Pol Sci 630: Problem Set 4 - Regression Model Estimation

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

Due Date: Wed, September 28, 2015 (Beginning of Class)

1 Subset data frame

1.1 Download data

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

1.2 Subsetting

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

2 Build linear model

2.1 Download

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

2.2 Calculate the regression coefficients WITHOUT using 'lm'

Use the mathematical formula of the regression coefficients you saw in class and implement it in R. Is this result the same as the result output by 'lm'?

2.3 Model output

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.

3 Calculate sum of squares and RMSE

1. Extract the residuals and predicted values (fitted values) from the model object (from the linear model built above)
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.