Lab04: Matrix Operations & Model Diagnostics

**Handed out:** Thursday, October 15, 2020

**Return date:** Friday, October 30, 2020 by midnight into the link **Lab04Submit** in eLearning.

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

**Objectives:** This lab practices in part 1 operations with matrices which are relevant to regression analysis and explore properties of different factor coding schemes. Part 2 focuses on model building and standard model diagnostics.

**Format of answer:** Your answers (statistical figures and verbal description) should be submitted as ***hardcopy***. Add a running title with the following information: Lab04, 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 1: Matrix Operations (4 points)

## Task 1: Manual matrix operations and regression analysis with matrices [1 points]

You are given a vector of the dependent variable and the design matrix

[a] Enter and into . Write your own OLS  function using the dependent vector and the associated design matrix as input. Your function should return the vector of the estimated regression coefficients. (0.5 point)

[b] Use 's matrix operations to calculate for a dependent variable , the design matrix and the diagonal weights matrix the weighted regression coefficients with the formula . (0.5 points)

[c] Compare the estimated regression coefficients from task 1 [a] with those from task 1 [b]. Explain why they are identically. Hint: what is the effect of the weights matrix . (0.5 points)

### Task 2: Coding schemes of categorical variables (3 points)

Provide the  syntax code of your answers. You can either use the **lm(…)** or your coded ordinary least squares function for this task

[a] Enter the matrix and the design matrices to separate matrix objects into  and show these object in your answer (0.5 points):

and are given in the ***indicator coding*** scheme ( codes it as **contrasts(factor) <- "contr.treatment"**) whereas and are given in the ***centered coding*** scheme ( codes it as **contrasts(factor) <- "contr.sum"** and Hamilton p 99 calls it ***effect*** coding). In and the last category is suppressed, whereas in and the second category is suppressed due to the redundancy among a full set of indicator variables.

[b] Calculate the three group means of the observations , and as well as the global mean for all observations . (0.5 points)

[c] Find the four sets of estimated regression coefficients for the intercept and group coefficients by regressing on the four design matrices , , and with your linear regression function that you have developed in task 1 [a] and enter these estimates into the table below (see columns *Assign Estimated Regression Coefficients*). (0.5 points)

Hints: (i) in the ***centered*** coding scheme the coefficient for the missing category can be calculated as the ***negative sum*** of the two other estimated parameters, i.e., . (ii) For the cornered coding scheme the values for the ***dashed*** cells cannot be calculated from the regression results.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Assign Estimated Regression Coefficients | | | | Give Expressions for the Means in Terms of the Estimate Regression Coefficients | | | |
| Model | **Coding** |  |  |  |  |  |  |  |  |
| y~X1 | ***cornered*** |  |  |  | ─ | ─ |  |  |  |
| y~X2 | ***cornered*** |  |  | ─ |  | ─ |  |  |  |
| y~X3 | ***centered*** |  |  |  |  |  |  |  |  |
| y~X4 | ***centered*** |  |  |  |  |  |  |  |  |

[d] For each design matrix the global mean and group means , and can be expressed as a function of the estimated regression coefficients in the columns *Assign Estimated Regression Coefficients*. (0.5 point)  
Find the expressions for the means and write them into columns labels by “*Give Expressions…*” using the parameter symbols, e.g., and note that this is an invalid expression.

[e] Which coding scheme has a more *intuitive interpretation*? Justify your answer. (0.5 points)

## Part 2: Model Building and Diagnostics (4 points)

Open the **CPS1985** data-frame with **data("CPS1985",package="AER")**. Assign new row-names with the statement **rownames(CPS1985) <- 1:nrow(CPS1985)** to the data-frame. ***Study the description*** of the variable **experience** in the associated online help.

### Task 3: Multicollinearity diagnostics (2 points)

[a] For the variables **~log(wage)+education+age+experience** generate a scatterplot matrix. (0.5 points)

Based on the definition of the variables and the scatterplot matrix, which variables do you expect to be multicollinear? Justify your decisions.

[b] Estimate the model **log(wage)~education+experience** and calculate the ***variance inflation factors***. Fully interpret the estimated model and the ***VIF***s. (0.5 points)

[c] Estimate the augmented model **log(wage)~education+experience+age** and show the output. (1 points)

Address the following points:

1. What do the ***VIF***s tell you?
2. ***What*** happened to the significances of the *t*-tests for the estimated regression parameters of the augmented model and ***why***?
3. Why does the global *F*-test still remain significant?

### Task 4: Refined model specification (1 point)

[a] Estimate the model: **log(wage)~education+experience+gender+occupation+union** and *fully interpret* the estimated regression model. (0.5 point)

[b] Test whether the factor **occupation** is significant and if necessary refine the model specification accordingly. (0.25 points)

[c] Investigate the model with **car::residualPlots( )**. Discuss the output and decide whether it is advisable to refine the model. (0.25 point)

### Task 5: Case statistics of the final model (1 point)

[a] Generate the following plots and *interpret* them for your final model. (0.75 points)

1. Identify the two most extreme observations with a **car::qqPlot( )** and interpret it.
2. Identify *potential* extreme observations with a **car::influenceIndexPlot( )** and interpret the plots.
3. Identify the two most extreme observation with a **car::avPlots( )** and interpret the plots.

[b] Inspect the ***two*** most extreme observations in the data-frame by examining their records. (0.25 points)

1. Discuss their attributes and argue if they are representative of the underlying population.
2. Drop them from the data-frame and show your code for doing so.