals for a very similar development, and the respective office and warehouse areas, measured in square feet, are provided below.

(i) Fit the models:

$$Y = f(X_1, X_2), \quad Y = f(X_1, X_2, X_3), \quad Y = f(X_1, X_2, X_1X_2),$$

$$Y = f(X_1, X_2, X_1^2, X_2^2), \quad Y = f(X_1, X_2, X_1^2, X_2^2, X_1X_2)$$

Which model is preferred, use the model selection criteria in Table 11.4.

- (ii) Repeat the estimation for the models in part (i) using $\ln(Y)$ as the dependent variable. Does transforming Y represent an improvement in capturing the underlying relationship between the variables?

Unit Number	Office Area X_1	Warehouse Area X_2	Office & Warehouse X_3	Annual Rent Y	Log of rent $\ln(Y)$
1	1815	2310	4125	1200	7.0901
2	360	2235	2595	5900	8.6827
3	420	2700	3120	7200	8.8818
4	350	2050	2400	5700	8.6482
5	350	1850	2200	5400	8.5942
6	1097	1103	2200	6750	8.8173
7	280	1320	1600	3600	8.1887
8	280	1320	1600	3750	8.2295
9	280	1160	1440	3400	8.1315
10	880	560	1440	4560	8.4251
11	350	1250	1600	3900	8.2687
12	450	150	600	4120	8.3236
13	274	4318	4592	5450	8.6034
14	250	1190	1440	3360	8.1197
15	302	1426	1728	4200	8.3428
16	676	1340	2016	5420	8.5979
17	690	750	1440	4250	8.3547
18	264	1896	2160	4800	8.4764
19	274	2606	2880	6450	8.7718
20	260	1180	1440	3450	8.1461

Source: Gene Dilmore, 1981, *Quantitative Techniques in Real-Estate Counseling*, Lexington Books, D.C. Heath and Company.

9. LAND VALUATION USING CENSUS DATA: The following regression uses as independent variables (1) population; (2) retail sales, city; (3) apparel sales, metro; (4) women's, girls' clothing sales, metro; and (5) bank deposits, county. The dependent variable is the estimated 100% downtown land value for the city. The purpose was to provide a method for a rough estimate of 100% land values for economic analyses, based solely on available census data. The raw data are given below.

(i) Estimate:

$$\hat{Y} = b_0 + b_1X_1, \quad \hat{Y} = b_0 + b_1X_1 + b_2X_2$$

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$$

- (ii) If your objective was estimating land value only, how many variables would you use?

- (iii) If you were concerned mainly with the contributory factors for land value, and wanted to retain variables significant at the 95% level, which regression would you use?
- (iv) Considering S_e , R^2 , $\text{Adj.}R^2$, F and the t -ratios for each equation what conclusions may be drawn? Are these models really informative?

City	X_1 1,000	X_2 \$1,000,000	X_3 \$1,000,000	X_4 \$1,000,000	X_5 \$1,000,000	Y
1	30.1	153.685	11.716	10.002	137.6	6.0
2	299.2	1266.227	142.161	128.661	1154.6	45.0
3	36.7	132.108	18.403	13.067	86.1	6.0
4	54.7	187.107	13.640	12.660	114.3	6.4
5	147.2	472.996	38.586	43.309	172.2	15.0
6	140.1	526.567	47.924	47.099	347.1	27.5
7	69.5	211.349	19.893	12.893	111.1	16.0
8	147.6	680.510	42.166	64.002	545.1	22.5
9	129.0	447.311	36.161	40.482	216.1	16.0
10	117.1	404.361	39.237	35.518	315.8	25.0
11	23.1	197.960	12.202	15.541	193.6	12.0
12	454.7	1349.097	94.661	128.505	1348.2	75.0
13	490.9	2219.170	232.341	396.496	2536.7	150.0
14	356.5	1478.200	287.285	324.921	2825.2	100.0
15	64.6	253.014	25.696	32.332	179.9	7.5
16	303.9	1498.284	165.335	227.127	878.2	45.0
17	587.2	1551.132	187.546	201.401	1775.6	150.0
18	44.9	356.576	28.207	23.421	332.0	8.0
19	55.6	387.796	33.359	40.191	268.7	12.0
20	121.7	440.033	41.389	51.179	338.6	15.0
21	43.3	206.141	17.836	12.262	181.5	10.0
22	174.8	689.917	37.758	64.485	726.8	30.0
23	193.7	588.481	55.155	56.824	518.4	35.0

Source: Gene Dilmore, 1981, *Quantitative Techniques in Real-Estate Counseling*, Lexington Books, D.C. Heath and Company.

10. The owner of an apartment building in Minneapolis believed that her 1990 property tax bill was too high due to an over assessment of the property's value by the city tax assessor. The owner hired an independent real estate appraiser to investigate the appropriateness of the city's assessment. The appraiser used regression analysis to explore the relationship between the sale prices of apartments sold in Minneapolis during 1990 and various characteristics of the properties. Twenty-five apartment buildings were randomly sampled from all apartment buildings that were sold during 1990. The table below lists the data collected by the appraiser. The real estate appraiser hypothesised that the sale price (i.e., market value) of an apartment building is related to the other variables in the table according to the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

- (i) Fit the real estate appraiser's model to the data in the table. Report the least squares prediction equation.
- (ii) Find the standard deviation of the regression model and interpret its value in the context of this problem.
- (iii) Do the data provide sufficient evidence to conclude that value increases with the number of units in an apartment building? Report the observed significance level, and reach a conclusion using $\alpha = .05$.

- (iv) Interpret the value of β_1 in terms of these data. Remember that your interpretation must recognise the presence, of the other variables in the model.
- (v) Construct a scatter gram of sale price versus age. What does your scatter gram suggest about the relationship between these variables?
- (vi) Test $H_0: \beta_2 = 0$ against $H_1: \beta_2 < 0$ using $\alpha = .01$. Interpret the result in the context of the problem. Does the result agree with your observation in part (v) ? Why is it reasonable to conduct a one-tailed rather than a two-tailed test of this null hypothesis?
- (vii) What is the observed significance level of the hypothesis test of part (vi) ?

Code No	Sale Price Y (\$)	No. of Apartment Units X ₁	Age of Structure X ₂ (Years)	Lot Size X ₃ (sq. ft.)	No. on-site Parking Spaces X ₄	Gross Building Area X ₅	Condition of Apartment Building
0229	90,300	4	82	4,635	0	4,266	F
0094	384,000	20	13	17,798	0	14,391	G
0043	157,500	5	66	5,913	0	6,615	G
0079	676,200	26	64	7,750	6	34,144	E
0134	165,000	5	55	5,150	0	6,120	G
0179	300,000	10	65	12,506	0	14,552	G
0087	108,750	4	82	7,160	0	3,040	G
0120	276,538	11	23	5,120	0	7,881	G
0246	420,000	20	18	11,745	20	12,600	G
0025	950,000	62	71	21,000	3	39,448	G
0015	560,000	26	74	11,221	0	30,000	G
0131	268,000	13	56	7,818	13	8,088	F
0172	290,000	9	76	4,900	0	11,315	E
0095	173,200	6	21	5,424	6	4,461	G
0121	323,650	11	24	11,834	8	9,000	G
0077	162,500	5	19	5,246	5	3,828	G
0060	353,500	20	62	11,223	2	13,680	F
0174	134,400	4	70	5,834	0	4,680	E
0084	187,000	8	19	9,075	0	7,392	G
0031	155,700	4	57	5,280	0	6,030	E
0019	93,600	4	82	6,864	0	3,840	F
0074	110,000	4	50	4,510	0	3,092	G
0057	573,200	14	10	11,192	0	23,704	E
0104	79,300	4	82	7,425	0	3,876	F
0024	272,000	5	82	7,500	0	9,542	E

Source: J.T. McClave and P.G. Benson, 1991, *Statistics for Business and Economics*, Fifth edition, MacMillan Publishing Company, page 636.

11. Using the data from question 10:

- (i) Fit a first-order model to the data. (You may already have done this for question 10.)
- (ii) Do the data provide sufficient evidence to conclude that the model of part (i) is useful for predicting sale price ? Test using $\alpha = .05$.
- (iii) Drop X_3 and X_4 from the model of part (i) and refit the model to the data.
- (iv) Do the data provide sufficient evidence to conclude that the model of part (iii) is useful for predicting sale price ? Test using $\alpha = .05$.

12. The data in the following table provide personal savings and personal income (in billions of dollars) for the time period 1935 to 1949.

Year	Personal Savings Y	Personal Income X ₁	War Year
1935	2	60	
1936	4	69	
1937	4	74	
1938	1	68	
1939	3	73	
1940	4	78	
1941	11	96	
1942	28	123	yes
1943	33	151	yes
1944	37	165	yes
1945	30	171	yes
1946	15	179	
1947	7	191	
1948	13	210	
1949	9	207	

- (i) Plot the data as a scatter diagram showing peacetime years with dots and war time years with crosses.
- (ii) Using values for the dummy variable $X_2 = 0$ for peacetime and $X_2 = 1$ for wartime, determine the estimated regression equation:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

- (iii) Plot the two lines obtained from this equation corresponding to wartime and peacetime on your scatter diagram.
- (iv) Using a model of the form described by equation (11.29) in the text, test the hypothesis that the slope of X_1 is different in war years to peacetime.
13. Six women and four men have taken a test that measures their manual dexterity and patience in using their hands with tiny objects. Each has then gone through a week of intensive training as electronics assemblers, followed by a month at actual assembly, during which their productivity was measured by a relative index having values ranging from 0 to 10 (with 10 the most productive worker). The results obtained are provided in the following table.

SUBJECT	Productivity Index Y	Test Score X ₁	Sex
A	5.2	5.8	F
B	6.0	8.5	M
C	6.5	8.2	F
D	2.0	3.5	F
E	2.7	6.5	M
F	10.0	9.5	F
G	6.4	9.8	M
H	6.6	9.2	M
I	3.5	4.0	F
J	4.0	5.5	F

- (i) Plot the scatter diagram, using dots for women and crosses for men
- (ii) Using a dummy variable having value $X_2 = 1$ for women and $X_2 = 0$ for men, determine the coefficients for the equation

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

Draw the lines corresponding to $X_2 = 0$ and $X_2 = 1$ on your scatter diagram.

- (iii) State in words the meaning of the partial regression coefficients (i.e. b_0, b_1, b_2).
- (iv) Determine the estimated regression line obtained when the sex of the subjects is ignored, and plot it on your scatter diagram.
- (v) Use an interaction dummy to test that the slope of X_1 is the same for both sexes. What does your conclusions show?
14. In questions 10 and 11 a real estate appraiser used regression analysis to explore the relationship between the sale prices of apartments and various characteristics of the apartments. The last column of this data set contains data on the physical condition of each apartment building (E: excellent; G: good; F: fair).
- (i) Write a model that describes the relationship between sale price and number of apartment units as three parallel lines, one for each level of physical condition. Be sure to specify the dummy variable coding scheme you use.
- (ii) Plot Y against X_1 (number of apartment units) for all buildings in excellent condition. On the same graph, plot Y against X_1 for all buildings in good condition. Do this again for all buildings in fair condition. Does it appear that the model you specified in part (i) is appropriate? Explain.
- (iii) Fit the model from part (i) to the data. Report the least squares prediction equation for each of the three building condition levels.
- (iv) Plot the three prediction equations of part (iii) on a scatter gram of the data.
- (v) Does the data provide sufficient evidence to conclude that the relationship between the mean sale price and number of units differs depending on the physical condition of the apartments? Test using $\alpha = .05$.
15. (i) Distinguish between regression analysis and correlation analysis.
- (ii) What is the correlation coefficient and what does it measure? What is the coefficient of determination and what does it measure?
- (iii) What is a dummy variable? When is it used? How is it interpreted?
- (iv) What is a variable transformation? When is it employed?
16. The data (provided below) was collected to conduct a study of the relation of amount of body fat (Y) to several possible explanatory, independent variables, based on a sample of 20 healthy females 25-34 years old. The possible independent variables are:
- X_1 triceps skinfold thickness
 X_2 thigh circumference
 X_3 midarm circumference
- (i) Estimate the models:
- $Y = f(X_1)$ $Y = f(X_1, X_2)$
 $Y = f(X_2)$ $Y = f(X_1, X_2, X_3)$

What is the change in SSR and SSE and what does this tell us?

What is the marginal effect of adding X_3 to the model given that X_1 and X_2 are already included? What is the effect on the coefficients of the explanatory variables as new variables are added?

- (ii) Is it appropriate to drop X_2 and X_3 from the model? Use an F test of the form

$$F = \frac{SSR(X_2, X_3 | X_1)/df}{SSE(X_1, X_2, X_3)/df}$$

where $SSR(X_2, X_3 | X_1)$ is the extra sum of squares associated with the regression when X_2 and X_3 are included in the model given that X_1 is already included. $SSE(X_1, X_2, X_3)$ is the error sum of squares from the regression with all three variables included. The degrees of freedom (df) for the numerator is $k_2 - k_1 = 2$ and the denominator is $n - (k+1)$.

Is your conclusion consistent with a t-test for the coefficients of X_2 and X_3 ?

- (iii) Should any independent variables not yet included in the model be considered for possible inclusion? To test for omitted variables consider the model of the form:

$$Y_i = \beta_0 + \beta_1 X_i + \sum_{k+j} \beta_{k+j} Z_j^i \text{ for } i = 1, 2, \dots, k, \text{ and } j = 2, 3, \dots, p$$

where Z_j^i represent powers of Z : Z^2, Z^3, Z^4 , etc. and the variable Z represents fitted values, \hat{Y}_i , from the regression of $Y_i = \beta_0 + \beta_1 X_i$. Test at the 1% level using a joint F-test defined as follows.

$$F = \frac{(SSE_1 - SSE_2)/(k_2 - k_1)}{SSE_2/(n - k_2)}$$

Consider also the regression of $\hat{u} = \beta_0 + \sum_j \beta_j Z_j^i$ for $j = 1, 2, \dots, p$, and test that the coefficients of this model are significantly different to zero.

X_1	X_2	X_3	Y
19.5	43.1	29.1	11.9
24.7	49.8	28.2	22.8
30.7	51.9	37.0	18.7
29.8	54.3	31.1	20.1
19.1	42.2	30.9	12.9
25.6	53.9	23.7	21.7
31.4	58.5	27.6	27.1
27.9	52.1	30.6	25.4
22.1	49.9	23.2	21.3
25.5	53.5	24.8	19.3
31.1	56.6	30.0	25.4
30.4	56.7	28.3	27.2
18.7	46.5	23.0	11.7
19.7	44.2	28.6	17.8
14.6	42.7	21.3	12.8
29.5	54.4	30.1	23.9
27.7	55.3	25.7	22.6
30.2	58.6	24.6	25.4
22.7	48.2	27.1	14.8
25.2	51.0	27.5	21.1

17. Many companies must accurately estimate their costs before a job is begun in order to acquire a contract and make a profit. A heating and plumbing contractor, for example, may base cost estimates for new homes on the total area of the house and whether central air conditioning is to be installed.

- (i) Write a main effects model relating the mean cost of material and labour, $E(Y)$, to the area and central air conditioning variables.
 - (ii) Write a complete second-order model for the mean cost as a function of the same two variables.
 - (iii) What hypothesis would you test to determine whether the second-order terms are useful for predicting mean cost ?
 - (iv) Explain how you would compute the F statistic needed to test the hypothesis of part (iii).
18. Investors are interested in knowing the relationship between the behaviour of a mutual fund and the behaviour of the stock market as a whole. Researchers in finance have hypothesised that the model that appropriately characterises this relationship is

$$E(Y) = \beta_0 + \beta_1 X$$

where $Y =$ Monthly rate of return of a mutual fund
 $X =$ Monthly rate of return of the stock market as a whole as measured by the monthly rate of return to a market index such as Standard & Poor's 500 Composite Index.

The value of β_1 in the above model is referred to as the mutual fund's **beta coefficient**. Assuming the preceding model is true, investors can predict how the returns of an individual mutual fund will react to changes in the behaviour of the market. For example, if $\beta_1 > 1$, the implication is that the return to the mutual fund will be greatly influenced by the behaviour of the market and will move in the same direction as the change in the market return. If $0 < \beta_1 < 1$, the return to the mutual fund will be less sensitive to changes in market behaviour but will also move in the same direction as the change in the market return.

In studying mutual funds, Alexander and Stover (1980)⁸ included a dummy variable in the above model to determine whether the beta coefficient for an individual mutual fund depends on whether the market is moving generally upward (a **bull market**) or generally downward (a **bear market**).

- (i) Modify the above regression model (as Alexander and Stover did) to reflect the possibility that $E(Y)$ may depend on whether the market is bullish or bearish. Include an interaction term in your model and carefully define the dummy variable coding scheme you use.
 - (ii) Using the model you developed in part (i), describe the differences that may exist between the response curves of $E(Y)$ under bull and bear markets.
 - (iii) Specify the hypothesis you would test to determine whether a mutual fund's beta coefficient is different during bull and bear markets.
 - (iv) Specify the hypothesis you would test to determine whether $E(Y)$ should be characterized as $E(Y) = \beta_0 + \beta_1 X$ or as in the modified model you developed in part (i).
19. To explain the movements in domestic and international passenger travel through Ngurah Rai International Airport Bali during the past 17 years the following time series data has been collected.

⁸ Alexander, G.J., & Stover, R.D., "Consistency of mutual fund performance during varying market conditions", *Journal of Economics and Business*, Spring 1980, 32, 219-226.

Year	Per Capita GNP (Rps) X_1	Jet Fuel Price (Rps) X_2	Hotel Rooms X_3	Domestic Passengers Y_1	International Passengers Y_2
1981	96,854.7	150	1,425	687,056	317,256
1982	1,084,012.4	240	4,036	678,290	395,120
1983	1,006,573.9	300	4,897	764,940	285,360
1984	929,135.5	300	6,412	906,036	292,966
1985	812,977.8	330	9,402	925,868	312,163
1986	851,697.1	250	11,689	1,037,755	371,813
1987	812,977.8	250	12,843	1,175,818	502,442
1988	788,897.4	250	14,427	1,086,830	939,795
1989	970,608.4	250	16,872	1,487,728	864,310
1990	1,043,844.1	330	21,812	1,850,096	730,846
1991	1,193,880.9	400	25,177	1,905,605	948,021
1992	1,341,191.2	400	27,021	1,829,435	1,476,168
1993	1,691,054.4	420	27,945	1,607,909	2,294,581
1994	1,950,802.9	420	28,942	1,789,593	2,801,869
1995	2,276,892.9	420	28,983	1,929,473	2,726,604
1996	2,633,762.0	420	31,348	2,096,645	2,966,460
1997	3,028,111.7	420	32,608	1,996,800	2,985,934

- (i) Estimate a model for both domestic and international travel demand using the explanatory variables provided. Are the significant explanatory variables, and their level of significance, the same for both models? Explain why this is so.
- (ii) For each type of travel demand, estimate what you consider the best model. Explain your choice.
20. In question 18, the relationship between the behaviour of an individual mutual fund and the behaviour of the stock market as a whole was discussed. The table lists the monthly rates of return for the Dreyfus Fund (a mutual fund) and the monthly rates of return for Standard & Poor's 500 Composite Index (S&P) for the period January 1966 to December 1971. The bear market periods were from January 1966 through September 1966 and from December 1968 through May 1970. The bull market periods were from October 1966 through November 1968 and from June 1970 through December 1971 (Alexander and Stover, 1980).

	Time Period	RETURNS			Time Period	RETURNS	
		Dreyfus	S&P			Dreyfus	S&P
1	Jan-66	0.008	0.006	37	Jan-69	-0.001	-0.007
2	Feb-66	0.067	-0.013	38	Feb-69	-0.070	-0.043
3	Mar-66	-0.008	-0.021	39	Mar-69	0.015	0.036
4	Apr-66	0.021	0.022	40	Apr-69	0.014	0.023
5	May-66	-0.074	-0.049	41	May-69	-0.003	0.003
6	Jun-66	0.024	-0.015	42	Jun-69	-0.066	-0.054
7	Jul-66	-0.011	-0.012	43	Jul-69	-0.059	-0.059
8	Aug-66	0.099	-0.073	44	Aug-69	0.057	0.045
9	Sep-66	-0.011	-0.005	45	Sep-69	-0.001	-0.024
10	Oct-66	0.015	0.049	46	Oct-69	0.050	0.046
11	Nov-66	0.079	0.010	47	Nov-69	-0.027	-0.030
12	Dec-66	0.008	0.000	48	Dec-69	-0.027	-0.018
13	Jan-67	0.086	0.080	49	Jan-70	-0.083	-0.074
14	Feb-67	0.010	0.007	50	Feb-70	-0.044	0.059
15	Mar-67	0.041	0.041	51	Mar-70	-0.007	0.003
16	Apr-67	0.048	0.044	52	Apr-70	-0.110	-0.089
17	May-67	-0.039	-0.048	53	May-70	-0.060	-0.055
18	Jun-67	0.027	0.019	54	Jun-70	-0.042	-0.048
19	Jul-67	0.073	0.047	55	Jul-70	0.051	0.075
20	Aug-67	-0.019	-0.007	56	Aug-70	0.051	0.051
21	Sep-67	0.010	0.034	57	Sep-70	0.047	0.035
22	Oct-67	-0.029	-0.028	58	Oct-70	-0.026	-0.010

23	Nov-67	0.011	0.007	59	Nov-70	0.040	0.054
24	Dec-67	0.025	0.028	60	Dec-70	0.046	0.058
25	Jan-68	-0.063	-0.043	61	Jan-71	0.044	0.042
26	Feb-68	-0.042	-0.026	62	Feb-71	0.025	0.014
27	Mar-68	0.022	0.011	63	Mar-71	0.031	0.038
28	Apr-68	0.100	0.083	64	Apr-71	0.035	0.038
29	May-68	0.021	0.016	65	May-71	-0.034	-0.037
30	Jun-68	0.001	0.011	66	Jun-71	-0.008	0.002
31	Jul-68	-0.038	-0.017	67	Jul-71	-0.026	-0.040
32	Aug-68	0.021	0.010	68	Aug-71	0.045	0.041
33	Sep-68	0.056	0.040	69	Sep-71	-0.028	-0.006
34	Oct-68	0.010	0.000	70	Oct-71	-0.039	-0.040
35	Nov-68	0.062	0.050	71	Nov-71	0.019	0.003
36	Dec-68	-0.025	-0.040	72	Dec-71	0.075	0.090

Sources: Standard & Poor's Composite Index returns from Ibbotson, R. G. and Singuefield, R. A., *Stocks, Bonds, Bills, and Inflation: The Past (1926-1976) and the Future (1977-2000)*, Financial Analysts Research Foundation, 1977; and Dreyfus returns from *The Wall Street Journal*.

- (i) Fit the model you developed in part (i) of Question 18 to the data shown in the table.
- (ii) Using the fitted model, estimate the Dreyfus Fund's beta coefficient for bull markets. Estimate the corresponding parameter for bear markets. Describe the relative responsiveness of the mutual fund to the market during bullish and bearish periods.
- (iii) Test the hypothesis you specified in part (iii) of Question 18. Draw the appropriate conclusions. Test using $\alpha = .05$.
- (iv) Test the hypothesis you specified in part (iv) of Question 18. Draw the appropriate conclusions. Test using $\alpha = .05$.

21. Show that $\text{Var}(\underline{b}) = \sigma^2(\underline{X}'\underline{X})^{-1}$. Assume $\underline{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$

22. Given that $\text{SST} = \text{SSE} + \text{SSR}$, and $R^2 = \frac{\text{SSR}}{\text{SST}}$

where: $\text{SST} = \sum (Y_i - \bar{Y})^2$ $\text{SSR} = \sum (\hat{Y}_i - \bar{Y})^2$ $\text{SSE} = \sum (Y_i - \hat{Y}_i)^2$

Express each of these in matrix notation and show that:

$$R^2 = \frac{\underline{b}'\underline{X}'\underline{Y} - n\bar{Y}^2}{\underline{Y}'\underline{Y} - n\bar{Y}^2}$$

23. Spreadsheet Skill Building Exercise

The data provided below refers to a lending company operating in 10 sectors within Melbourne. The number of competing loan companies operating in a particular sector (X) and the number per hundred of the firm's loans made in that sector that are currently delinquent (Y).

	1	1	1	1	1	1	1	1	1
X:	5	6	4	1	2	3	3	4	5
Y:	19	21	15	5	10	15	13	22	23

Enter the data in the spreadsheet as three columns, the X matrix contains a column of 1s to take account of the regression constant. Use matrix algebra to estimate a model of the form: $Y_i = b_0 + b_1X_i + b_2$

- (i) Create the following matrices in Excel and assign the **names** allocated to them (for an example of naming cell ranges in Excel refer to the last subsection in chapter 4.):

$$\mathbf{A} = [\mathbf{I} - \left(\frac{1}{n}\right)\mathbf{J}] \quad \mathbf{M} = [\mathbf{I} - \mathbf{H}] \quad \mathbf{D} = [\mathbf{H} - \left(\frac{1}{n}\right)\mathbf{J}]$$

\mathbf{I} is the identity matrix of order $(n \times n)$, \mathbf{J} is an $(n \times n)$ matrix of 1s ($n=10$) and $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

- (i) Determine (and assign names for) $\mathbf{Y}'\mathbf{Y}$ $\mathbf{X}'\mathbf{X}$ $\mathbf{X}'\mathbf{Y}$ $(\mathbf{X}'\mathbf{X})^{-1}$
- (ii) Obtain the vector of estimated regression coefficients, $\underline{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, and the vector of residuals, $\underline{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\underline{\mathbf{b}}$ using the matrix $(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{M}\mathbf{Y}$.
- (iii) Obtain the standard error of the estimate, S_e^2 , and the estimated variance-covariance matrix, $S_e^2(\mathbf{X}'\mathbf{X})^{-1}$
- (iv) Using the matrices defined in (i) obtain:
 $\text{SST} = \mathbf{Y}'\mathbf{A}\mathbf{Y}$, $\text{SSE} = \mathbf{Y}'\mathbf{M}\mathbf{Y}$ ($= \underline{\mathbf{e}}'\mathbf{M}\underline{\mathbf{e}}$), $\text{SSR} = \mathbf{Y}'\mathbf{D}\mathbf{Y}$, R^2 and F .
- (v) Obtain \mathbf{H} and show that \mathbf{H} and $(\mathbf{I} - \mathbf{H})$ are idempotent.
- (vi) Verify that the $\text{tr}(\mathbf{I}) = n$ and $\text{tr}(\mathbf{H}) = k+1$. The trace is equal to the sum of the diagonal elements of a matrix (refer chapter 3, Section 3.6).
- (vi) Determine the standard errors and t-ratios for the coefficients.
- (vii) Obtain a point estimate of $E(\mathbf{Y}/\mathbf{X}=5)$.

11. Multiple Regression.....	311
11.1 The Multiple Regression Model.....	311
11.2. An Example of Multiple Regression.....	316
11.3 Validation of the Estimated Equation.....	319
11.4 Analysis of Variance for the Regression.....	323
11.5 Model Selection Criteria.....	325
11.6 Dummy Variables In Regression.....	328
11.7 Comparing Two Regressions.....	333
11.8 Non Linear Regression.....	336
Appendix 11.1 The Covariance Matrix.....	339
Appendix 11.2 Data Set for Residential Units.....	341
References.....	345
Exercises.....	346

- adjusted R2, 314
AIC, 325
bid rent curve, 336
BLUE, 313
coefficient of multiple determination, 313, 320
corrected coefficient of determination, 314
covariance matrix, 315
covariance, 315
criteria for model comparison, 325
criterion statistic, 325
curvilinear regression, 337
degrees of freedom, 314
diagonal elements, 315
dummy variable trap, 330
dummy variables, 328
Error ratio, 320
F statistic, 329
F test, 320
F-distribution, 336
fitted model, 311
FPE, 325
GCV, 325
general linear model, 340
Gerbish, 339
goodness of fit, 314
hat matrix, 341
HQ, 325
Hypotheses test, 321
hypothesis tests, 315
hypothesised population model, 316
idempotent, 341
Interpretation of Results, 322
least squares, 313
log-linear transformation, 338
Mass Valuation, 326
mass valuation models, 319
matrix functions in Excel, 319
matrix notation, 317
Model Selection Criteria, 325
model validation, 319
multicollinearity, 327
multiple regression, 311
Non Linear Regression, 336
non linear in variables, 336
normal equations, 312
OLS assumptions, 313
OLS, 312
ordinary least squares, 312
parameter estimates, 312
perfect multicollinearity, 330
population parameter, 315
population parameters, 311, 314
projection matrix, 341
qualitative variables, 328
regime shift, 332
Regression Results, 322
regression application, 316
regression plane, 319, 322
residential units, 337
restricted model, 335
RICE, 325
sample covariance matrix, 320
SCHWARZ, 325
second derivative, 312
SGMASQ, 325
SHIBATA, 325
shopping centre leases, 339
simultaneous linear equations, 312
Slope Dummies, 332
slope dummy, 333
SSE, 313
SSR, 313
SST, 313
standard error, 319
symmetric, 341
t-distribution, 321
tenancy mix, 339
Testing A Linear Combination, 335
Testing for Structural Change, 332
theoretical considerations, 329
theoretical model, 328
Trans-log model, 339
transformations, 337
unrestricted model, 335