Lab03: Multivariate Regression

**Handed out:** Thursday, October 1, 2020

**Return date:** Friday, October 16, 2020 as by midnight at eLearning’s **Lab03Submit** tab.

**Grading:** This lab counts 8 % towards your final grade

**Objectives:** This lab exercises building and interpreting multivariate regression models, the use of the partial *F*-test and the concept of dummy variable as well as their interaction with metric variables.

**Format of answer:** Your answers (statistical figures and verbal description) should be submitted as ***hardcopy***. Add a running title with the following information: Lab03, your name and page numbers. You may use this document as template. Copy the requested statistical figures into your document. Trial and error answers will lead to a deduction of points. Label each answer properly with the bold task headings. You are expected to hand in professionally formatted answers: use a fixed pitch font, like **Courier New**, for any  code the use mathematical type-setting when equations are required. Copy and paste figures into your document. Make sure that each figure has a proper ***caption*** describing its content.

# Part I: Partial Regression Coefficients (3.5 points)

The SPSS dataset **StateSchool.sav** evaluates the impact of the average state’s expenditure (variable **expend**) in primary education per student onto their average state-wide performance in the SAT test (variable **sat**). A potentially confounding variable is the state-wide participation rate of students in the SAT test (variable **pctsat**). Note some states encourage the participation of all their students in the SAT test while in other states only good students with university aspirations take the SAT test; therefore, a selection bias can be expected. See Gruber (1999) at <https://www.researchgate.net/publication/327474149_Getting_What_You_Pay_For_The_Debate_Over_Equity_in_Public_School_Expenditures> for more information.

Note: please should ***not*** perform variable transformations in the task.

Task 1: Generate a scatterplot matrix of the three variables **sat**, **expend** and **pctsat**. Interpret their pairwise relationships. (0.5 points)

Task 2: Formulate ***explicit hypotheses*** about the ***direction*** in which the two independent variables **expend** and **pctsat** may influence the SAT performance **sat**.   
Use common sense arguments to ***justify*** your hypotheses. Explain, if you are unable to decide ***a priori*** into which the direction a particular independent variable may influence the dependent variable. (0.5 points)

Task 3: Evaluate each hypothesis individually with a bivariate regression models (0.5 points)  
 [a] and   
 [b] .   
Discuss the results in terms of the signs of the estimated regression coefficients and .

Task 4: Perform a multiple regression analysis with the two independent variables   
 [c] and ***fully interpret*** the results. (1.0 point)  
Pay in particular attention to any changes in the estimated slope parameters and of the bivariate models and the partial slope parameters and of the multiple regression model.   
If a regression coefficient changes substantially, why may this be the case?

Task 5: Plot the regression residuals from the multiple model [4] against the independent variable **pctsat** (placed on the *x*-axis). (1.0 points)

Does the scatterplot indicate a deviation from what we would expect if the model would be properly specified? Would it be advisable to add the SAT participation rate square to the model, i.e., ?

## Part II: A Multiple Regression Model with Factors and Partial *F*-test (3.5 points)

You will find the data needed for this part in the zipped data file **Italy.zip** from Lab01. (read the attribute file **PROVINCES.DBF** into  with the function **foreign::read.dbf**)

You will experiment with regression models that aim at explaining the 1994 fertility rate **TOTFERTRAT** (number of children born by a woman during her lifetime) within the 95 Italian provinces. The following independent variables are:[a] the metric illiteracy rate **ILLITERRAT** , [b] the metric average woman’s age at first marriage **FEMMARAGE**, [c] the metric divorce rate **DIVORCERAT**, [d] the metric televisions per household **TELEPERFAM** and [e] a regional factor **REGION** denoting whether a province is located on the islands of Sicily or Sardina, or in the southern, central or northern parts of Italy’s mainland.

Note: please do ***not*** perform variable transformations in the task.

Task 6: Use common sense arguments ***how*** the four metric variables will influence the provincial fertility rates. Use one or two sentences per explanation and formulate one or two-sided null and alternative hypotheses based on your explanation. Format everything in at table. (0.5 points)

|  |  |  |
| --- | --- | --- |
| Variable | Common Sense Arguments | Statistical Hypotheses |
| ILLITERRAT |  |  |
| FEMMARAGE |  |  |
| DIVORCERAT |  |  |
| TELEPERFAM |  |  |

Task 7: Generate a scatterplot matrix showing the dependent variable and the four metric independent variables. Also generate a parallel boxplot of **TOTFERTRAT~ REGION**. *Briefly interpret the scatterplot matrix and the boxplot.* (0.5 points)

Task 8: Run a base model multiple regression with the four metric variables to explain the variation of the fertility rates. Interpret this model [a] in the light of your earlier stated hypotheses in task 6, [b] the significances of the estimate regression coefficients and [c] the goodness of fit. (0.5 points)

Task 9: Calculate the *beta-coefficients* for the multiple model in task 8. Rank the independent variables *according to the absolute strength* *of their effects* on the fertility rates and plot the beta coefficients with the **coefplot( )** function. (1 point)

Task 10: Run the multiple regression model with the four metric variables plus the **REGION** factor to explain the variation of the fertility rates. (0.5 points)   
From the perspective of interpreting the augmented total fertility rate model, which model, i.e., task 8 or task 10, is more informative?

Task 11: Use a partial *F*-test to check whether the model in task 10 has improved the model fit of the base model in task 8 significantly. (0.5 points)  
That is, test the null hypothesis: against the alternative hypothesis is for at least one .

## Part III. Identification of the Underlying Model Structure (1 point)

Use the five data-frames **df1** to **df5** in the workspace **DummyInteraction.RData** for this task (the workspace can be imported with the **load( )** function). These data-frames comprise of three variables: **y** for the dependent variables, **d** for a binary ***dummy variable***, and **x** for a ***metric*** variable. Each of these data-frames is best described by one of these competing models:

|  |  |
| --- | --- |
| Name | Models Structure |
| Full interaction model | **mod.full <- lm(y~d\*x, data=df)** |
| Different intercept model | **mod.int <- lm(y~d+x, data=df)** |
| Different slope model | **mod.slope <- lm(y~d:x, data=df)** |
| Just different means model | **mod.means <- lm(y~d, data=df)** |
| Plain bivariate model | **mod.plain <- lm(y~x, data=df)** |

For each of the data-frame generate an informative scatterplot showing the regression regimes for both groups. You can employ the syntax **car::scatterplot(y~x|d, data=df)**. Then identify which of the competing model structures best describes the given datasets.

Task 12: Plot the data and identify the underlying model structure for **df1**. (0.2 points)

Task 13: Plot the data and identify the underlying model structure for **df2**. (0.2 points)

Task 14: Plot the data and identify the underlying model structure for **df3**. (0.2 points)

Task 15: Plot the data and identify the underlying model structure for **df4**. (0.2 points)

Task 16: Plot the data and identify the underlying model structure for **df5**. (0.2 points)