Lab06: Regression Analysis

**Handout date:** Wednesday, October 07, 2019

**Due date:** Thursday, October 15, 2020 by midnight in the Lab06Submit eLearning folder  
 *This lab counts 4 % toward your total grade.*

**Objectives:** In this lab you   
[a] formulate explicit hypotheses in a multiple regression model;  
[b] explore the meaning of regression coefficients in a multiple model;

[c] explain the why a bivariate direction of influence may change in a multiple regression model.

**Format of answer:** Your answers (graphs and verbal description) should be handed in as ***hard-copy*** in ***one*** document. Add a running title into the header of the document with the following information: ***your name***, ***Lab06*** and ***page numbers***. Label each answer properly starting with its task number. Maintain the sequence of questions. For your answers, use ***Word's Equation Editor***. Format any code and computer output properly before inserting it into the document with your answer. -code and text output need to be in a ***monospaced*** font (i.e., fixed-pitch font) such as Courier New so proper spacing and alignments are preserved. Excessive, but irrelevant, output will lead to a deduction of your accumulated points.

# Part I: Multiple Regression Analysis Tasks (2.8 points)

Attach the libraries **car** and **effects** to your  session. Read information on 506 communities in the Boston area from the internet with the statement:

**hprice <- foreign::read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/hprice2.dta")**

We will focus on the variables:

|  |  |
| --- | --- |
| Variable | Description |
| price | Median home value in $100 in the community (dependent variable) |
| nox | Nitrogen-oxide (measure of traffic related air pollution) |
| dist | Weighted distance from 5 major employment centers |
| stratio | Student-teacher ratio in the community |
| rooms | Average number of rooms |

Your ***response variable*** is the **price** and the remaining variables are your ***exogenous variables***.

Task 1: Formulate ***explicit hypotheses*** and your ***common-sense*** rationale on how four exogenous variables will influence home price. Place your answer in the table below (0.4 points)

|  |  |  |
| --- | --- | --- |
| Independent Variable | Direction of Influence | Rationale |
| nox |  |  |
| dist |  |  |
| stratio |  |  |
| rooms |  |  |

Task 2: Generate a scatterplot matrix of the five variables **price**, **nox**, **dist**, **stratio** and **rooms**. Make sure that the dependent variable is the first one in the list. ***Thoroughly interpret*** the individual distributions of the variables and their pairwise relationships. (0.5 points)

Task 3: Run a multiple regression model of the **price** onto the four independent variables. Interpret the estimated regression coefficients with regards to your stated hypothesis in Task 1. (0.4 points)  
Also interpret the statistics. (0.3 points)

Task 4: Generate a scatterplot of the model’s regression residuals from Task 3 against each independent variable. Place the independent variables on the x-axis. Are the residuals independent of the exogenous variables? (0.4 points)

Task 5: Is there potentially a quadratic relationship with regards to independent variable **rooms**? (0.1 points)

Task 6: Rerun the model in Task 3 by augmenting it with the squared number of rooms as fifth exogenous variable.   
Note: the -formula statement in the **lm( )** function needs to wrap the squared number of rooms inside the inhibit function **I(rooms^2)**.  
Plot the residuals of the augmented model against the number of rooms. Has the potentially non-linear relationship been fixed? (0.2 points).

Task 7: Why does it become difficult to interpret the regression parameters associated with the number of rooms in the augmented quadratic model? Hint: look at the signs of the linear and the quadratic terms. (0.1 points)  
To enable the interpretation, visualize the quadratic relationship between the number of rooms and the home price with an effects plot. Note that the number of rooms ranges from 3 to 9. (0.2 points)

Interpret your non-linear effects plot. (0.2 points)

# Part II: Partial Regression Effects (1.2 points)

Open the SPSS file **Concord1.sav** in your  session.

Task 8: Formulate hypotheses on how the water consumption in 1981 (**water81**) is influenced by the household’s income (**income**) and the education level (**educat**) of the household’s head. Justify your hypotheses with common sense arguments. If you are not sure, then explain why. (0.4 points)

Task 9: Generate a scatterplot matrix of the three variables. Do the bivariate relationships between **water81~income** and **water81~educat** support your hypotheses from Task 8? (0. 4 points)

Task 10: Run the multiple regression model **water81~income+educat**. In comparison to the bivariate relationship in Task9 why does the interpretation of the education effect in the multiple regression model change? Consider the correlation between income and education in your argument. (0.4 points)