**A9Wl Introduction to multiple regression analysis**

In many situations it is unlikely that a dependent variable (y) would depend upon only one predictor variable (x), but on a number of independent variables (x1, x2, x3, ….., and xn). To solve problems with more than one independent variable we would be required to undertake **multiple regression analysis**. For example, house price might depend not only upon the land value but also upon the value of home improvements made to a property. The form of the population regression equation with ‘n’ independent variables can be written as:

 (8.23)

The multiple regression models can be found using the Excel ToolPak Regression tool to provide the coefficients, assumption and reliability checks, and conduct appropriate inference tests.

**Example**

Table 1 consists of data that has been collected by an estate agent who wishes to model the relationship between house sales price (£'s) and the independent variables: land value, LV (£'s) and the value of home improvements, IV (£'s). In order to fit the model the estate agent selected a random sample of size 20 properties from the 2000 properties sold in that year.

|  |  |  |
| --- | --- | --- |
| Selling Price (£’s), Y | Land Value (£’s), X1 | Home Improvements (£’s), X2 |
| 68900 | 5960 | 44967 |
| 48500 | 9000 | 27860 |
| 55500 | 9500 | 31439 |
| 62000 | 10000 | 39592 |
| 140000 | 18000 | 72827 |
| 45000 | 8500 | 27317 |
| 115000 | 15000 | 60000 |
| 144000 | 23000 | 65000 |
| 59000 | 8100 | 39117 |
| 47500 | 9000 | 29349 |
| 40500 | 7300 | 40166 |
| 40000 | 8000 | 31679 |
| 135800 | 20000 | 75000 |
| 45500 | 8000 | 23454 |
| 40900 | 8000 | 20897 |
| 80000 | 10500 | 56248 |
| 56000 | 4000 | 20859 |
| 37000 | 4500 | 22610 |
| 50000 | 3400 | 35948 |
| 22400 | 1500 | 5779 |

Table 1

The solution process includes:

* Construct a scattergram to identify relationships between the variables
* Fit multiple regression models to the sample data
* Check model assumptions
* Test model reliability using the multiple Coefficient of Determination and **adjusted r2**
* Test whether the predictor variables are a significant contributor to the overall model, F-test
* Test whether the predictor variables are significant contributors, t-tests
* Provide a 95% confidence interval for the population slopes

Construct scatter plot to identify possible model

Figure 1 Scatter plot of sales price versus land value



Figure 1

Figure 2 Scatter plot of sales price versus the value of home improvements



Figure 2

The two scatter plots suggest that a linear model would be appropriate for y vs. x1 and y vs. x2. It should be noted that in both scatter plots we do have some evidence that non-linear models may be more appropriate. This is due to the observation that the data points are starting to decrease in y value at the top range for x. It should also be noted that the sample sizes are quite small, and we will assume that both relationships are linear within the multiple regression model. From this analysis, we can identify three possible models identified in Table 2.

|  |  |  |
| --- | --- | --- |
|  | Population | Sample |
| Model 1 |  |  |
| Model 2 |  |  |
| Model 3 |  |  |

Table 2

Table 3 shows the results of applying least squares regression for models 1, 2 and 3. The results show that model 3 represents a better fit to the data than models 1 and 2.

|  |  |  |
| --- | --- | --- |
| Model | Equation | COD |
| 1 |  | 84% |
| 2 |  | 86% |
| 3 |  | 92% |

Table 3

 From the summary in Table 3, we observe that the third model is the best fit as 92% of variations in selling price are explained by the combined effect of both the land value and home improvements. Clearly, this is the most superior model.

To complete the solution, you would then need to check the model assumptions and undertake an appropriate t-test (or F-test) to test whether the independent variable is a significant contributor to the dependent variable. The examples given here serve just as an illustration to indicate that there is much more depth to regression analysis technique.