Pol Sci 630: Problem Set 5 - Regression Model Interpretation

Prepared by: Jan Vogler (jan.vogler@duke.edu)

Due Date: Wednesday, October 5th, 2016, 1.25 PM (Beginning of Class)

Note 1: It is absolutely essential that you show all your work, including intermediary steps, in your (mathematical) calculations and that you comment on your R code to earn full credit (you can comment on your R code both with the use of # in the R code and in the LaTeX code). Showing all steps and commenting on code will also be required in future problem sets.

Note 2: Please submit a PDF file created through knitr containing all your answers to the problem set. knitr allows you to combine R code and LaTeX code in one document, meaning that you can include both the answers to R programming and math problems. Also submit the source code that generates the PDF file (i.e. the .Rnw file).

Note 3: Make sure that the PDF files you submit do not include any references to your identity. The grading will happen anonymously. You can submit your answer at the following website: `http://ps630-f15.herokuapp.com/`

R Programming

Problem 1 (4 points)

Do the following in R:

a) Load the *swiss* dataset in R via the command *data(swiss)*. According to the documentation this is data on "Standardized fertility measure and socio-economic indicators for

each of 47 French-speaking provinces of Switzerland at about 1888". Display the summary statistics for every variable in the dataset.

We are interested in how the level of "Education" is related to the other variables. For this purpose, use the *lm* (linear model) function to estimate a regression of *Education* on all other variables in the dataset.

b) Produce a table of the results via the R package *stargazer* and include it in your document. Interpret the results for every variable in the linear model. Be specific about marginal effects and the meaning of p-values. State at which levels of significance each estimated coefficient is statistically significant. Then, interpret the R^2 statistic and the F-statistic.

Problem 2 (4 points)

Do the following in R:

a) Load the *VOTE1* dataset that you can find on the course website under "meetings". This dataset contains information on elections of the United States House of Representatives. Each election has one incumbent party A and one challenger party B. Display the summary statistics of *all variables* and explicitly describe the summary statistics of the *following variables*:

1. VoteA (the vote share that party A, the incumbent party, gained in the election)
2. expendA (the expenditures that party A, the incumbent, had in the election)
3. expendB (the expenditures that party B, the challenger, had in the election)
4. partystrA (the strength of party A, the incumbent party)

We are interested in the relationship between *the vote share that party A received*, and other variables in the dataset.

Which of the above factors may influence the vote share that party A receives and in which way? Consider all of the above variables. Choose one of them *before* running any model and explain which kind of relationship—positive, negative, or no influence—you would expect. Make a succinct claim and formulate a hypothesis that you can test empirically.

Note: Your claim does not have to be true. Just do your best to make a plausible claim. Your claim itself will not be evaluated, just your empirical test of it.

b) Run a linear regression with the variable you chose as the main predictor variable (independent variable). Include the other variables as control variables.

c) Produce a table of the results via the R package *stargazer* and include it in your document. Then, interpret the results with respect to your variable, the R^2 statistic, and the F-statistic.

d) Show the marginal effect of your main predictor variable graphically, holding all other variables at their mean value. Include confidence intervals in your graphic.

Statistical Theory: Linear Regression Models

Problem 3 (4 points)

Answer the following questions.

a) What types of relationships between variables can be adequately modeled by OLS regression without including polynomials of higher order than one (i.e. without including squared terms etc.)? Explain carefully.

b) What can one generally say about causality regarding the statistical relationships identified by OLS regression? Explain carefully.

c) How can you calculate the total sum of squares (TSS), residual sum of squares (RSS), and the regression sum of squares (RegSS) if you have knowledge of the following three values (1.) the root mean squared error (RMSE), (2.) the degrees of freedom, and (3.) the R^2 statistic? Explain carefully and show mathematically.

Problem 4 (4 points)

Do the following problems. Show every step.

a) You have empirical data on two variables, X and Y. For both variables you have several different values. You estimate two OLS regressions: Y $= \beta_0 + \beta_1 x + u$ and X $= \beta_0' + \beta_1' y + u$. Is it true or false that the slopes of β_1 and β_1' will always have the same sign (i.e. positive, negative, or zero)? Explain carefully and show mathematically.

b) Following the task above, is it true or false that the intercepts β_0 and β_0' will always have the same sign (i.e. positive, negative, or zero)? Explain carefully and show mathematically.