Pol Sci 630: Problem Set 4 - Regression Model Estimation

Prepared by: Anh Le (anh.le@duke.edu)

Due Date: Tuesday, September 22nd, 2015, 10 AM (Beginning of Class)

Note 1: It is absolutely essential that you show all your work, including intermediary steps, and comment on your R code to earn full credit.

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. .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/`

1. Create a data frame (4 points)

a)

First, `set.seed(2)`. Then, create a data frame with 1000 rows and 3 variables as follows:

1. `var_norm`: a normal variable with mean = 5, sd = 10
2. `var_binom`: a binomial variable with number of trial = 10, probability of success = 0.5
3. `var_poisson`: a Poisson variable with $\lambda = 4$

(Recall how to generate random sample from various distributions from previous labs.)

b)

Plot the histograms of the three variables, arranging them nicely (with `fig.width()`, `fig.height()`, `par(mfrow)` as you see fit). Brownie point if you plot using a for loop instead of writing `hist` three times.

2. Subset data frame (4 points)

a)

Download the following data from `WDI` and clean it as follows. Briefly comment on what each command does.

```
library(WDI)

## Loading required package:  RJSONIO

d_wdi <- WDI(indicator = c("NY.GDP.PCAP.CD", "SP.DYN.IMRT.IN", "SH.MED.PHYS.ZS"),
             start = 2005, end = 2010, extra = TRUE)
d_wdi <- d_wdi[d_wdi$region != "Aggregates",
       c("country", "year", "NY.GDP.PCAP.CD", "SP.DYN.IMRT.IN", "SH.MED.PHYS.ZS")]
colnames(d_wdi)[3:5] <- c('gdppc', 'infant_mortality', 'number_of_physician')
d_wdi <- na.omit(d_wdi)
```

`infant_mortality`: number of mortality per 1000 live births
`number_of_physician`: number of physician per 1000 people

b)

Use subsetting techniques to do the following:

1. Show the GDP per capita of Brazil across years
2. Show the country-years where infant mortality $>$ 100 per 1000 live birth
3. Show the country-years where GDP per capita is above average
4. Show the country-years where GDP per capita is above average, but number of physician is below average

3. Build linear model (4 points)

a)

Download 2 variables of interest from `WDI` and build a linear model of their relationship using `lm()`. Show the `summary()` of results

b)

Show the result with `stargazer`, customizing:

- The labels of the independent variables (i.e. the covariate)
- The label of the dependent variable

- Make the model name (i.e. OLS) show up

Hint: The options to do those things are in `help(stargazer)`. I have worded the task in a way that should help you find the relevant options.

4. Calculate sum of squares and RMSE (4 points)

1. Extract the residuals and predicted values (fitted values) from the model object (from Question 3)
2. Calculate three "sum of squares" (TSS, RegSS, RSS)
3. Calculate the root mean square error and compare with R. (In R and stargazer, RMSE is called "Residual standard error".)

Note: the data you feed to `lm()` may have missing data, so R has to modify the data a little before using it. To extract the data that are actually used by `lm()`, use `my_model$model`. Use this data to calculate $\bar{y}$ in the sum of squares.