San Francisco State University **Michael Bar**
ECON 312 **Fall 2018**

Problem set 2

Due Thursday, September 20, in class

Names of collaborators (type below):

Name 1 MINJIN PARK

Name 2 NISHANLANG KHONGLAH

Assignment Rules

1. Homework assignments must be typed. For instruction on how to type equations and math objects please see notes "Typing Math in MS Word".
2. Homework assignments must be prepared within this template. Save this file on your computer and type your answers following each question. Do not delete the questions.
3. Your assignments must be stapled.
4. No attachments are allowed. This means that all your work must be done within this word document and attaching graphs, questions or other material is prohibited.
5. Homework assignments must be submitted at the end of the lecture, in class, on the listed dates.
6. Late homework assignments will not be accepted under any circumstances, but the lowest homework score will be dropped.
7. The first homework assignment cannot be dropped.
8. You are encouraged to work on this homework assignment in groups of up to 3 people, and submit one assignment with up to 3 names typed on this page. Sharing the electronic version of your assignment with other teams is absolutely prohibited.
9. All the graphs should be fully labeled, i.e. with a title, labeled axis and labeled curves.
10. In all the questions that involve calculations, you are required to show all your work. That is, you need to write and explain the steps that you made in order to get to the solution.
11. This page must be part of the submitted homework.

Properties of Estimators

1. (6 points). Define the following concepts:
   a. **Random Sample**. A random sample of size n on a random variable X is a collection of independent random variables $X_1,\ldots., X_n$ each with the same distribution of X.

   b. **Estimator**. An estimator is a function of a random sample $X_1,\ldots., X_n$.

   c. **Estimate**. An estimate is a particular realization of an estimator.

2. (10 points). Suppose $\mu_n = \sum_{i=1}^{n} \lambda_i X_i$ is an estimator of the population mean μ, with $\lambda_i s$ being some numbers.
   a. Prove that this estimator is unbiased if and only if $\sum_{i=1}^{n} \lambda_i = 1$.

   $Let\ \hat{\theta} = \mu_n$

   For the estimator $\hat{\theta}$ to be unbiased, the condition $E(\hat{\theta}) = \mu$. must be satisfied

   $$E(\hat{\theta}) = E\left(\sum_{i=1}^{n} \lambda_i X_i\right)$$

   $$= \sum_{i=1}^{n} E(\lambda_i X_i) \quad \text{Expectation of a sum} = \text{The sum of expectations}$$

   $$= \sum_{i=1}^{n} \lambda_i\, E(X_i) \quad \text{Constants factor out}$$

   $$= \sum_{i=1}^{n} \lambda_i\, \mu \qquad \text{Expectation of } X_i = \text{ the population mean } \mu$$

   Therefore, we can say that $\hat{\theta}$ is unbiased if and only if $\sum_{i=1}^{n} \lambda_i = 1$

   b. Prove that if $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \frac{1}{n}$, then the estimator is efficient among all unbiased estimators of the form $\mu_n = \sum_{i=1}^{n} \lambda_i X_i$.

   Plugging for $\lambda_n = \frac{1}{n}\ in\ \mu_n = \sum_{i=1}^{n} \lambda_i X_i$

   $$Therefore, \mu_n = \sum_{i=1}^{n} \frac{1}{n} X_i$$

   $$E(\mu_n) = E\left(\sum_{i=1}^{n} \frac{1}{n} X_i\right)$$

   $$= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) \qquad \text{Constants factor out}$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \qquad \text{Expectation of a sum} = \text{Sum of expectations}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu \qquad E(X_i) = \text{the population mean } \mu$$

$$= \frac{1}{n} \cdot n\mu = \mu$$

Therefore, μ_n is unbiased when $\lambda_n = \frac{1}{n}$

We now determine the variance of the estimator μ_n *when* $\lambda_n = \frac{1}{n}$

$$Var(\mu_n) = Var\left(\sum_{i=1}^{n} \frac{1}{n} X_i\right) \qquad \text{for } \lambda_n = \frac{1}{n}$$

$$= \frac{1}{n^2}\, Var\left(\sum_{i=1}^{n} X_i\right) \qquad \text{Constants factor out as squares}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \qquad \text{Independent random variables of same distibution}$$

$$= \frac{1}{n^2} \cdot n\,\sigma^2 = \frac{\sigma^2}{n}$$

We now determine the variance of the estimator $\hat{\theta} = \mu_n$, which has been defined in section A.

$$Var(\hat{\theta}) = Var\left(\sum_{i=1}^{n} \lambda_i X_i\right)$$

$$= \sum_{i=1}^{n} Var(\lambda_i X_i)$$

$$= \lambda_i^2 \sum_{i=1}^{n} Var(X_i) \qquad \text{Constants factor out as squares}$$

$$= \lambda_i^2 \sum_{i=1}^{n} \sigma^2 \qquad \text{Independent random variables of same distibution}$$

$$= \lambda_i^2 \cdot n\sigma^2$$

Comparing $\text{Var}(\mu_n)$ and $\text{Var}(\hat{\theta})$,

$\text{Var}(\mu_n) = \frac{\sigma^2}{n}$ and $\text{Var}(\hat{\theta}) = \lambda_i^2 \cdot n\sigma^2$

We observe that $\text{Var}(\mu_n) < \text{Var}(\hat{\theta})$

Therefore, when $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \frac{1}{n}$

μ_n is efficient among all unbiased estimators of the form $\mu_n = \sum_{i=1}^{n} \lambda_i X_i$

3. (10 points). Suppose A and B are two estimators, with the following properties:

	A	B
Bias	-2	1
Variance	5	8

   a. Based on our definition of efficient estimator, can we determine which of these two estimators is more efficient?

      We cannot determine which of the two estimators is more efficient because both estimators are biased and the definition of an efficient estimator holds for unbiased estimators.

   b. Based on the MSE criterion, which estimator would you choose?

      According to the MSE criterion, MSE = variance + bias^2

      $$\text{MSE (A)} = 5 + (-2^2) = 5 + 4 = 9$$
      $$\text{MSE (B)} = 8 + 1^2 = 9$$

      In this case, either estimator can be considered as they have the same MSE.

4. (12 points). The following are estimators of the population mean μ, based on a random sample of n observations. For each estimator, determine whether it is biased or not, and if it is biased, find the bias.
   a. $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$E(A_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$
$$= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) \quad \text{The constant } \frac{1}{n} \text{ factors out}$$
$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \quad \text{Expectation of a sum} = \text{sum of expectations}$$
$$= \frac{1}{n}\sum_{i=1}^{n} \mu \quad E(X_1) = \text{the population mean } \mu$$
$$= \frac{1}{n}\, n \cdot \mu = \mu$$

Therefore, A_n is unbaised

b. $B_n = \frac{1}{n-1}\sum_{i=1}^{n} X_i$

$$E(B_n) = E\left(\frac{1}{n-1}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n-1}\, E\left(\sum_{i=1}^{n} X_i\right) \quad \text{The constant } \frac{1}{n-1} \text{ factors out}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} E(X_i) \quad \text{Expectation of a sum} = \text{sum of expectations}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \mu \quad E(X_1) = \text{the population mean } \mu$$

$$= \frac{n\,\mu}{n-1}$$

Therefore, B_n is biased.

$$\text{Bias } (B_n) = E(B_n) - \mu$$

$$= \frac{n\,\mu}{n-1} - \mu$$

$$= \frac{n\,\mu - n\,\mu + \mu}{n-1}$$

$$= \frac{\mu}{n-1}$$

c. $C_n = \frac{1}{n+1}\sum_{i=1}^{n} X_i$

$$E(C_n) = E\left(\frac{1}{n+1}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n+1} E\left(\sum_{i=1}^{n} X_i\right) \quad \text{The constant } \frac{1}{n+1} \text{ factors out}$$

$$= \frac{1}{n+1}\sum_{i=1}^{n} E(X_i) \quad \text{Expectation of a sum} = \text{sum of expectations}$$

$$= \frac{1}{n+1}\sum_{i=1}^{n} \mu \quad E(X_i) = \text{the population mean } \mu$$

$$= \frac{n\mu}{n+1}$$

Therefore C_n is biased

$$\text{Bias } (C_n) = E\,(C_n) - \mu$$

$$= \frac{n\mu}{n+1} - \mu$$

$$= -\frac{\mu}{n+1}$$

d. $D_n = \frac{1}{2}X_1 + \frac{1}{n-1}\sum_{i=1}^{n-1} X_i$

$$E(D_n) = E\left(\frac{1}{2}X_1 + \frac{1}{n-1}\sum_{i=1}^{n-1} X_i\right)$$

$$= E\left(\frac{1}{2}X_1\right) + E\left(\frac{1}{n-1}\sum_{i=1}^{n-1} X_i\right) \quad \text{Expectation of a sum}$$

$$= \text{sum of expectations}$$

$$= \frac{1}{2}E(X_1) + \frac{1}{n-1}\sum_{i=1}^{n-1} E(X_i) \quad \text{Constants factor out}$$

$$= \frac{1}{2}\mu + \frac{1}{n-1}\sum_{i=1}^{n-1} \mu \qquad \mathrm{E(X_i)} = \text{the population mean } \mu$$

$$= \frac{1}{2}\mu + \frac{1}{n-1}\cdot(n-1)\mu$$

$$= \frac{1}{2}\,\mu + \mu = \frac{3}{2}\,\mu$$

Therefore, $\mathrm{D_n}$ is biased

$$\text{Bias } (\mathrm{D_n}) = \mathrm{E(D_n)} - \mu$$

$$= \frac{3}{2}\mu - \mu = \frac{1}{2}\mu$$

5. (12 points). For each estimator in the previous question, determine whether it is consistent or not.

For an estimator to be consistent, its variance and bias must be 0 as the sample size n approaches ∞. Therefore, we determine the variance for each estimator and using the bias calculated in the previous question, we verify if the variance and bias of each estimator are 0 as n approaches ∞.

$$Var\ (A_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\ Var\left(\sum_{i=1}^{n} X_i\right) \quad \text{Constants factor out as squares}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \qquad \mathrm{Var(X_i)} = \sigma^2$$

$$= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

$$\lim_{n\to\infty} Var(A_n) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

$Bias(A_n) = 0$ Shown in question 4 a.

Therefore, A_n is consistent.

$$Var(B_n) = Var\left(\frac{1}{n-1}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{(n-1)^2} Var\left(\sum_{i=1}^{n} X_i\right)$$ Constants factor out as squares

$$= \frac{1}{(n-1)^2} \sum_{i=1}^{n} Var(X_i)$$

$$= \frac{1}{(n-1)^2} \sum_{i=1}^{n} \sigma^2$$ $\mathrm{Var}(X_i) = \sigma^2$

$$= \frac{n\,\sigma^2}{(n-1)^2}$$

$$\lim_{n\to\infty} \mathrm{Var}(B_n) = \lim_{n\to\infty} \frac{n\,\sigma^2}{(n-1)^2} = 0$$

$$\lim_{n\to\infty} \mathrm{Bias}\,(B_n) = \lim_{n\to\infty} \frac{\mu}{n-1} = 0$$

Therefore, B_n is consistent.

$$Var(C_n) = Var\left(\frac{1}{n+1}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{(n+1)^2} Var\left(\sum_{i=1}^{n} X_i\right)$$ Constants factor out as squares

$$= \frac{1}{(n+1)^2} \sum_{i=1}^{n} Var(X_i)$$

$$= \frac{1}{(n+1)^2} \sum_{i=1}^{n} \sigma^2$$ $\mathrm{Var}(X_i) = \sigma^2$

$$= \frac{n\sigma^2}{(n+1)^2}$$

$$\lim_{n\to\infty} \mathrm{Var}(C_n) = \lim_{n\to\infty} \frac{n\sigma^2}{(n+1)^2} = 0$$

$$\lim_{n\to\infty} \mathrm{Bias}\,(C_n) = \lim_{n\to\infty} -\frac{\mu}{n+1} = 0$$

Therefore, C_n is consistent.

$$Var(D_n) = Var\left(\frac{1}{2}X_1 + \frac{1}{n-1}\sum_{i=1}^{n-1} X_i\right)$$

$$= Var\left(\frac{1}{2}X_1\right)$$

$$+ Var\left(\frac{1}{n-1}\sum_{i=1}^{n-1} X_i\right) \text{ Covariance is 0 for independent random variables}$$

$$= \frac{1}{4}Var(X_1) + \frac{1}{(n-1)^2}Var\sum_{i=1}^{n-1} X_i \text{ Constants factor out as squares}$$

$$= \frac{1}{4}Var(X_1) + \frac{1}{(n-1)^2}\sum_{i=1}^{n-1} Var(X_i)$$

$$= \frac{1}{4}\sigma^2 + \frac{1}{(n-1)^2}\sum_{i=1}^{n-1} \sigma^2$$

$$= \frac{1}{4}\sigma^2 + \frac{(n-1)\sigma^2}{(n-1)^2}$$

$$= \frac{1}{4}\sigma^2 + \frac{\sigma^2}{n-1}$$

$$\lim_{n\to\infty} \text{Var}(\text{D}_\text{n}) = \lim_{n\to\infty}\left(\frac{1}{4}\sigma^2 + \frac{\sigma^2}{n-1}\right) \neq 0$$

$$\lim_{n\to\infty} \text{Bias}(\text{D}_\text{n}) = \lim_{n\to\infty}\frac{1}{2}\mu \neq 0$$

Therefore, D_n is inconsistent.

Simple Regression Model

6. (10 points). Consider the simple regression model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Derive the OLS estimators of the unknown population parameters β_1, β_2. Denote these estimators by b_1, b_2.

The fitted equation for the estimators b_1 and b_2 is defined as follows

$\widehat{Y_i} = b_1 + b_2 X_i$ where $\widehat{Y_i}$ is the fitted value of Y_i

The residual (prediction error) associated with each observation is

$e_i = Y_i - \widehat{Y_i} = Y_i - b_1 - b_2 X_i$

According to the OLS method, the estimators b_1 and b_2 should be chosen such that the residual sum of squares (RSS) is minimized.

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\mathrm{Y_i} - b_1 - b_2 X_i)^2$$

$$\text{Formally, } b_1^{OLS}, b_2^{OLS} = \arg\min_{b_1 b_2} \sum_{i=1}^{n} (\mathrm{Y_i} - b_1 - b_2 X_i)^2$$

We derive the OLS estimators using Calculus as follows

$$|b_1|: \quad \frac{\partial RSS}{\partial b_1} = -\sum_{i=1}^{n} 2(\mathrm{Y_i} - b_1 - b_2 X_i) = 0 \qquad (1)$$

$$|b_2|: \quad \frac{\partial RSS}{\partial b_2} = -\sum_{i=1}^{n} 2X_i(\mathrm{Y_i} - b_1 - b_2 X_i) = 0 \quad (2)$$

Rearranging equation (1) after eliminating -2

$$\sum_{i=1}^{n} \mathrm{Y_i} - \sum_{i=1}^{n} b_1 - b_2 \sum_{i=1}^{n} X_i = 0$$

$$\sum_{i=1}^{n} \mathrm{Y_i} - nb_1 - b_2 \sum_{i=1}^{n} X_i = 0$$

$$nb_1 = \sum_{i=1}^{n} \mathrm{Y_i} - b_2 \sum_{i=1}^{n} X_i$$

$$b_1 = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Y_i} - \frac{1}{n} b_2 \sum_{i=1}^{n} X_i$$

$$b_1 = \bar{Y} - b_2 \bar{X}$$

Rearranging equation (2) after eliminating -2

$$\sum_{i=1}^{n} Y_i X_i - b_1 \sum_{i=1}^{n} X_i - b_2 \sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} Y_i X_i - b_1 n\bar{X} - b_2 \sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} Y_i X_i - n\bar{X}\bar{Y} + b_2 n\bar{X}^2 - b_2 \sum_{i=1}^{n} X_i^2 = 0 \qquad \text{After plugging in } \mathrm{b_1}$$

$$\sum_{i=1}^{n} Y_i X_i - n\bar{X}\bar{Y} = b_2 \sum_{i=1}^{n} X_i^2 - b_2 n\bar{X}^2$$

$$b_2 = \frac{\sum_{i=1}^{n} Y_i X_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

Create an R script, which performs all the analysis for question 7. You can name the script `HW2.R`. You can either print out the script and attach it as a separate page at the end of this assignment, or copy and paste its content at the end of your assignment. Make sure to add comments explaining which question you are solving.

7. (40 points). This question uses the wage21 data set, posted on the course webpage in several formats. One way to read the data set is to use the following command in R:

```
wage <- read.csv("http://online.sfsu.edu/mbar/ECON312_files/wage21.csv")
```

The above command reads the data set in `.csv`, and stores it as data frame named "`wage`". For loading data in other formats into R, see the script `LabIntroR.R`.

The key variables that you need to know for this assignment are:
`EARNINGS` – hourly earnings, in $ per hour.
`S` – schooling, in years.

a. Consider the model:

$$EARNINGS_i = \beta_1 + \beta_2 \cdot S_i + u_i$$

In this model, the dependent variable is EARNINGS and the regressor is SCHOOLING.

b. What is the interpretation of the error term u_i in this model?

The error term u_i in this model represents all influences on the dependent variable, $EARNINGS_i$, other than the regressor, S_i. In other words, the error term in this model represents all factors that affect earnings other than schooling. For example, skill set, experience, occupation, etc.

c. Estimate the model's coefficients using OLS, and report the estimates b_1, b_2. The commands in R are:

```
model1 <- lm(EARNINGS ~ S, data = wage) #OLS estimation
b <- coef(model1) #Storing the OLS coefficients in vector b
b  #Displaying the estimated coefficients
```

Based on the OLS estimation performed using the command above, the estimates for b_1 = -15.59 and for b_2 = 2.57

d. Interpret the regression coefficients.

b_2 = 2.57 implies that $2.57 is added to hourly earnings for every year of schooling.

b_1 = -15.59 implies that the predicted hourly earnings for someone with 0 years of schooling is $-15.59, which is not plausible.

e. Based on your estimates, what is your predicted hourly earnings for an individual with 16 years of schooling? You can calculate this in R as follows:

```
EARN_pred <- b[1] + b[2]*16 #Calculating the prediction
EARN_pred #Display the prediction
```

Using the command above in R, the predicted hourly earnings for an individual with 16 years of schooling is $25.66/hour.

f. Based on your estimates, what is your predicted hourly earnings for an individual with 20 years of schooling? You can calculate this in R as follows:

```
EARN_pred <- b[1] + b[2]*20 #Calculating the prediction
EARN_pred #Display the prediction
```

Using the command above in R, the predicted hourly earnings for an individual with 20 years of schooling is $35.97/hour.

g. Calculate and interpret the value of R^2 in this model. Use the command `summary(model1)` in R. This gives a table with various statistics, including the `Multiple R-squared`.

Using the summary command, the value of R^2 was calculated to be 0.20, which implies that only 20% of the variations in earnings in the data can be explained by schooling.

h. Plot the graph of actual values of `EARNINGS` vs `S`, and add the fitted equation $\widehat{EARNINGS_i} = b_1 + b_2 \cdot S_i$. In R, use the following commands:

```
plot(EARNINGS ~ S, data=wage, main="Earnings vs Schooling",
     col="blue",lwd=2)
abline(coef(model1),col="red",lwd=2)
legend("topright", legend = c("data","fitted equation"),
       col=c("blue","red"), lty = c(0,1), pch = c(1,NA),lwd=2)
grid() #Optional, add grid to figure
```