PROBLEM SET 1: ECONOMETRICS I

CEF EXERCISES

a) If $e = Xu$, where X and u are independent $N(0,1)$, prove that, conditional on X, the error has a distribution $N(0, X^2)$.

b) What does the above imply in terms of the independence between e and X? Relate this to the concept of homoscedasticity.

c) Prove that Q_{XX} is a positive semi-definite matrix.

d) Consider the linear projection:

$$\mathcal{P}[log(wage)|X] = 0.046experience - 0.07experience^2 + 0.11education + 2.3$$

Formally (using derivatives), what is the effect of one extra year of experience on log(wage)?

e) Consider the linear projection:

$$\begin{aligned}\mathcal{P}[log(wage)|X] = 0.046experience - 0.07experience^2 - 0.09female \\ +0.11education - 0.07education * female + 1.06\end{aligned}$$

Formally (using derivatives), what is the return to one extra year of education on log(wage) for males? and for females?

f) Draw a two-dimensional graph showing the linear projections onto education by gender.

g) Consider the linear projection:

$$\begin{aligned}\mathcal{P}[log(wage)|X] = 0.046experience - 0.07experience^2 - 0.09female + 0.05education - 0.04education * female \\ +0.05north - 0.03north * female + 0.08north * education - 0.05north * education * female + 0.98\end{aligned}$$

where *north* identifies with 1 the population living in northern Mexico and zero otherwise.

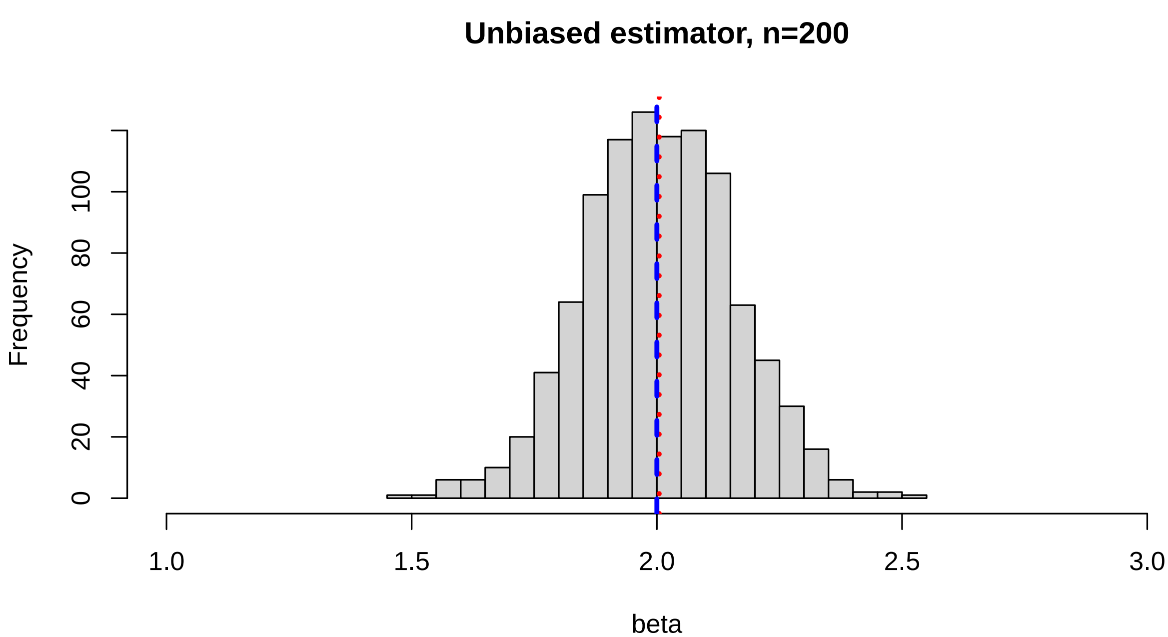
Formally (using derivatives), what is the return to one extra year of education on log(wage) for males in the north of Mexico? and for females in the south?

h) This is a piece of code simulating an unbiased estimator β_1 and β_2.

```
# Parameters ====
repet <- 1000
n <- 200
beta <- NULL

# Simulation ====
for (i in 1:repet){
  x <- rnorm(n)
  x2 <- rnorm(n)
  u <- rnorm(n,0,1)
  y=2+2*x+2*x2+u
  beta[i] <- lm(y~x)$coef[2]
}

# Plot ====
hist(beta,
     main="Unbiased estimator, n=200",
     xlim = c(1,3),
     breaks = 25)
abline(v = mean(beta), col="red", lwd=3, lty=3)
abline(v = 2, col="blue", lwd=3, lty=2)
```



- Modify the code so that it shows a **biased** estimator of β_1.
- While keeping this last modification make another (different) modification to the code so that you can get rid of the bias.
- Give a concrete example of a linear projection on which we most likely suffer from OVB.

LS EXERCISES

a) Use the salary dataset provided by Wooldridge:

```
library(wooldridge)
data("wage1")
```

- Construct the vector X with education, experience, and experience squared, and let the vector Y be the logarithm of salary. Obtain the following:

$$\hat{\beta} = \hat{Q}_{XX}^{-1} \hat{Q}_{XY}$$

- Explicitly display the vector $\sum_{i=1}^{n} X_i Y_i$ and the matrix $(\sum_{i=1}^{n} X_i X_i')^{-1}$.
- Interpret β_1 and the effect of experience on salaries.
- Are the *residuals* orthogonal to education? Are the errors orthogonal to education? Explain.
- Formally demonstrate whether there is bias in the estimator $\hat{\beta}_1$.
- Obtain ESS and TSS, as well as R^2, and interpret it.

b) Generate two parameters, x and $residual$ (n=100, mean=0, and sd=5) where:

```
y <- 15 + (7 * x) + residual
```

- Perform a regression of x on y.
- Create a scatter plot of the estimated regression showing each observation Y_i and the LS line.
- Obtain ESS, TSS, and R^2.
- Demonstrate, by modifying the data created in b), that if RSS increases, R^2 decreases. Discuss whether $\hat{\beta}$ is unbiased.
- Create a variable x_2 and add it to the estimated regression in b). What happens to ESS, TSS, and R^2?
- Modify your code to significantly bias $\hat{\beta}_1$ and re-estimate the same regression. Does this change R^2? Discuss.

e) Consider two LS regressions:

$$Y = X_1 \tilde{\beta} + \tilde{e}$$

$$Y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{e}$$

where R_1^2 and R_2^2 correspond to each regression, respectively. Show that $R_2^2 \geq R_1^2$. In what case would they be equal?