# Lab02: Multiple Regression Analysis, Factors and Interaction Effects

**Handed out:** Monday, March 1, 2021

**Return date:** Friday, March 12, 2021, at eLearning’s **Lab02Submit** link.

**Grading:** This lab counts 13 % towards your final grade

**Objectives:** In this lab you will explore the meaning of the partial regression coefficients, and the partial *F*-test and you will build and analyze a full-fledge multiple regression model which includes a factor.

**Format of answer:** Your answers (statistical figures and verbal description) should be submitted as ***hardcopy***. Add a running title with the following information: Lab02, 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 and sub-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.

## Task 1. Partial Regression Coefficient [3 points]

Use the **Concord1.sav** file for this task. You will demonstrate that in multiple regression the partial effect of an independent variable is free from any linear effects of the remaining independent variables in a regression model.

**Task 1.1:** Run the multiple model **water81~income+water80+educat** and *interpret* its regression coefficients. [0.5 points]

**Task 1.2:** Calculate the residuals of the two models [a] **water81~income+water80** and [b] **educat~income+water80**. *What are these residuals specifically measuring*? [1 point]

**Task 1.3:** Generate the partial regression leverage scatterplot of the water residuals against the education residuals. Make sure to use properly labeled axes. *Briefly interpret the scatterplot*. [0.5 points]

**Task 1.4:** Estimate a regression model of the *water residuals* on the *education residuals* and *compare* its estimate slope coefficient against the slope coefficient for **educat** of the multiple model from task 1.1. *Why are you allowed to suppress the intercept in this model*? [1 point]

## Task 2: A Multiple Regression Model with Factors and Partial *F*-test [7 points]

Use the dBase data file **provinces.dbf** (read it into  with the function **foreign::read.dbf**)

You will experiment with regression models that aim at explaining the 1994 total 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 Sardinia, or in the southern, central or northern parts of Italy’s mainland.

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

**Task 2.1:** Use common sense arguments ***how*** these four metric variables will influence the provincial fertility rates. Use one or two sentences per explanation, and formulate preferably *one* sided null and alternative hypotheses based on your explanation. The statistical hypotheses should be *type-set* properly, for instance, as against . Format everything in the table shown below. [1 point]

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

**Task 2.2:** Generate a scatterplot matrix showing the dependent variable and the four metric independent variables. Also generate a boxplot of the fertility rate against the regions. *Briefly interpret the scatterplot matrix.* [1 point]

**Task 2.3:** 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 2.1, [b] the significances of the estimate regression coefficients and [c] the goodness of fit. [1 point]

**Task 2.4:** Calculate the standardized *beta-coefficients* for the multiple model in task 2.3. 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. Use proper options for the **coefplot( )** function. [1 point]

**Task 2.5:** Run five separate regressions on the [a] independent variables as well as [b] the dependent fertility rate using the factor **REGION** as independent variable.   
Does the **REGION** factor *explain variation* of the four independent variables as well as the fertility rate, i.e., what are the factor’s ’s? [1 point]

Hint: To calibrate all five models with one function call you can use the regression formula syntax **cbind(TOTFERTRAT,ILLITERRAT,FEMMARAGE,DIVORCERAT,TELEPERFAM)~REGION**.  
The **summary** function gives you the results for all five models.

**Task 2.6:** Run the multiple regression model with the four metric variables plus the **REGION** factor to explain the variation of the fertility rates.   
*Speculate* in an informed way why some independent metric variables are no longer significant? [1 point]

**Task 2.7:** Use a partial *F*-test to check whether the model in task 2.6 has improved the model fit of the base model in task 2.3 significantly. [1 point]   
That is, test the null hypothesis: against the alternative hypothesis is *for at least one* .

## Task 3. Identification of the Underlying Model Structure [3 points]

Use the workspace **ModelSpecs.RData** for this task. It contains the six data-frames **mod1** to **mod6**. Each data-frame is comprised of three variables: **y** for the dependent variables, **g** for a binary ***factor***, and **x** for a ***metric*** variable. Each of these data-frames is best ***statistically*** described by one of these competing models:

|  |  |
| --- | --- |
| Name | Models Structure |
| Full interaction model | **lm(y~g+x+g:x, data=mod?)  lm(y~g\*x, data=mod?)** |
| Intercept model | **lm(y~g+x, data=mod?)** |
| Slope model | **lm(y~g:x, data=mod?)** |
| Means model | **lm(y~g, data=mod?)** |
| Plain regression model | **lm(y~x, data=mod?)** |

For each of the data-frame generate an informative scatterplot showing the regression regimes for both groups of observations. You can employ the syntax:

**car::scatterplot(y~x|g,smoother=F,boxplots="xy",data=mod???,main="Model???")**

Then identify, which of the competing model structures best describes the given data-frame. By visual inspection extend one of the regression lines to to check if both lines share an identical intercept. If several competing model structures seem to be reasonably relevant, then try to eliminate inferior models by looking at their or the -values of the most elaborated model **lm(y~g\*x, data=mod?)**.

**Task 3.1:** Identify the underlying model structure for **mod1**. [0.5 points]

**Task 3.2:** Identify the underlying model structure for **mod2**. [0.5 points]

**Task 3.3:** Identify the underlying model structure for **mod3**. [0.5 points]

**Task 3.4:** Identify the underlying model structure for **mod4**. [0.5 points]

**Task 3.5:** Identify the underlying model structure for **mod5**. [0.5 points]

**Task 3.6:** Identify the underlying model structure for **mod6**. [0.5 points]