Pol Sci 630: Problem Set 5 - Regression Model Interpretation

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

Due Date: Tuesday, October 2nd, 2015, 10 AM (Beginning of Class)

Note 1: It is absolutely essential that you show all your work, including interme-diary steps, and comment on your R code to earn full credit. Showing all steps and commenting on code them will also be required in future problem sets.

Note 2: Please use a *single* PDF file created through knitr to submit your answers. 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. either .Rnw or .Rmd files)

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

Do the following in R:

a) (1 point) 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. Note that R is case-sensitive.

b) (3 points) 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 relationship is statistically significant.

Problem 2

Do the following in R:

a) (2 points) Load the *LDC_IO_replication* dataset that was introduced in the tutorial. We are interested in the relationship between *fdignp*, which is defined as "Net change in foreign investment in the reporting country", and other variables in the dataset. This variable shows changes in terms of investment coming from other countries. Which political and economic factors may influence foreign investment? Consider all the variables that we used in the tutorial when we predicted the level of trade barriers. Choose one of them *before* running any model and explain which kind of relationship — positive, negative, or no influence — you would expect. Formulate a succinct hypothesis that you can test empirically.

b) (2 points) Run the regression with the variable you chose as the main predictor variable (independent variable). Include all other variables that we used previously as control variables. Interpret the results with respect to your variable. Then, graphically show the marginal effect of your main predictor variable when you hold all other variables at their mean value and include confidence intervals in your graphic.

Regression Model Interpretation Issues

Problem 3

Answer the following questions.

a) (1 point) What types of relationships between variables can be adequately modeled by OLS regression?

b) (1 point) What is a probabilistic relationship and how is it different than a deterministic relationship? Which type of relationship can be modeled by OLS?

c) (2 points) What can you say about causality with respect to the statistical relationships identified by OLS regression?

Statistical Theory: Regression Model Estimation

Problem 4

Do the following problems. Show every step.

a) (2 points) Assume that you have the following linear model
$$Y = 10 + 5 * X_1 + (-2) * X_2 + (-1) * X_1 * X_2 + \epsilon$$
Calculate the first derivative of Y with respect to X1 and X2. Then plot the effect that X_1 has on Y at different levels of X_2. Assume that $X2 \in [-10, 10]$. You may use R to plot the effect of X_1 on Y.

b) (2 points) Assume that you have two sets of points from the data on which the model was estimated:

1. $Y = 16, X_1 = 6, X_2 = 3$

2. $Y = 17, X_1 = 7, X_2 = 3$

Please plug X_1 and X_2 into the model and write down the value of Y that you obtain.

Is any of the two sets of points inconsistent with the linear model estimated above? Explain your answer carefully.