concordance=TRUE

Pol Sci 630: Problem Set 4 Solution - 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

Solution

```
library(WDI)

# Download data from WDI, specifying the indicators and start / end year
d_wdi <- WDI(indicator = c("NY.GDP.PCAP.CD", "SP.DYN.IMRT.IN", "SH.MED.PHYS.ZS"),
             start = 2008, end = 2010, extra = TRUE)

# Remove aggregates rows, selecting wanted columns by name
d_wdi <- d_wdi[d_wdi$region != "Aggregates",
       c("country", "year", "NY.GDP.PCAP.CD", "SP.DYN.IMRT.IN", "SH.MED.PHYS.ZS")]

# Rename some of the columns
colnames(d_wdi)[3:5] <- c('gdppc', 'infant_mortality', 'number_of_physician')

# Remove all rows that have missing data
d_wdi <- na.omit(d_wdi)
```

```
# 1. Show the GDP per capita of Brazil across years
d_wdi[d_wdi$country == "Brazil", c("country", "year", "gdppc")]

##    country year     gdppc
## 94  Brazil 2008  8706.819
## 95  Brazil 2009  8474.881
## 96  Brazil 2010 11121.421

# 2. Show the country-years where infant mortality > 100 per 1000 live birth
d_wdi[d_wdi$infant_mortality > 100, c("country", "year", "infant_mortality")]

##                      country year infant_mortality
## 34                    Angola 2009            112.2
## 119 Central African Republic 2008            105.5
## 120 Central African Republic 2009            103.6
## 568             Sierra Leone 2010            107.0
## 570             Sierra Leone 2008            116.2

# 3. Show the country-years where GDP per capita is above average
d_wdi[d_wdi$gdppc > mean(d_wdi$gdppc), c("country", "year", "gdppc")]

##                  country year     gdppc
## 16               Andorra 2009  42701.45
## 17               Andorra 2010  39639.39
## 19  United Arab Emirates 2009  32905.05
## 20  United Arab Emirates 2008  45720.02
## 21  United Arab Emirates 2010  34341.91
## 44               Austria 2010  46659.84
## 46             Australia 2009  42715.13
## 47             Australia 2010  51845.65
## 63              Barbados 2010  15901.43
## 67               Belgium 2010  44382.88
## 68               Belgium 2008  48424.59
## 76               Bahrain 2010  20386.02
## 77               Bahrain 2008  23043.03
## 78               Bahrain 2009  19166.71
## 88     Brunei Darussalam 2009  27726.48
## 89     Brunei Darussalam 2010  31453.22
## 90     Brunei Darussalam 2008  37798.39
## 99          Bahamas, The 2008  23657.37
## 112               Canada 2008  46596.34
## 114               Canada 2010  47445.76
## 124          Switzerland 2010  74277.12
## 154               Cyprus 2008  34950.35
## 155               Cyprus 2009  31673.46
## 156               Cyprus 2010  30438.90
## 157       Czech Republic 2010  19763.96
## 158       Czech Republic 2008  22649.38
## 160              Germany 2010  41788.04
## 162              Germany 2008  45699.20
## 166              Denmark 2010  57647.67
## 167              Denmark 2009  57895.50
## 168              Denmark 2008  64181.99
## 181              Estonia 2008  18094.55
## 183              Estonia 2010  14641.40
## 190                Spain 2009  32333.47
## 191                Spain 2010  30737.83
## 192                Spain 2008  35578.74
## 202              Finland 2010  46205.17
## 203              Finland 2008  53401.31
## 204              Finland 2009  47107.16
## 214               France 2008  45413.07
## 215               France 2010  40705.77
## 222       United Kingdom 2010  38292.87
## 247               Greece 2010  26919.36
## 248               Greece 2008  31997.28
## 268              Croatia 2008  15893.86
## 275              Hungary 2008  15649.72
## 280              Ireland 2008  61189.73
## 282              Ireland 2010  48260.67
## 283               Israel 2010  30736.36
## 298              Iceland 2010  41620.07
## 299              Iceland 2008  55229.61
## 300              Iceland 2009  40362.04
## 301                Italy 2009  36976.85
## 302                Italy 2010  35851.51
## 303                Italy 2008  40640.18
## 313                Japan 2010  42935.25
## 315                Japan 2008  37865.62
## 337          Korea, Rep. 2008  20474.89
## 338          Korea, Rep. 2010  22151.21
## 339          Korea, Rep. 2009  18338.71
## 340               Kuwait 2009  36754.95
## 341               Kuwait 2008  54484.30
## 342               Kuwait 2010  37725.14
## 371            Lithuania 2008  14961.57
## 373           Luxembourg 2010 103267.28
## 377               Latvia 2008  16323.77
## 381                Libya 2008  14231.60
## 425                Malta 2010  19694.08
## 426                Malta 2009  19636.01
## 460          Netherlands 2010  50341.25
## 462          Netherlands 2008  56928.82
## 463               Norway 2010  87646.27
## 464               Norway 2009  80017.78
## 465               Norway 2008  96880.51
## 472          New Zealand 2009  28200.94
## 473          New Zealand 2008  31287.61
## 474          New Zealand 2010  33692.17
## 478                 Oman 2009  17518.83
## 479                 Oman 2010  19920.65
## 480                 Oman 2008  22963.38
## 501               Poland 2008  13906.22
## 508             Portugal 2008  24815.61
## 509             Portugal 2010  22540.00
## 510             Portugal 2009  23063.97
## 517                Qatar 2008  82990.07
## 518                Qatar 2010  70870.23
## 519                Qatar 2009  61463.90
## 544         Saudi Arabia 2008  19436.86
## 545         Saudi Arabia 2009  15655.08
## 546         Saudi Arabia 2010  18753.98
## 556               Sweden 2008  55746.84
## 557               Sweden 2010  52076.43
## 558               Sweden 2009  46207.06
## 559            Singapore 2009  38577.56
## 560            Singapore 2010  46569.68
## 561            Singapore 2008  39721.05
## 562             Slovenia 2010  23438.85
## 564             Slovenia 2008  27501.82
## 565      Slovak Republic 2009  16460.22
## 566      Slovak Republic 2010  16554.88
## 650  Trinidad and Tobago 2010  15840.44
## 664        United States 2010  48374.09
## 665        United States 2009  47001.56
## 666        United States 2008  48401.43
```

```
# 4. Show the country-years where GDP per capita is above average,
# but number of physician is below average
d_wdi[d_wdi$gdppc > mean(d_wdi$gdppc) &
        d_wdi$number_of_physician < mean(d_wdi$number_of_physician),
      c("country", "year", "gdppc")]
```

```
##                 country year    gdppc
## 76              Bahrain 2010 20386.02
## 77              Bahrain 2008 23043.03
## 78              Bahrain 2009 19166.71
## 88    Brunei Darussalam 2009 27726.48
## 89    Brunei Darussalam 2010 31453.22
## 90    Brunei Darussalam 2008 37798.39
## 341              Kuwait 2008 54484.30
## 561           Singapore 2008 39721.05
## 650 Trinidad and Tobago 2010 15840.44
```

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.

Solution

Build the linear model

```
m1 <- lm(infant_mortality ~ gdppc, data = d_wdi)
summary(m1)

##
## Call:
## lm(formula = infant_mortality ~ gdppc, data = d_wdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.697 -16.248  -6.166  11.606  80.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.633e+01  1.384e+00   26.25   <2e-16 ***
## gdppc       -7.307e-04  5.933e-05  -12.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.9 on 350 degrees of freedom
## Multiple R-squared:  0.3024,Adjusted R-squared:  0.3004
## F-statistic: 151.7 on 1 and 350 DF,  p-value: < 2.2e-16
```

Calculate the regression coef by hand, using covariance and variance:

```
cov(d_wdi$infant_mortality, d_wdi$gdppc) / var(d_wdi$gdppc)

## [1] -0.00073073
```

or fully by hand (the result will be slightly different because R's ‘cov‘ and ‘var‘ divided by ‘n - 1‘ while ‘mean‘ divides by ‘n‘) (read more at http://nebula.deanza.edu/~bloom/math10/m10divideby_nminus1.pdf):

```
mean((d_wdi$infant_mortality - mean(d_wdi$infant_mortality)) *
       d_wdi$gdppc - mean(d_wdi$gdppc)) /
  mean((d_wdi$gdppc - mean(d_wdi$gdppc))**2)

## [1] -0.0007699624
```

```
library(stargazer)

##
## Please cite as:
##  Hlavac, Marek (2015).  stargazer:  Well-Formatted Regression and
Summary Statistics Tables.
##  R package version 5.2.  http://CRAN.R-project.org/package=stargazer

stargazer(m1,
          covariate.labels = c("GDP per capita"),
          dep.var.labels = c("Infant Mortality (per 1000 births)"),
          model.names = TRUE)
```

Table 1:

	Dependent variable:
	Infant Mortality (per 1000 births)
	OLS
GDP per capita	-0.001^{***}
	(0.0001)
Constant	36.332^{***}
	(1.384)
Observations	352
R^2	0.302
Adjusted R^2	0.300
Residual Std. Error	20.905 (df = 350)
F Statistic	151.705^{***} (df = 1; 350)
Note:	*p<0.1; **p<0.05; ***p<0.01

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.

Solution

```
res <- m1$residuals # Residuals
pred <- m1$fitted.values # Predicted values
y <- m1$model$infant_mortality # Data of Y that is used by lm()

# Calculate 3 sum of squares
TSS <- sum( (y - mean(y)) ** 2)
RegSS <- sum( (pred - mean(y)) ** 2)
RSS <- sum( res ** 2 )

# Calculate root mean square error
N <- nrow(d_wdi)
k <- 1 # We only have 1 predictor, which is log_gdppc
rmse <- sqrt(RSS / (N - k - 1))
```

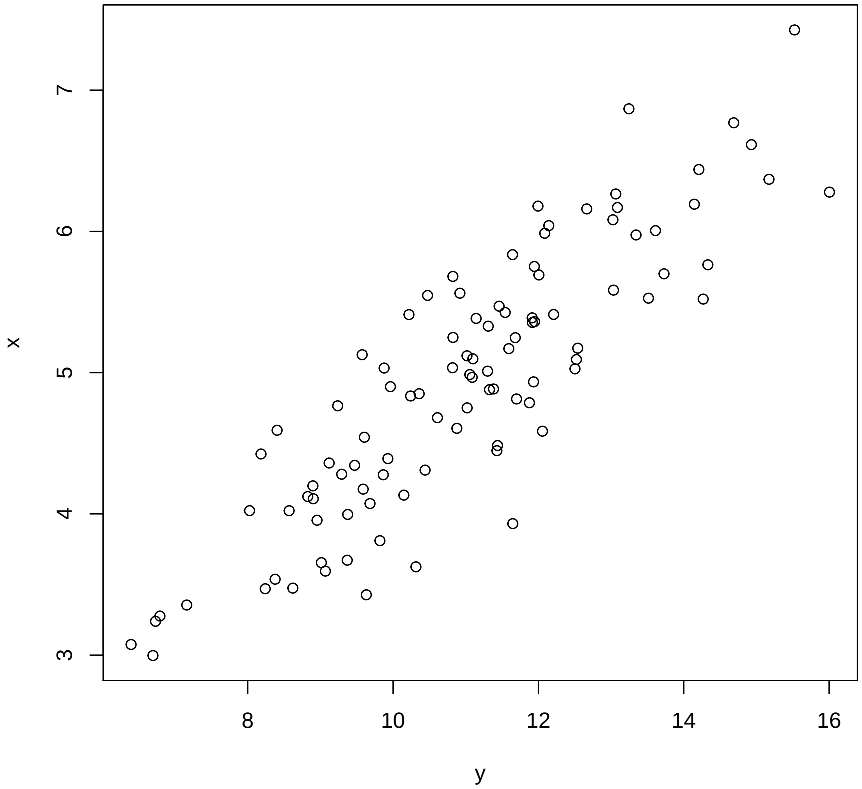
The calculated root mean square error is 20.9048032, the same as reported by R in `summary(m1)`.

4 Explore why we have standard errors

```
x <- rnorm(mean = 5, 100)
y <- 1 + 2 * x + rnorm(100)
plot(y, x)

c_nsim <- 100
model_data <- vector("list", length = c_nsim)
model_results <- vector("list", length = c_nsim)
for (i in 1:c_nsim) {
  sample_index <- sample(100, 10, replace = TRUE)
  sample_x <- x[sample_index]
  sample_y <- y[sample_index]
  model_data[[i]] <- list(sample_x, sample_y)
  model_results[[i]] <- lm(sample_y ~ sample_x)
}

plot(y, x)
```



```
for (i in 1:c_nsim) {

}
```