# 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)

It is absolutely essential that you show all your work, including intermediary 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.

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 (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 *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 investment coming from other countries (foreign direct investment) in terms of GDP. For a more detailed description of the variable please see the codebook.

Which political and economic factors may influence foreign investment in developing countries and in which way? 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. Make a succinct claim and formulate a hypothesis that you can test empirically.

Note: Your claim does not have to be true. If you are not familiar with topics of political economy, then 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 all other variables that we used previously as control variables.

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. Finally, 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: $\hat{\mathrm{Y}} = a + bx + \epsilon$ and $\hat{\mathrm{X}} = a' + b'y + \epsilon$. Is it true or false that the slopes of b and b' 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 a and a' will always have the same sign (i.e. positive, negative, or zero)? Explain carefully and show mathematically.