

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} .

(i) $XX' = \begin{matrix} 10 & 30 & 5 \\ 30 & 110 & 15 \\ 5 & 15 & 5 \end{matrix}$ (ii) $\det.(X'X) = 500$

(iii) $(X'X)^{-1} = \begin{matrix} 0.650 & -0.150 & -0.200 \\ -0.150 & 0.050 & 0.000 \\ -0.200 & 0.000 & 0.400 \end{matrix}$ (iv) $(X'X)^{-1}X'Y = \begin{matrix} 1 \\ 2 \\ 3 \end{matrix}$

(v) $(X'X)\underline{b} = \begin{matrix} 85 \\ 295 \\ 50 \end{matrix} = X'Y$ (vi) $Y'Y = 829$ (vii) $\underline{b}'X'Y = 825$

(viii) $R^2 = \frac{\underline{b}'X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2} = \frac{825 - 10(8.6364)^2}{829 - 10(8.6364)^2} = 0.952$

(ix) $S_e^2 = \frac{1}{n-(k+1)} (Y - X\underline{b})'(Y - X\underline{b})$
 $= \frac{1}{n-(k+1)} [Y'Y - \underline{b}'X'Y]$
 $= \left(\frac{1}{10 - 3}\right)[829 - 825] = 0.5714$

(x) $P = X(X'X)^{-1}X'$

$$P = \begin{vmatrix} 0.400 & 0.200 & 0.300 & 0.100 & 0.200 & 0.000 & 0.100 & -0.100 & 0.000 & -0.200 \\ 0.200 & 0.400 & 0.100 & 0.300 & 0.000 & 0.200 & -0.100 & 0.100 & -0.200 & 0.000 \\ 0.300 & 0.100 & 0.250 & 0.050 & 0.200 & 0.000 & 0.150 & -0.050 & 0.100 & -0.100 \\ 0.100 & 0.300 & 0.050 & 0.250 & 0.000 & 0.200 & -0.050 & 0.150 & -0.100 & 0.100 \\ 0.200 & 0.000 & 0.200 & 0.000 & 0.200 & 0.000 & 0.200 & 0.000 & 0.200 & 0.000 \\ 0.000 & 0.200 & 0.000 & 0.200 & 0.000 & 0.200 & 0.000 & 0.200 & 0.000 & 0.200 \\ -0.100 & -0.100 & 0.150 & -0.050 & 0.200 & 0.000 & 0.250 & 0.050 & 0.300 & 0.100 \\ -0.100 & 0.100 & -0.050 & 0.150 & 0.000 & 0.200 & 0.050 & 0.250 & 0.100 & 0.300 \\ 0.000 & -0.200 & 0.100 & -0.100 & 0.200 & 0.000 & 0.300 & 0.100 & 0.400 & 0.200 \\ -0.200 & 0.000 & -0.100 & 0.100 & 0.000 & 0.200 & 0.100 & 0.300 & 0.200 & 0.400 \end{vmatrix}$$

(xi) $M = M' = M^2 = I - P$

$$M = \begin{vmatrix} 0.600 & -0.200 & -0.300 & -0.100 & -0.200 & 0.000 & -0.100 & 0.100 & 0.000 & 0.200 \\ -0.200 & 0.600 & -0.100 & -0.300 & 0.000 & -0.200 & 0.100 & -0.100 & 0.200 & 0.000 \\ -0.300 & -0.100 & 0.750 & -0.050 & -0.200 & 0.000 & -0.150 & 0.050 & -0.100 & 0.100 \\ -0.100 & -0.300 & -0.050 & 0.750 & 0.000 & -0.200 & 0.050 & -0.150 & 0.100 & -0.100 \\ -0.200 & 0.000 & -0.200 & 0.000 & 0.800 & 0.000 & -0.200 & 0.000 & -0.200 & 0.000 \\ 0.000 & -0.200 & 0.000 & -0.200 & 0.000 & 0.800 & 0.000 & -0.200 & 0.000 & -0.200 \\ -0.100 & 0.100 & -0.150 & 0.050 & -0.200 & 0.000 & 0.750 & -0.050 & -0.300 & -0.100 \\ 0.100 & -0.100 & 0.050 & -0.150 & 0.000 & -0.200 & -0.050 & 0.750 & -0.100 & -0.300 \\ 0.000 & 0.200 & -0.100 & 0.100 & -0.200 & 0.000 & -0.300 & -0.100 & 0.600 & -0.200 \\ 0.200 & 0.000 & 0.100 & -0.100 & 0.000 & -0.200 & -0.100 & -0.300 & -0.200 & 0.600 \end{vmatrix}$$

(xii)
$$\begin{matrix} 3 \\ 6 \\ 5 \\ 8 \\ 7 \\ 10 \\ 9 \\ 12 \\ 11 \\ 14 \end{matrix} \underline{Xb} = \underline{PY} = \hat{Y}$$

(xiii) Find the product of $S_e\sqrt{(X'X)^{-1}}$ from the diagonal elements of $(X'X)^{-1}$ and calculate the t-ratios.

$$S_e = \sqrt{0.5714} = 0.7559 \quad (X'X)^{-1} = \begin{matrix} 0.650 & -0.150 & -0.200 \\ -0.150 & 0.050 & 0.000 \\ -0.200 & 0.000 & 0.400 \end{matrix}$$

$$S_e\sqrt{(X'X)^{-1}} = (0.7559) \begin{matrix} \mathbf{0.650} & -0.150 & -0.200 \\ -0.150 & \mathbf{0.050} & 0.000 \\ -0.200 & 0.000 & \mathbf{0.400} \end{matrix} \begin{matrix} 0.5 \\ 0.000 \\ 0.400 \end{matrix}$$

$$S_{b_0} = 0.7559\sqrt{0.650} = 0.6094 \quad t_0 = \frac{1}{0.6094} = 1.64$$

$$S_{b_1} = 0.7559\sqrt{0.050} = 0.169 \quad t_1 = \frac{2}{0.169} = 11.83$$

$$S_{b_2} = 0.7559\sqrt{0.400} = 0.4781 \quad t_2 = \frac{3}{0.4781} = 6.27$$

$$(xiv) F = \frac{(b'X'Y - n\bar{Y}^2)/k}{(Y'Y - \underline{b}'X'Y)/(n-k-1)} = \frac{(825 - 10(8.6364)^2)/2}{(829 - 825)/(10-3)} = 69.235$$

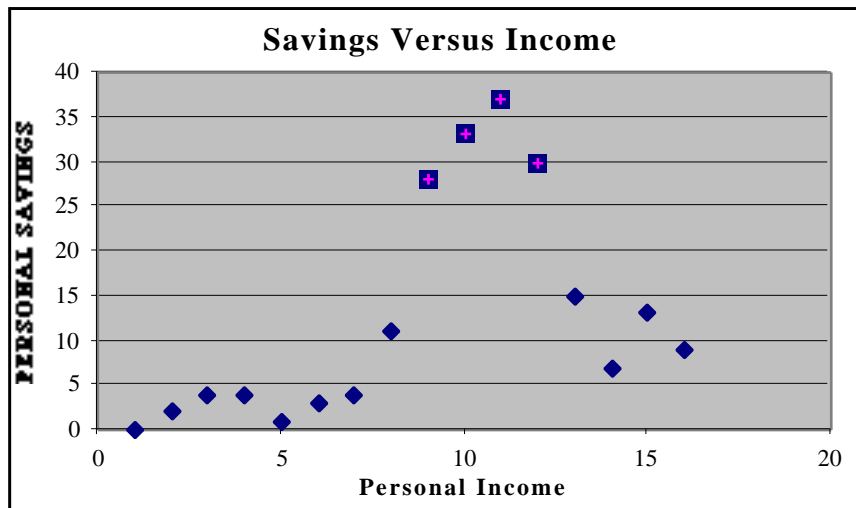
(xv) \hat{Y} , being the predicted value of Y - see (xii) above.

(xvi) Project \hat{Y} for December half year in year 6.

$$\begin{aligned} \hat{Y} &= b_0 + b_1X_1 + b_2X_2 + b_3X_3 \\ &= 1 + 2(6) + 3(1) = 16 \end{aligned}$$

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

- (i) Plot the data as a scatter diagram showing peacetime years with dots and war time years with crosses. War time \oplus , Peace time \bullet



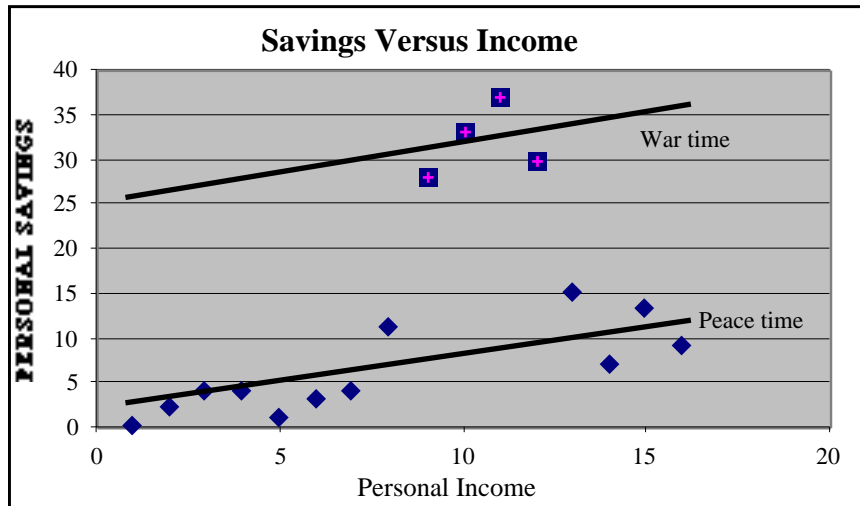
- (ii) Using values for the dummy variable $X_2 = 0$ for peacetime and $X_2 = 1$ for wartime, determine the estimated regression equation:

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

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Standard Error</i> | |
|-----------|---------------------|-----------------------|---------------|----------------|-----------------------|--------|
| Intercept | -0.415 | 2.110 | -0.197 | 0.84736 | R Square | 0.943 |
| X_1 | 0.059 | 0.016 | 3.760 | 0.00272 | Adj. R Square | 0.933 |
| X_2 | 23.351 | 1.947 | 11.995 | 0.00000 | F | 98.888 |

- (iii) Plot the two lines obtained from this equation corresponding to wartime and peacetime on your scatter diagram.

$$\hat{Y}_{\text{peace}} = -0.415 + 0.059X_1 \qquad \hat{Y}_{\text{war}} = 22.936 + 0.059X_1$$



- (iv) Using the model described by equation 10.19 in the text, describing slope and intercept dummies, test the hypothesis that the slope of X_1 is different in war years to peacetime.

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3(X_2X_1)$$

$$\hat{Y}_{\text{peace}} = b_0 + b_1X_1$$

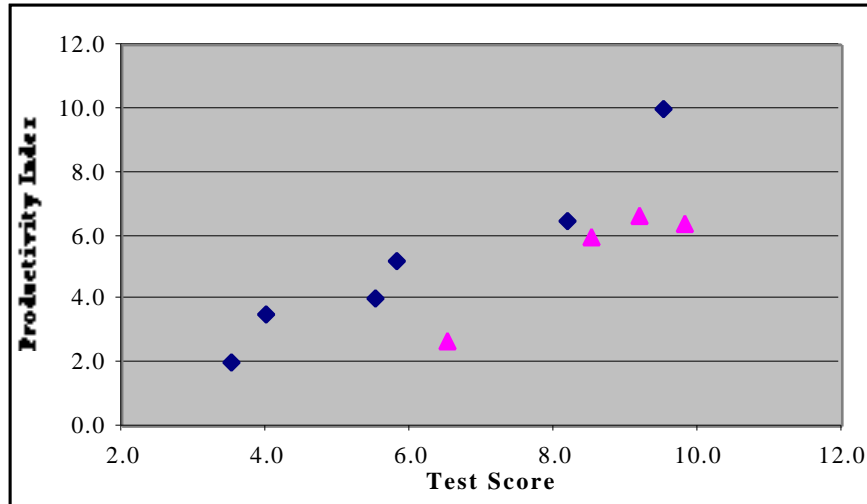
$$\hat{Y}_{\text{war}} = (b_0 + b_2) + (b_1 + b_3)X_1$$

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | | |
|-----------|---------------------|-----------------------|---------------|----------------|----------------|--------|
| Intercept | -0.234 | 2.208 | -0.106 | 0.91739 | Standard Error | 3.310 |
| X_1 | 0.058 | 0.017 | 3.488 | 0.00508 | R Square | 0.944 |
| X_2 | 16.439 | 13.911 | 1.182 | 0.26221 | Adj. R Square | 0.929 |
| X_2X_1 | 0.046 | 0.091 | 0.502 | 0.62548 | F | 61.901 |

The t-ratio for the coefficient of X_3 is not significant indicating that the slope of X_1 in war time is not significantly different from its slope in peace time. The introduction of this additional variable, to take account of changes in the slope, has made the variable X_2 , the slope dummy, insignificant. The preferred model is therefore a dummy variable that accounts for the shift in the intercept during war years with the slope remaining unchanged throughout the entire period – this is the model estimated in part (ii) of the question.

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.

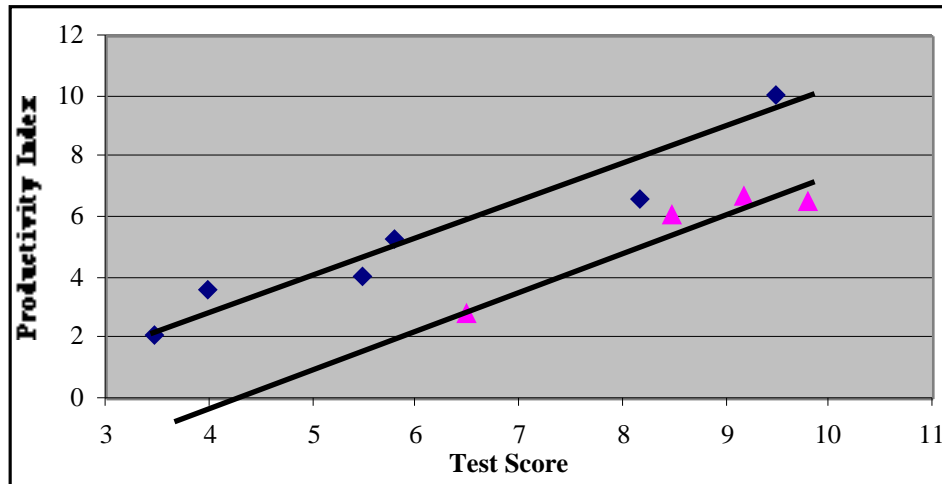
- (i) Plot the scatter diagram, using ♣ for women and ♠ for men



- (ii) Using a dummy variable having value $X_2 = 1$ for women and $X_2 = 0$ for men, determine the coefficients for the equation: $\hat{Y} = b_0 + b_1X_1 + b_2X_2$

| | Coefficients | Standard Error | t Stat | P-value | Standard Error | 0.745 |
|-----------|--------------|----------------|--------|---------|----------------|--------|
| Intercept | -4.442 | 1.151 | -3.859 | 0.00622 | R Square | 0.922 |
| X_1 | 1.161 | 0.128 | 9.058 | 0.00004 | Adj. R Square | 0.899 |
| X_2 | 2.580 | 0.572 | 4.512 | 0.00276 | F | 41.136 |

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).

b_0 this in the regression constant. When $X_2 = 0$, b_0 captures the effect of male productivity, thus male productivity is the benchmark against which women may be compared in this model. When $X_2 = 1$, the regression constant becomes $-4.442 + 2.580 = -1.862$.

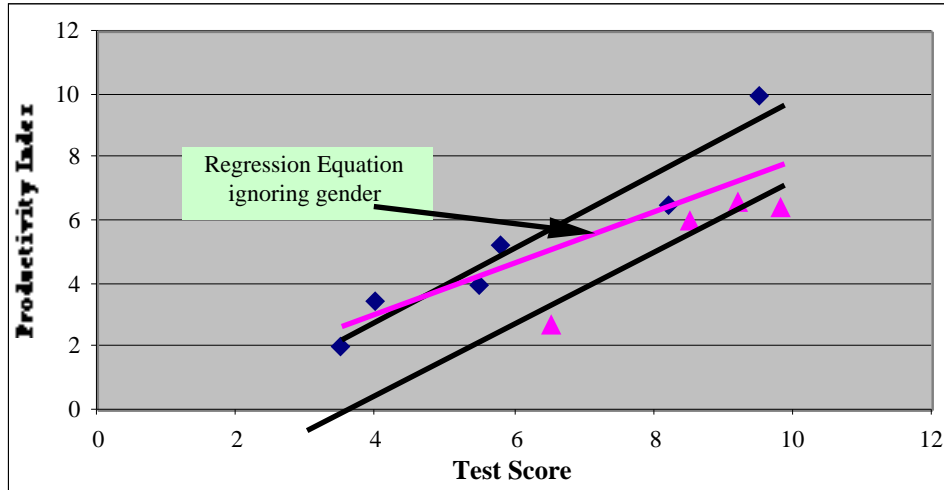
b_1 represents the marginal effect of test score on the Productivity index. The value of $b_1 = 1.161$, indicating that this effect is positive and one unit increase in the test score results in a 1.161 increase in the productivity index, holding all other influences constant.

b_2 is the gender effect; compared to the performance of males, females performance is a constant amount of 2.580 better.

- (iv) Determine the estimated regression line obtained when the sex of the subjects is ignored, and plot it on your scatter diagram.

If gender is ignored the regression equation is:

| | Coefficients | Standard Error | t Stat | P-value | Standard Error |
|----------------|--------------|----------------|--------|---------------|----------------|
| Intercept | -0.686 | 1.471 | -0.466 | 0.6533 | |
| X ₁ | 0.848 | 0.199 | 4.254 | 0.0028 | |
| | | | | R Square | 0.693 |
| | | | | Adj. R Square | 0.655 |
| | | | | F | 18.10 |



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. Some of the data from those exercises are reproduced in the following table, which also 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.

Coding scheme:

| | X ₂ | X ₃ |
|---|----------------|----------------|
| E | 1 | 0 |
| G | 0 | 1 |
| F | 0 | 0 |

Apartment buildings that are rated fair are used as the benchmark against which buildings that are rated excellent and good are compared. Since buildings rated excellent and good are expected to have a higher sale price than buildings rated fair, the signs of the coefficients for X₂ and X₃ should be positive.

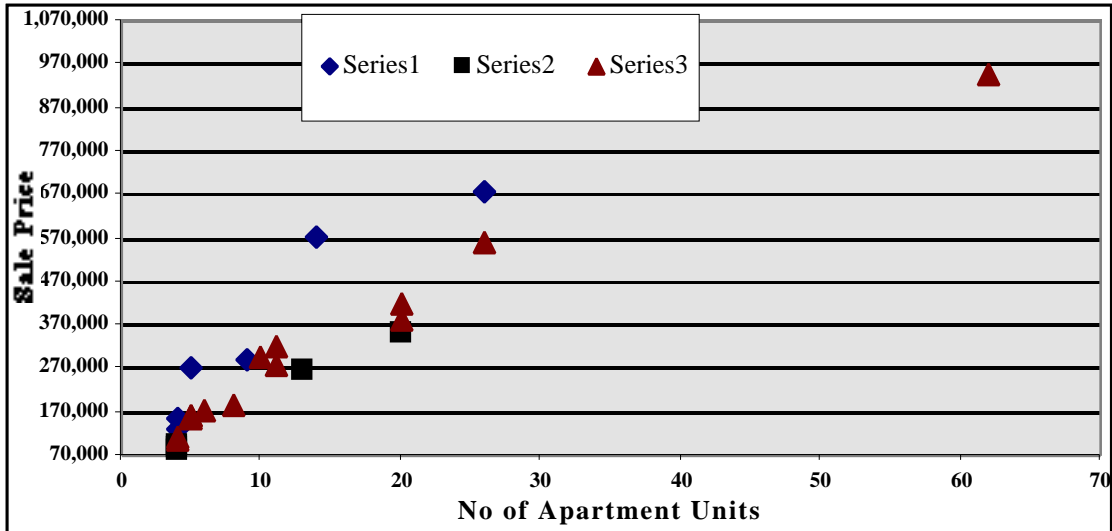
General model: $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$

$\hat{Y}_E = (b_0 + b_2) + b_1X_1$

$\hat{Y}_G = (b_0 + b_3) + b_1X_1$

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

- (ii) Plot Y against X₁ (number of apartment units) for all buildings in excellent condition. On the same graph, plot Y against X₁ 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.



Series 1: Apartment buildings rated Excellent
 Series 2: Apartment buildings rated Good
 Series 3: Apartment buildings rated Fair

From the scatter diagram Apartment buildings rated excellent tend to have a steeper slope than properties that are rated Fair. Property Code No. 0025 has a sale price of \$950,000 and is rated Good. This property is likely to bias the results from the model, it is shown as an outlier in the top right hand side of the diagram. With the exception of this building, there is some evidence to support the categorisation of Apartment buildings based on the rating system in (i)

- (iii) Fit the model from part (i) to the data. Report the least squares prediction equation for each of the three building condition levels.

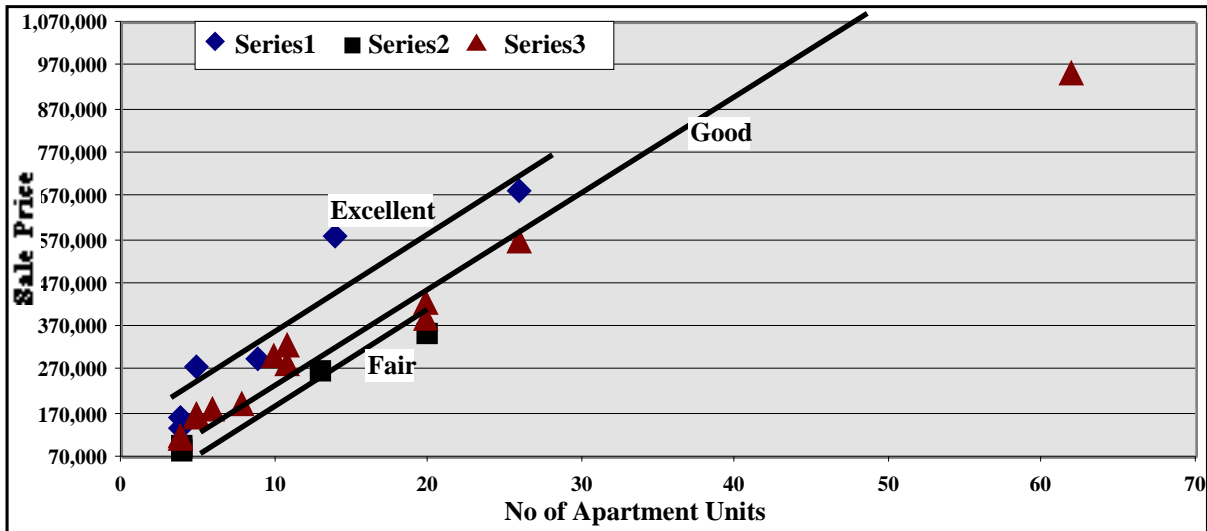
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | | |
|----------------|---------------------|-----------------------|---------------|----------------|----------------|----------|
| Intercept | 36387.640 | 30450.087 | 1.195 | 0.2454 | Standard Error | 64623.61 |
| X ₁ | 15616.929 | 1065.540 | 14.656 | 0.0000 | R Square | 0.918 |
| X ₂ | 152487.428 | 39157.324 | 3.894 | 0.0008 | Adj. R Square | 0.907 |
| X ₃ | 49441.146 | 34099.037 | 1.450 | 0.1619 | F | 78.71 |

$$\hat{Y}_E = 188,875 + 15616.929X_1$$

$$\hat{Y}_G = 85,829 + 15616.929X_1$$

$$\hat{Y}_F = 36,388 + 15616.929X_1$$

- (iv) Plot the three prediction equations of part (iii) on a scatter gram of the data.



- (v) Do 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 $\alpha = .05$.

The critical t-value for $\alpha = .05$ and $df = 25 - 4 = 21$ from tables is: 2.08

The t-ratio for X_3 is less than the critical t-value suggesting that Apartment buildings rated Good are not significantly different from other types of buildings. When the outlier is removed (property code 0025) from the data set and the model reestimated the t-ratio for X_3 is very close to 2, however, it still fails the 5% test. The constant term was also expected to be significant since it captures the category Fair, the results from this sample indicates that it is not. Both X_1 and X_2 are significant with X_1 contributing most of the explanation in the model. There is sufficient evidence to form the view that physical condition has a role to play in determining sale price.

15. (i) Distinguish between regression analysis and correlation analysis.

Regression analysis is the process of determining a relationship between a dependent variable and one or more independent variables. Correlation analysis is used to describe the **degree** to which the dependent variable is related to the independent variable(s).

- (ii) What is the correlation coefficient and what does it measure? What is the coefficient of determination and what does it measure?

Correlation coefficient: measures the strength of the relationship between two variables.
Coefficient of determination: also known as R^2 , is a measure of the proportion of the variance in the dependent variable explained by the independent variable(s).

- (iii) What is a dummy variable? When is it used? How is it interpreted?

Dummy variables are coded variables enabling the quantification of qualitative information, the coding is typically 0 or 1. They may be used to account for weather conditions affecting crop yields or the condition of a building. Dummy variables are included in a regression equation and are evaluated in exactly the same manner as other variables in the model. The interpretation of dummy variables depends on their use, typically they describe a shift upwards or downwards in the regression constant, indicating a positive or negative impact due to some qualitative attribute captured by the variable(s). Dummy variables may also be used to account for slope shifts; after some event the slope of the regression function may be steep or flatter.

(iv) What is a variable transformation? When is it employed?

Variable transformations are used to create a new variable from an existing variable. Ordinary least squares regression (OLS) is a linear estimation technique, therefore if variables are non linearly related they may be transformed before applying OLS. Question 6 provides some examples of variable transformations.

16. The data following 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

| | | | |
|-------|-------|-------|-------|
| | X_1 | X_2 | X_3 |
| X_1 | 1 | | |
| X_2 | 0.924 | 1 | |
| X_3 | 0.458 | 0.085 | 1 |
| Y | 0.843 | 0.878 | 0.142 |

(i) Estimate the models:

Model 1: $Y = f(X_1)$ SSR = 352.270 SSE = 143.120

| | Coefficients | Standard Error | t Stat | P-value | Standard Error |
|-----------|--------------|----------------|--------|---------|---------------------|
| Intercept | 2.820 | | | | |
| X_1 | -1.496 | 3.319 | -0.451 | 0.6576 | R Square 0.711 |
| | 0.857 | 0.129 | 6.656 | 0.0000 | Adj. R Square 0.695 |
| | | | | | F 44.305 |

Model 2: $Y = f(X_2)$ SSR = 381.966 SSE = 113.424

| | Coefficients | Standard Error | t Stat | P-value | Standard Error | |
|-----------|--------------|----------------|--------|---------|----------------|--------|
| Intercept | -23.634 | 5.657 | -4.178 | 0.00057 | R Square | 0.771 |
| X_2 | 0.857 | 0.110 | 7.786 | 0.00000 | Adj. R Square | 0.758 |
| | | | | | F | 60.617 |

Model 3: $Y = f(X_1, X_2)$ SSR = 385.439 SSE = 109.951

| | Coefficients | Standard Error | t Stat | P-value | Standard Error | |
|-----------|--------------|----------------|--------|---------|----------------|--------|
| Intercept | -19.174 | 8.361 | -2.293 | 0.0348 | R Square | 0.778 |
| X_1 | 0.222 | 0.303 | 0.733 | 0.4737 | Adj. R Square | 0.752 |
| X_2 | 0.659 | 0.291 | 2.265 | 0.0369 | F | 29.797 |

Model 4: $Y = f(X_1, X_2, X_3)$ SSR = 396.985 SSE = 98.405

| | Coefficients | Standard Error | t Stat | P-value | Standard Error | |
|-----------|--------------|----------------|--------|---------|----------------|--------|
| Intercept | 117.085 | 99.782 | 1.173 | 0.2578 | Standard Error | 2.480 |
| X_1 | 4.334 | 3.016 | 1.437 | 0.1699 | R Square | 0.801 |
| X_2 | -2.857 | 2.582 | -1.106 | 0.2849 | Adj. R Square | 0.764 |
| X_3 | -2.186 | 1.595 | -1.370 | 0.1896 | F | 21.516 |

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

If SSR increases then explained variance improves and the unexplained variance, represented by SSE decreases.

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?

$$\begin{aligned} \text{SSR}(X_3/X_1, X_2) &= \text{SSE}(X_1, X_2) - \text{SSE}(X_1, X_2, X_3) \\ &= 109.951 - 98.405 = 11.546 \end{aligned}$$

or

$$\begin{aligned} SSR(X_3/X_1, X_2) &= SSR(X_1, X_2, X_3) - SSR(X_1, X_2) \\ &= 396.985 - 385.439 = 11.546 \end{aligned}$$

- (ii) Is it appropriate to drop X_2 and X_3 from the model?

$$\begin{aligned} SSR(X_2, X_3/X_1) &= SSR(X_1, X_2, X_3) - SSR(X_1) \\ &= 396.985 - 352.270 = 44.715 \end{aligned}$$

$$F = \frac{SSR(X_2, X_3 | X_1)/df}{SSE(X_1, X_2, X_3)/df} = \frac{44.715/2}{98.405/16} = 3.635$$

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 ?

From F-distribution tables; $F_{0.05,2,16} = 3.63$ and $F_{0.01,2,16} = 6.23$

The extra sum of squares attributed to X_2 and X_3 is barely significant at the 5% level but not at the 1% level. In model 5 it can be observed that the t-ratios for X_2 and X_3 are not significant which tends to support the calculated F statistic at least at the 1% level.

- (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:

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

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.

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

Based on the analysis from parts (i) and (ii) Model 2 is the preferred model, this includes only X_2 . The correlation matrix indicates that X_1 and X_2 are highly correlated and X_2 has a stronger correlation with Y, the F-statistic for Model 2 is also significantly better than for Model 1. Testing for omitted variables produces the following results.

| SSE_1 | SSE_2 | $(n - k_2)$ | F | |
|----------|----------|-------------|--------|------------------------|
| 113.4237 | 113.4237 | 17 | 0.0000 | $F_{0.01,1,17} = 8.40$ |
| 113.4237 | 113.3454 | 16 | 0.0055 | $F_{0.01,2,16} = 6.23$ |
| 113.4237 | 111.0707 | 15 | 0.1059 | $F_{0.01,3,15} = 5.42$ |

We reject the hypothesis of omitted variables in each case.

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

| | Coefficients | Standard Error | t Stat | P-value |
|---------------|--------------|----------------|--------|---------|
| Intercept | 0.176 | 17.880 | 0.010 | 0.9922 |
| \hat{Y}_1^2 | -0.005 | 0.306 | -0.015 | 0.9884 |

| | | | | |
|---------------|-------|-------|--------|--------|
| \hat{Y}_i^3 | 0.000 | 0.021 | 0.017 | 0.9865 |
| \hat{Y}_i^4 | 0.000 | 0.000 | -0.020 | 0.9847 |

Since none of the t-ratios are significant we reject the hypothesis of omitted variables.

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.

Y - material and labour costs
 X_1 - area of building
 $X_2 = 1$ if air conditioning is to be installed
 $X_2 = 0$ if air conditioning is not to be installed

$$E(Y) = b_0 + b_1X_1 + b_2X_2$$

- (ii) Write a complete second-order model for the mean cost as a function of the same two variables.

$$E(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_1^2 + b_4X_1X_2 + b_5X_2^2$$

- (iii) What hypothesis would you test to determine whether the second-order terms are useful for predicting mean cost ?

$$H_0: b_3 = b_4 = b_5 = 0$$

$$H_1: \text{At least one of the } b_i \neq 0, i = 3, 4, 5.$$

- (iv) Explain how you would compute the F statistic needed to test the hypothesis of part (iii).

Denote the error sum of squares from the equation described in (i) as SSE_R and the error sum of squares from the equation in part (ii) as SSE_{UR} , then the F-statistic may be calculated as follows:

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

If F_c is less than the critical value from the F-distribution table for the required degrees of freedom and level of significance then H_0 is accepted.

19. 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 hypothesized 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)¹ 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.

Since the market is either bull or bear a single dummy variable captures these effects.

Define $X_2 = 1$ for a bull market and $X_2 = 0$ for a bear market

Model: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 X_1$

$$\hat{Y}_{\text{Bull}} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1 \qquad \hat{Y}_{\text{Bear}} = \beta_0 + \beta_1 X_1$$

- (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.

Under the bull market it is expected that the mutual funds return is higher, hence, β_2 should be a positive number. The coefficient β_3 determines whether the slope of the response function is different from a bull market to a bear market. If the market as a whole, represented by X_1 , is bullish then the mutual fund's performance should be better, if this is the case then the coefficient β_3 will be positive.

- (iii) Specify the hypothesis you would test to determine whether a mutual fund's beta coefficient is different during bull and bear markets.

$$H_0: \beta_2 = \beta_3 = 0$$

$$H_1: \text{Either } \beta_2 \text{ or } \beta_3 \text{ or both not equal zero.}$$

20. In question 19, the relationship between the behaviour of an individual mutual fund and the behaviour of the stock market as a whole was discussed. The table following 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).

- (i) Fit the model you developed in part (i) of Question 19 to the data shown in the table.

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | | |
|-----------|---------------------|-----------------------|---------------|----------------|----------------|--------|
| Intercept | -0.011 | 0.005 | -2.330 | 0.02277 | Standard Error | 0.024 |
| X_1 | 0.405 | 0.097 | 4.171 | 0.00009 | R Square | 0.658 |
| X_2 | 0.014 | 0.006 | 2.196 | 0.03149 | Adj. R Square | 0.643 |
| X_3 | 0.426 | 0.134 | 3.179 | 0.00222 | F | 43.566 |

¹ 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.

- (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.

$$\hat{Y}_{\text{Bull}} = (-0.011 + 0.014) + (0.405 + 0.426)X_1 = 0.003 + 0.831X_1$$

$$\hat{Y}_{\text{Bear}} = -0.011 + 0.405X_1$$

The beta coefficients are: $\beta_{\text{Bull}} = 0.831$ and $\beta_{\text{Bear}} = 0.405$

- (iii) Test the hypothesis you specified in part (iii) of Question 17. Draw the appropriate conclusions. Test using $\alpha = .05$.

The critical t-value with $df = 72 - 4 = 68$ for $\alpha = .05$ is 1.98 (for $\alpha = .01$ it is 2.617).

Since the t-ratios for β_2 and β_3 are greater than this critical value reject H_0 , the beta coefficient for bull markets is different to that for bear markets based on the sample data.