Revision 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

13.4 A financial planner believes that retirement trends reflect the trends in the labour market and also share market returns. She collects data on 15 OECD countries' retirement rates (%), unemployment rates (%) and share market returns for 2011. **[Dataset: RETIRE.XLSX]**

Country	Retirement rate	Unemployment rate	Share market return
Australia	20	5	8
Canada	34	9	6
Denmark	22	8	3
Finland	16	6	6
France	32	8	6
Germany	17	7	5
Italy	20	8	2
Netherlands	18	9	4
New Zealand	19	6	7
Norway	24	5	3
Portugal	25	11	3
Spain	34	16	4
Sweden	14	4	7

United Kingdom	18	7	6
United States	15	8	7

a. State the multiple regression equation.

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

c. Explain why the regression coefficient, b0, has no practical meaning in the context of this problem.

d. Predict the mean retirement rate when unemployment is 6% and there is a share market return of 5%.

e. Construct a 95% confidence interval estimate for the mean retirement rate when unemployment is 6% and there is a share market return of 5%.

f. Construct a 95% prediction interval for the retirement rate when unemployment is 6% and there is a share market return of 5%.

13.40 A traffic engineer wants to predict the number of road accidents at intersections in a major urban area. He believes the main determinants are volume of traffic and whether or not the intersection has traffic lights. Traffic lights have the dummy value 1 and no traffic lights the value 0. A sample of 15 intersections is placed into a table.

Intersection	Number accidents per month	Traffic volume ('000/day)	Traffic lights
1	12	13	1
2	25	25	0
3	18	17	1
4	10	22	1
5	20	33	0
6	8	10	1
7	11	44	1
8	22	21	0
9	24	28	0
10	16	33	1
11	9	19	1
12	15	23	1
13	8	13	0
14	18	16	1
15	23	28	0

a. State the multiple regression equation.

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

c. Predict the number of accidents for an intersection with lights and 18,000 cars/day 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 the number of accidents and the two independent variables (traffic volume and traffic lights) 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 the number of accidents and traffic volume, and between accidents and traffic lights.

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

i. Calculate the adjusted R^2 and interpret the result.

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

k. What assumption about the slope of the number of accidents with volume of traffic do you need to make?

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.

13.62 What process should you follow to determine which variables should be included in a multiple regression?

13.70 The owner of a moving company typically has his most experienced manager predict the total number of labour hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labour hours. In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved and the number of pieces of large furniture as the independent variables, and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York, with the travel time an insignificant portion of the hours worked. The data are organised and stored in **[Dataset: MOVING.XLSX].**

a. State the multiple regression equation.

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

c. Predict the mean labour hours for moving 500 cubic feet with two large pieces of furniture.

d. Perform a residual analysis and determine whether the regression assumptions are valid.

e. Determine whether there is a significant relationship between labour hours and the two independent variables (the number of cubic feet moved and the number of pieces of large furniture) 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 r^2.

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

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

k. Construct a 95% confidence interval estimate of the population slope between labour hours and the number of cubic feet moved.

l. Calculate and interpret the coefficients of partial determination.

m. What conclusions can you reach concerning labour hours?

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.