MIS770 Foundation Skills in Business Analysis

Department of Information Systems and Business Analytics
Deakin Business School
Faculty of Business and Law

Tutorial Topic 9

Introduction to Multiple Regression

Introduction

In this topic we will be looking at Multiple Regression which is an extension of the Simple Regression we learned in the last topic. The basic difference is the number of Independent variables. Remember, these independent variables attempt to explain the variation in the dependent variable.

A good example used to explain multiple regression is House Price, which we would classify as our dependent variable. There are many, many variables which would have an influence on the price of a house. These variables are known as the independent variables.

For those of you who have had the experience of a real estate agent coming to your house to give an estimate of its value, you would actually be experiencing multiple regression in action. The agent will take note of the land size, house size, number of bedrooms, number of bathrooms, the building material, the suburb you live in, even if there are trees in your street, plus many other variables. All of these elements, or variables, have an influence on the price of your house. There are some negative variables as well: how close you live to high voltage transmission lines, how close to a rubbish dump and how far away from public transport. They too have an effect on the price of a house, albeit negative.

Therefore, the aims of this tutorial are to:

- construct a multiple regression model and analyse model output
- determine which independent variables to include in the regression model, and decide which are more important in predicting a dependent variable
- incorporate categorical and interactive variables into a regression model
- detect collinearity using the variance inflationary factor (VIF)

Textbook Questions

13.3 A marketing analyst for a shoe manufacturer is considering the development of a new brand of tennis shoes. The marketing analyst wants to determine which variables to use in predicting durability (i.e. the effect of long-term impact). Two independent variables under consideration are X_1 (Foreimp), a measurement of the forefoot shock-absorbing capability, and X_2 (Midsole), a measurement of the change in impact properties over time. The dependent variable, Y, is LTIMP, a measure of the shoe's durability after a repeated impact test. A random sample of 15 types of currently manufactured tennis shoes was selected for testing, with the following results.

ANOVA	df	SS	MS	F	Significance F
Regression	2	12.61020	6.30510	97.69	0.0001
Error	12	0.77453	0.06454		
Total	14	13.38473			

Variable	Coefficients	Standard Error	t stat	p-value
Intercept	0.02686	0.06905	0.39	0.7034
Foreimp	0.79116	0.06295	12.57	0.0000
Midsole	0.60484	0.07174	8.43	0.0000

a. State the multiple regression equation.

b. Interpret the meaning of the slope coefficients b1 and b2 in this problem.

13.37 Suppose X_1 is a numerical variable and X_2 is a dummy variable and the following is the regression equation for a sample of *n* = 35 is:

$$\hat{Y}_i = 12 + 5X_{1i} + 0.5X_{2i}$$

a. Interpret the meaning of the slope for variable X_1.

b. Interpret the meaning of the slope for variable X_2.

c. Suppose that the *t* statistic for testing the contribution of variable X_2 is 2.67. At the 0.05 level of significance, is there evidence that variable X_2 makes a significant contribution to the model?

13.39 A real estate association in Melbourne would like to study the relationship between the size of a family house (measured by the number of rooms) and the selling price of the house (in thousands of dollars). Two different neighbourhoods are included in the study, one on the east Dandenong side (= 0) and the other on the west Sunshine side (= 1). A random sample of 20 houses was selected with the following results given in the file: [Dataset: HOUSE_SIZE.XLSX]

House	Selling Price ($,000)	Number of Rooms	Location
1	345	8	0
2	655	9	0
3	325	7	1
4	824	12	0
5	432	10	1
6	233	6	1
7	567	9	1
8	988	13	0
9	199	6	1
10	934	12	0
11	258	6	1
12	379	10	1
13	355	8	1
14	643	10	0
15	710	11	0
16	585	9	0
17	677	14	1
18	870	12	0
19	670	9	0
20	280	7	1

a. State the multiple regression equation.

b. Interpret the meaning of the slopes in this problem.

c. Predict the selling price for a house with nine rooms located in Melbourne's east and construct a 95% confidence interval estimate and a 95% prediction interval.

d. Perform a residual analysis on the results and determine the adequacy of the model.

e. Is there a significant relationship between selling price and the two independent variables (rooms and neighbourhood) at the 0.05 level of significance?

f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.

g. Construct 95% confidence interval estimates of the population slope for the relationship between selling price and number of rooms, and between selling price and neighbourhood.

h. Interpret the meaning of the coefficient of multiple determination.

i. Calculate the adjusted R2.

j. Calculate the coefficients of partial determination and interpret their meaning.

k. What assumption do you need to make about the slope of selling price with number of rooms?

l. Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.

m. On the basis of the results of (f) and (l), which model is more appropriate? Explain.

13.63 Academics have noticed that attendance at lectures has declined over the last decade or so. A marketing lecturer decides to collect data on 12 students' lecture attendance over a 13-week session. She believes that attendance is related to the distance a student has to travel to campus (in km) and how many weeks students access the video recording on the lecture attendance. **[Dataset: ATTENDANCE.XLSX]**

Lecture	Distance (km)	Video
12	5	4
4	35	6
2	45	8
6	12	7
5	23	7
13	2	5
8	12	4
7	25	5
3	40	7
2	25	9
9	21	4
8	8	5

a. State the multiple regression equation.

b. Interpret the meaning of the slopes in this equation.

c. Predict the attendance for a student who lives 5 km from campus and who accessed three video lectures.

d. Perform a residual analysis and determine the adequacy of fit of the model.

e. Is there a significant relationship between student attendance and the two independent variables (distance from campus and number of video lectures) at the 0.05 level of significance?

f. Determine the p-value in (e) and interpret its meaning.

g. Interpret the meaning of the coefficient of multiple determination.

h. Determine the adjusted R2.

i. At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.

j. Determine the p-values in (i) and interpret their meaning.

k. Calculate a measure of collinearity and interpret the result.

TEXTBOOK REFERENCE:

Basic Business Statistics: Concepts and Applications. *Berenson, M.L. Levine, D.M. Szabat, K.A. O'Brien, M. Jayne, N. Watson, J.* 5th edition. 2019. Pearson Australia Group Pty Ltd. ISBN 9781488617249. Chapter 13, sections 13 to 13.7.