R computing for Business Data Analytics

Homework 4 (Due date: December 08, 2014)

Please e-mail your homework (.pdf) and the associated *R* code (.*R*) to hchuang.om@gmail.com.

The email title must be R_HW4_GroupName. NO late homework will be accepted.

Q1. Import the library *AER* in *R*, and attach the data set *CPS1988*.

> data("CPS1988")

> attach(CPS1988)

Focus on the four variables – wage, education, experience, and ethnicity (African-American versus Caucasian). Your job is to finish the following tasks.

(a) Run the linear regression model below (using *lm*( )) and save the model as object "CPS_lm". Note that the ethnicity is a categorical/dummy variable.

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{experience} + \beta_2 \text{experience}^2$$
$$+\beta_3 \text{education} + \beta_4 \text{ethnicity} + \varepsilon$$

(b) Explain the results in detail. What is the statistical significance of each independent variable? What are the implications of the findings? Particularly, what is the association between wage and experience? Is the identified association linear? If not, what is the shape of the association?

(c) Based on the estimated coefficients, write out the two equations of predictive models for Africa-American and Caucasian respectively.

Q2. Monte-Carlo simulation experiments of linear regression (OLS)

Run the following codes in *R*

```
> library(mvtnorm)
> set.seed(121402)
# Create two correlated independent variables
> x.corr=matrix(c(1, mclvl, mclvl, 1), ncol=2)
> x=rmvnorm(n, mean=c(0, 0), sigma=x.corr) # n is the sample size
> x1=x[ , 1]
> x2=x[,  2]
```

(a) *Multicollinearity*

The *mclvl* refers to the level of collinearity between x1 and x2. Set n=1000, and for each *mclvl* in seq(0, 0.95, 0.05) simulate the dependent variable using

> y=b0+b1*x1+b2*x2+rnorm(n, 0, 1) #set b0=0.2; b1=0.5; b2=0.75

Estimate b1 from lm(y~x1+x2). Repeat the estimation for 1000 times and calculate the standard deviation of those 1000 estimated b1 (for a given *mclvl*).

Do the whole simulation again with n=5000 and save the standard deviation of estimated b1.

For n=1000 and n=5000, generate a plot (respectively) where the x-axis is *mclvl* and the y-axis is the corresponding *standard deviation* of b1. Discuss what you find.

(b) *Omitted variable*

Following the procedure in (a), for each *mclvl* in c(0, 0.5, 1), set n=1000 and simulate x1& x2. Then simulate the dependent variable using

> y=b0+b1*x1+b2*x2+rnorm(n, 0, 1) #set b0=0.2; b1=0.5; b2=0.75

Estimate b1 from lm(y~x1) – we intentionally omit x2 – and repeat the estimation for 1000 times (for a given *mclvl*). Save all of the estimated b1 and plot the three distributions of estimated b1 in each *mclvl*. Compare the distributions to the true b1=0.5. What is the impact of omitting x2? Discuss what you observe.

(c) Me*asurement error*

Run the following codes in *R*

> set.seed(385062)
>
> n=1000
>
> x=runif(n, -1, 1)

For each *errlvl* in c(0, 0.5, 1), generate x with measurement error

> xp=x+rnorm(n, 0, errlvl)

Then repeat the following process for 1000 times. First simulate the dependent variable using

> y=b0+b1*x+rnorm(n, 0, 1) #set b0=0.2; b1=0.5

Then estimate b1 from OLS regression (*lm*( ))

> lm(y~xp)

Save the estimated b1 in each of the 1000 replications (for this given *errlvl*).

Plot the three distributions of estimated b1s for *errlvl* in c(0, 0.5, 1). Compare the distributions to the true b1=0.5. What's the impact of measurement errors? Discuss what you observe.