Problem Set 1

Philip Nye

2/8/2022

Contents

Crime and Unemployment - Practical

Problem 1

1. Firstly let's look at our data. This is the most important part of econometrics, and it is often forgotten. Let's draw a histogram of the Violence data. Left click on the 'Violence' data to select it, then Variable → Frequency distribution, at the top of the Gretl GUI. Select the number of bins equal to 19, and select the then click 'ok'. A nice histogram should pop up as a figure.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

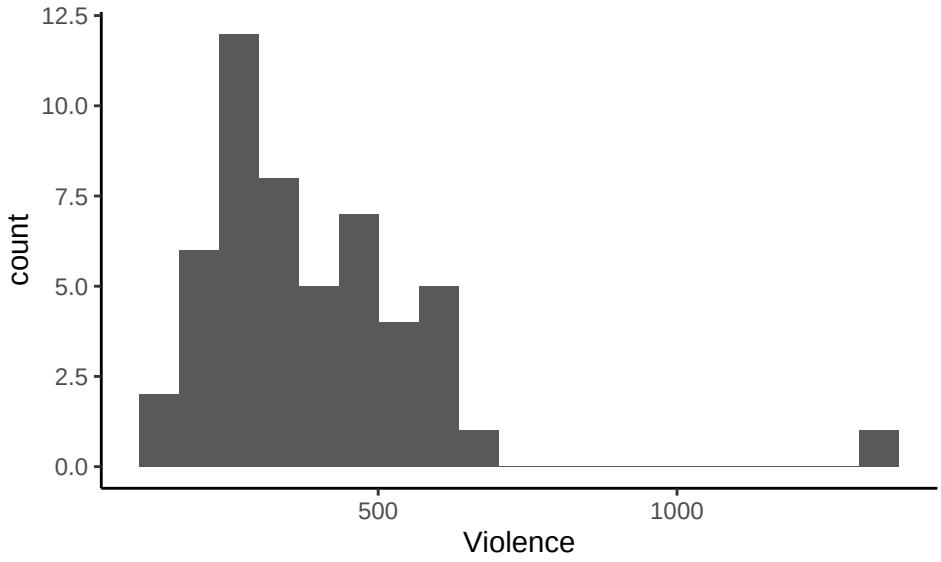
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

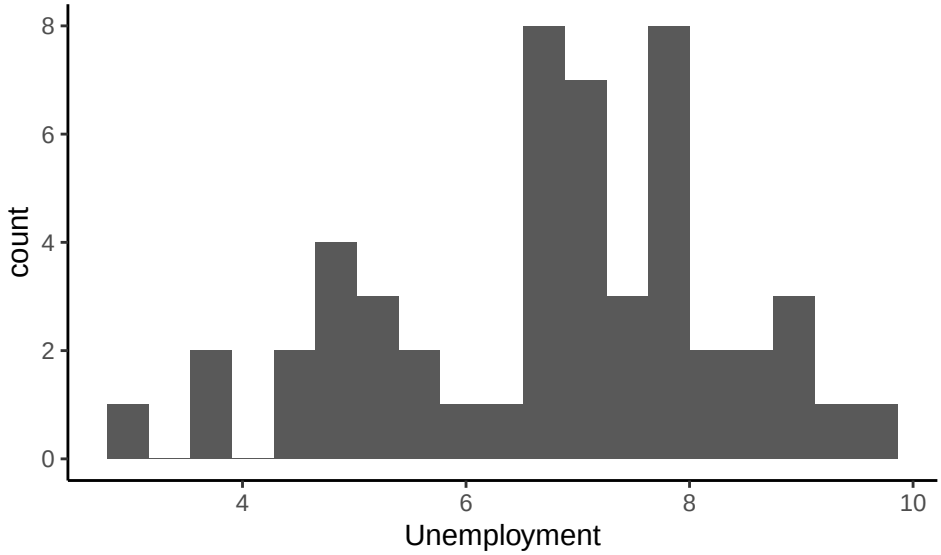
```
violence_data <- readxl::read_xls("problemset1.xls")

ggplot(violence_data, aes(Violence)) +
  geom_histogram(bins = 19) +
  theme_classic()
```



Problem 2

2. Draw a similar histogram (with 19 bins) of the Unemployment data, and report the unemployment rate bin which has the highest frequency.

```
ggplot(violence_data, aes(Unemployment)) +
  geom_histogram(bins=19) +
  theme_classic()
```



Approximating the intervals for the bins, the bins that share the highest frequency are the bins on the interval (6.5,6.9) and (7.75,8).

Problem 3

3. Can you find out which State has the highest rate of violent crime reported? To do this you just need to click on the 'Violence' variable, then look for the state with the highest violence rate.

The State with the highest rate of violent crime reported is District of Columbia (if D.C. counts). If D.C. doesn't count, then Nevada is the answer.

Problem 4

4. Another way of understanding a dataset it to look at its summary statistics. Gretl provides a nice, and simple way of doing this. In order to view this information for a given variable, just click 'Variable' → 'Summary Statistics'. This will provide a statistical summary of a given variable. Why not have a look at the Unemployment dataset's summary statistics?

```
skewness <- function(x, na.rm = FALSE){
  if(any(ina <- is.na(x))) {
    if(na.rm)
      x <- x[!ina]
    else
      return(NA)
  }
  skew <- sqrt(length(x)) * sum((x - mean(x)) ^ 3) / (sum((x - mean(x)) ^ 2) ^ (3/2))
  return(skew)
}

kurtosis <- function(x, na.rm = FALSE){
  if(any(ina <- is.na(x))) {
    if(na.rm)
      x <- x[!ina]
    else
      return(NA)
  }

  kurt <- length(x) * sum((x - mean(x)) ^ 4) / (sum((x - mean(x)) ^ 2) ^ 2)
  return(kurt)
}


summary_table <- function(data, round_to = 3){
  data.numeric <- data[,sapply(data, is.numeric)]
  cat("\\begin{table}[h!] \n")
  cat("\\resizebox{\\textwidth}{!} \n")
  cat("{ \n")
  cat("\\begin{tabular}{|l")
  for(i in seq(ncol(data.numeric)+1)){
    cat("|c")
  }
  cat("|} \n")
  cat("\\hline \n")
  stats <- matrix(nrow = 9,ncol = ncol(data.numeric) + 1)
  stats[,1] <- c(" ", "Mean", "Median", "Maximum", "Minimum", "Variance",
                 "Std. Dev.", "Skewness", "Kurtosis")
  for(i in seq.int(from=2,to=ncol(stats))){
    stats[,i] <- c(colnames(data.numeric)[i-1],
                   round(mean(data.numeric[[i-1]]),round_to),
                   round(median(data.numeric[[i-1]]),round_to),
                   round(max(data.numeric[[i-1]]),round_to),
                   round(min(data.numeric[[i-1]]),round_to),
                   round(var(data.numeric[[i-1]]),round_to),
                   round(sd(data.numeric[[i-1]]),round_to),
                   round(skewness(data.numeric[[i-1]]),round_to),
                   round(kurtosis(data.numeric[[i-1]]),round_to))
```

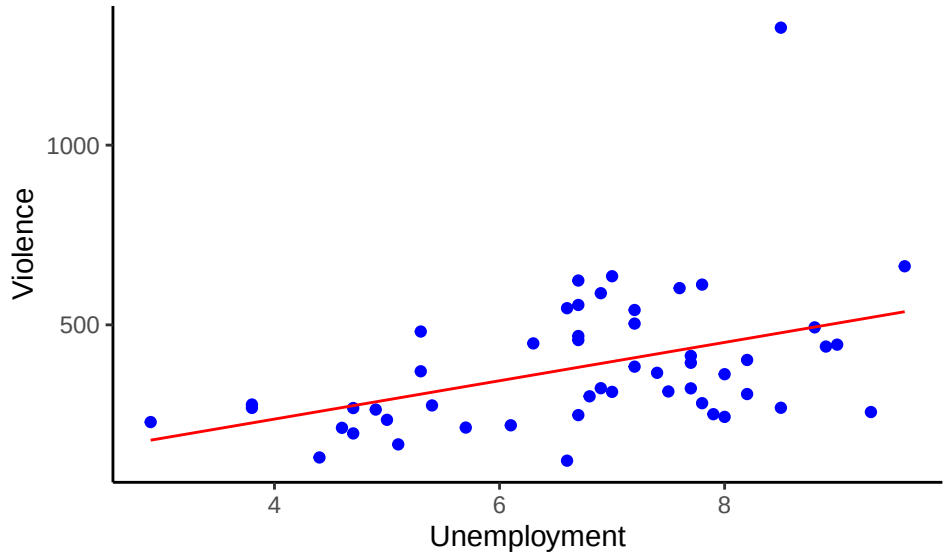
```
  }
  for(i in seq(nrow(stats))){
    cat("\\hline \n")
    for(ind in seq(ncol(stats)-1)){
      cat(paste(stats[i,ind], " & "))
    }
    cat(paste(stats[i,ncol(stats)], "\\\\ \n"))
  }
  cat("\\hline \n")
  cat("\\hline \n")
  cat("\\end{tabular} \n")
  cat("} \n")
  cat("\\end{table}")
}

summary_table(violence_data)
```

	Unemployment	Violence
Mean	6.765	385.28
Median	6.9	323.7
Maximum	9.6	1326.8
Minimum	2.9	122.1
Variance	2.37	38069.372
Std. Dev.	1.539	195.114
Skewness	-0.452	2.279
Kurtosis	2.61	11.637

Problem 5

5. Let's look at whether it we can visibly see if there is any relationship between unemployment and violent crime rates by drawing a scatterplot. To do this go to View → Graph specified vars → X-Y scatter. Then select 'Violence' as a Yaxis variable variable, and 'Unemployment' as an X-axis variable. This should produce a scatterplot with a fitted regression line. From this, there appears to be some sort of positive relationship between 'Violence' and 'Unemployment'.

```
slr1 <- lm(Violence ~ Unemployment, data = violence_data)
violence_data <- violence_data %>%
  mutate(
    pred_violence = unname(slr1[["coefficients"]][1]) + unname(slr1[["coefficients"]][2]) * Unemploymen
  )

ggplot(violence_data, aes(Unemployment, Violence))+
  geom_point(color="blue") +
  geom_line(aes(Unemployment, pred_violence), color="red") +
  theme_classic()
```



Problem 6

6. One way of quantifying the relationship between two variables is via their correlation coefficient. You can find this on Gretl by going to 'View' → 'Correlation matrix. If you then select both 'Violence' and 'Unemployment' you should see an outputted correlation (along with associated p values etc.) of around 0.42. What does this mean?

The correlation between 'Violence' and 'Unemployment' is 0.4208818. This value means there's a positive correlation between 'Violence' and 'Unemployment.' More specifically, when the value for 'Violence' exceeds its average, 'Unemployment' also tends to exceed its respective average.

Problem 7

7. Since we've inspected our variables sufficiently, it is now time to run our first regression. Let's run an ordinary least squares regression with 'Violence' as a dependent variable, and 'Unemployment' (and a constant) as an independent variable. To do this go to 'Model' → 'Ordinary Least Squares'. Then select 'Violence' as a dependent variable, and 'Unemployment' as an independent (a constant should

already be in the list of independent variables) and click 'ok'. This should provide a read-out of the results from your first OLS regression!

```
library(stargazer)

stargazer(slr1,
          title = "How Unemployment May Influence Violence",
          style = "all",
          summary = T,
          dep.var.labels.include=T,
          df=T,
          digits = 3,
          float=F,
          header=F,
          model.names=T)
```

	Dependent variable:
	Violence
	OLS
Unemployment	53.348***
	(16.426)
	t = 3.248
	p = 0.003
Constant	24.398
	(113.900)
	t = 0.214
	p = 0.832
Observations	51
R^2	0.177
Adjusted R^2	0.160
Residual Std. Error	178.788 (df = 49)
F Statistic	10.549*** (df = 1; 49) (p = 0.003)
Note:	*p<0.1; **p<0.05; ***p<0.01

Problem 8

8. What is the coefficient on 'Unemployment'? What is the interpretation of this value?

The coefficient of 'Unemployment' is 53.3478477. This means for every percentage (unit) increase in 'Unemployment,' one can expect a 53.3478477 unit increase in 'Violence.'

Problem 9

9. What would this model predict would be the increase in the rate of violent crime for a 1 standard deviation increase in unemployment? What is this increase in terms of standard deviations of the rate of violence?

This model would predict an increase of 82.1198135 cases for a 1 standard deviation (1.5393276 units) increase in unemployment. In terms of standard deviations of the rate of violence, this is a 0.4208818 standard deviation increase in the rate of violence.

Problem 10

10. What does a regression of the rate of unemployment on violent crime rates (the other way round to that in the last part) suggest would be the increase in the unemployment rate for a 1 standard deviation increase in the rate of violent crime?

```
slr2 <- lm(Unemployment ~ Violence, data = violence_data)
```

A regression of the rate of unemployment on violent crime rates suggests the increase in the unemployment rate for a 1 standard deviation increase in the rate of violent crime would be a 0.6478749% increase in the violent crime rate.

Problem 11

11. Can you use this regression to uncover what it suggests the increase in unemployment associated with a 1 standard deviation increase in violent crime? Is this the same as we found previously? Why is this the same/different?

Note that this is completely different to the results which we obtained from our OLS regression of 'Violence' on 'Unemployment'. This is because of the fact that the regression of y on x is not the same as the regression of x on y. The former minimizes square distances of 'y' from the line, whereas the latter minimizes square distances in 'x'. Rearranging the latter regression equation will hence not yield the former. For example this model suggests that a 1.58% increase in the unemployment rate will increase violence rates by 467.6! Very different to the previous estimate.

Problem 12

12. What can you conclude about the causal mechanism between violent crime and unemployment based on the two regressions you have run? Does violent crime cause unemployment or vice versa?

No, the regression certainly doesn't entail any causality. It's likely there are several other unmeasurable factors that are influencing the violence rate and are also correlated with unemployment.

Problem 13

13. Why might it be incorrect to conclude that increases in unemployment lead to increases in rates of violent crime?

This notion implies causality when the data provided doesn't allow us to draw causal inferences. An experiment where all other factors are held constant while allowing unemployment and violence to vary would allow us to draw such a conclusion.

Theory

Problem 1

1. For a pupil, i, selected at random from a school, the number of years of education of their parents, X_i, is given by:

$$X_i = \mu + \varepsilon_i$$

$\varepsilon_i \sim iid(0, \sigma^2)$. Here μ is the mean number of years of education completed by parents. For a sample of N students selected independently from the population:

(a) What is the expected value of the sample mean?

(b) Calculate the variance of the sample mean. What happens to the variance as the sample size increases?

(c) Is the sample mean consistent?

(d) Prove that the sample mean is a least-squares estimator for the population mean.

(e) Is the sample mean BLUE? Either way, prove it.

1. (a) $E[\bar{X}] = E[\frac{1}{N}\sum_{i=1}^{N} X_i]$

$$= \frac{1}{N} E[\sum_{i=1}^{N} X_i]$$

$$= \frac{1}{N} \sum_{i=1}^{N} E[X_i]$$

$$= \frac{1}{N} \sum_{i=1}^{N} E[\mu + \epsilon_i]$$

$$= \frac{1}{N} \sum_{i=1}^{N} (E[\mu] + E[\epsilon_i])$$

$$= \frac{1}{N} \sum_{i=1}^{N} (E[\mu] + 0)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \mu$$

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

(b) $Var(\bar{X}) = Var(\frac{1}{N}\sum_{i=1}^{N} X_i)$

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

$$= \frac{1}{N^2} \cdot N \cdot Var(X) = \frac{\sigma_X^2}{N}$$

Note: The sum of variances only works here (without the covariance terms) because of the independence of the X's.

(c) Since $\bar{X}$ was proven unbiased as an estimator of the population mean, to prove consistency for $\bar{X}$, it suffices to show the variance of $\bar{X}$ tends to zero as $N \to \infty$. In this case, $\lim_{N\to\infty} Var(\bar{X}) = \lim_{N\to\infty} \frac{\sigma_X^2}{N} = 0$.

(d) For the population mean, there exists an estimator, which we will denote as $\hat{\mu}$, that solves $\min S = \sum_{i=1}^{N}(X_i - \hat{\mu})^2$.

$$\frac{\partial S}{\partial \hat{\mu}} = \sum_{i=1}^{N} -2(X_i - \hat{\mu}) = 0$$

$$\to \sum_{i=1}^{N}(X_i - \hat{\mu}) = 0$$

$$\to \sum_{i=1}^{N} X_i = \sum_{i=1}^{N} \hat{\mu}$$

$$\to N \cdot \bar{X} = N \cdot \hat{\mu}$$

$$\to \hat{\mu} = \bar{X}$$

(e) Yes, the sample mean $\bar{X}$ is the best linear unbiased estimator (BLUE) of the population mean. In part(a), we proved $E[\bar{X}] = \mu$, and thus the sample mean is an unbiased estimator. Thus, we must show $\bar{X}$ is the variance minimizing estimator for μ. To prove this, take an estimator of $\mu, \hat{\mu} = \sum_{i=1}^{N}(\omega_i \cdot X_i)$, where $\sum_{i=1}^{N} \omega_i = 1$. The problem then becomes:

$$\min Var(\hat{\mu}) \; s.t. \; \omega_1 + ... + \omega_N = 1$$

$$\to \min Var(\sum_{i=1}^{N}(\omega_i \cdot X_i)) \; s.t. \; \omega_1 + ... + \omega_N = 1$$

$$\to \min Var(\omega_1 \cdot X_1 + ... + \omega_N \cdot X_N) \; s.t. \; \omega_1 + ... + \omega_N = 1$$

$$\to \min Var(\omega_1 \cdot X_1) + ... + Var(\omega_N \cdot X_N) \; s.t. \; \omega_1 + ... + \omega_N = 1$$

$$\to \min \omega_1^2 Var(X_1) + ... + \omega_N^2 Var(X_N) \; s.t. \; \omega_1 + ... + \omega_N = 1$$

$$\to \min \omega_1^2 \sigma_X^2 + ... + \omega_N^2 \sigma_X^2 \; s.t. \; \omega_1 + ... + \omega_N = 1$$

$$\to \mathscr{L} = \omega_1^2 \sigma_X^2 + ... + \omega_N^2 \sigma_X^2 - \lambda(\omega_1 + ... + \omega_N - 1)$$

$$\frac{\partial \mathscr{L}}{\partial \omega_1} = 2\sigma_X^2 \omega_1 - \lambda = 0$$

$$...$$

$$\frac{\partial \mathscr{L}}{\partial \omega_N} = 2\sigma_X^2 \omega_N - \lambda = 0$$

$$\to 2\sigma_X^2 \omega_1 - \lambda = ... = 2\sigma_X^2 \omega_N - \lambda$$

$$\to 2\sigma_X^2 \omega_1 = ... = 2\sigma_X^2 \omega_N$$

$$\to \omega_1 = ... = \omega_N$$

$$\to \omega_1 + ... + \omega_N = 1 = N \cdot \omega_1$$

$$\to \omega_1 = ... = \omega_N = \frac{1}{N}$$

$$\to \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X}$$

Problem 2

1. For each of the following state whether or not the estimator is biased, consistent, both or neither, when used to estimate the population mean:

   (a) $\tilde{X} = \frac{1}{N-1}\sum_{i=1}^{N} X_i$

   $$E[\tilde{X}] = E[\frac{1}{N-1}\sum_{i=1}^{N} X_i] = \frac{1}{N-1}\sum_{i=1}^{N} E[X_i] = \frac{1}{N-1}\sum_{i=1}^{N} \mu = \frac{1}{N-1}\cdot N\mu \neq \mu$$

   $$\lim_{N\to\infty} Var(\tilde{X}) = \lim_{N\to\infty} Var(\frac{1}{N-1}\sum_{i=1}^{N} X_i) = \lim_{N\to\infty} \frac{N}{(N-1)^2}\cdot Var(X) = 0$$

   $\therefore \tilde{X}$ displays consistency but not unbiasedness.

   (b) $\hat{X} = \frac{2}{N}\sum_{i=1}^{N/2} X_i$

   $$E[\hat{X}] = E[\frac{2}{N}\sum_{i=1}^{N/2} X_i] = \frac{2}{N}\sum_{i=1}^{N/2} E[X_i] = \frac{2}{N}\cdot\frac{N}{2}\cdot\mu = \mu$$

   $$\lim_{N\to\infty} Var(\hat{X}) = \lim_{N\to\infty} Var(\frac{2}{N}\sum_{i=1}^{N/2} X_i) = \lim_{N\to\infty} \frac{4}{N^2}\cdot\frac{N}{2}\cdot\sigma_X^2 = \lim_{N\to\infty} \frac{2}{N}\cdot\sigma_X^2 = 0$$

   $\therefore \hat{X}$ displays both consistency and unbiasedness.

   (c) Assuming N is even, $\bar{X} = \frac{2}{N}\sum_{i=1}^{\frac{N}{2}}(X_i+\mu) + \frac{2}{N}\sum_{i=\frac{N}{2}+1}^{N}(X_i-\mu) = \frac{2}{N}[\frac{N}{2}\mu - \frac{N}{2}\mu + \sum_{i=1}^{N} X_i] = \frac{2}{N}\sum_{i=1}^{N} X_i$

   $$E[\bar{X}] = E[\frac{2}{N}\sum_{i=1}^{N} X_i] = \frac{2}{N}\sum_{i=1}^{N} E[X_i] = \frac{2}{N}\cdot N\cdot\mu = 2\mu$$

   $$\lim_{N\to\infty} Var(\bar{X}) = \lim_{N\to\infty} Var(\frac{2}{N}\sum_{i=1}^{N} X_i) = \lim_{N\to\infty} \frac{4}{N^2}\cdot N\cdot Var(X) = \lim_{N\to\infty} \frac{4}{N}\sigma_X^2 = 0$$

   $\therefore \bar{X}$ displays consistency but not unbiasedness.

   (d) $Y \sim N(\mu, \sigma^2)$

   $$E[Y] = \mu$$
   $$\lim_{N\to\infty} Var(Y) = \sigma^2$$

   $\therefore$ Y is unbiased but not consistent.

   (e) $Z = \sum_{i=1}^{N} w_i X_i$ where $\sum_{i=1}^{N} w_i = 1$.

   $$E[Z] = E[\sum_{i=1}^{N} w_i X_i] = \sum_{i=1}^{N} E[w_i X_i] = \sum_{i=1}^{N} E[w_i]E[X_i] = \mu\sum_{i=1}^{N} E[w_i] = \mu$$

   $$\lim_{n\to\infty} Var(Z) = \lim_{n\to\infty} Var(\sum_{i=1}^{N} w_i X_i) = \lim_{n\to\infty} \sum_{i=1}^{N} w_i^2 Var(X_i) = 0$$

   $\therefore Z$ is both consistent and unbiased.

Problem 3

3. Examine the following economic model

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

(a) Derive the formula for the sample least squares estimator for the parameters α and β.

(b) In the regression of X on Y (the reverse of the above), what is the formula for the least squares estimator for the slope parameter on Y?

(c) If the slope parameter for the reverse regression is δ. Is the value of $\delta \times \beta = 1$? Explain your reasoning.

(d) Show that the geometric mean of δ and β is equal to the correlation coefficient.

3. (a) $\min SSR = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N}(Y_i - (\hat{\alpha} + \hat{\beta}X_i))^2$

$$\frac{\partial SSR}{\partial \alpha} = -2\sum_{i=1}^{N}(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$$

$$\rightarrow \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N}(\hat{\alpha} + \hat{\beta}X_i)$$

$$\rightarrow N\bar{Y} = N\hat{\alpha} + N\hat{\beta}\bar{X}$$

$$\rightarrow \bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$$

$$\rightarrow \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\frac{\partial SSR}{\partial \beta} = -2\sum_{i=1}^{N} X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$$

$$\rightarrow \sum_{i=1}^{N} X_iY_i = \sum_{i=1}^{N}(\hat{\alpha}X_i + \hat{\beta}X_i^2)$$

$$\rightarrow \sum_{i=1}^{N} X_iY_i = \hat{\alpha}N\bar{X} + \hat{\beta}\sum_{i=1}^{N} X_i^2$$

$$\rightarrow \sum_{i=1}^{N} X_iY_i = N(\bar{Y} - \hat{\beta}\bar{X})\bar{X} + \hat{\beta}\sum_{i=1}^{N} X_i^2$$

$$\rightarrow \sum_{i=1}^{N} X_iY_i - N\bar{X}\bar{Y} = \hat{\beta}(\sum_{i=1}^{N} X_i^2 - N\bar{X}^2)$$

$$\rightarrow \hat{\beta} = \frac{\sum_{i=1}^{N} X_iY_i - \bar{X}\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i^2 - \bar{X}\sum_{i=1}^{N} X_i}$$

$$\rightarrow \hat{\beta} = \frac{\sum_{i=1}^{N} X_iY_i - \bar{X}Y_i}{\sum_{i=1}^{N} X_i^2 - \bar{X}X_i}$$

$$\rightarrow \hat{\beta} = \frac{\sum_{i=1}^{N} Y_i(X_i - \bar{X})}{\sum_{i=1}^{N} X_i(X_i - \bar{X})}$$

$$\rightarrow \hat{\beta} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})}$$

$$\rightarrow \hat{\beta} = \frac{\hat{cov}(X,Y)}{\hat{var}(X)}$$

(b) Slope parameter on Y $= \hat{\delta} = \frac{\hat{cov}(Y,X)}{\hat{var}(Y)}$

(c) No, $\hat{\delta} \times \hat{\beta} = \frac{\hat{cov}(Y,X)}{\hat{var}(Y)} \times \frac{\hat{cov}(X,Y)}{\hat{var}(X)} = \frac{(\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}))^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2} \neq 1$

(d) $\hat{\delta} \times \hat{\beta} = \frac{\hat{cov}(X,Y)^2}{\hat{var}(X)\hat{var}(Y)}$

$$\hat{\rho} = \frac{\hat{cov}(X,Y)}{\hat{\sigma_X}\hat{\sigma_Y}}$$

$$\rightarrow \hat{\rho} = \sqrt{\hat{\delta} \times \hat{\beta}}$$

Problem 4

4. There are two populations of individuals called *samies varies* respectively. The height of individuals in the *samies* is given by:

$$X_i \sim \mu + \varepsilon_i$$

And the height of individuals in the *varies* is given by:

$$Y_i \sim \mu + \epsilon_i$$

Where $\varepsilon_i \sim iid(0, \sigma^2)$ and $\epsilon_i \sim iid(0, 4\sigma^2)$.

(a) Is the sample mean from the population of *samies* an unbiased estimator of μ?

(b) Is the sample mean from the population of *varies* an unbiased estimator of μ and consistent?

(c) Which of the two previous estimators is most efficient, and why?

(d) You have a sample of N individuals from each population. Is the average of the two sample means unbiased? Is this the best estimator you can construct?

(e) For the previous example, if relevant construct a BLUE estimator. If not, prove why the mean of the sample means is best.

4. (a) Yes, $E[\bar{X}] = E[\frac{1}{N}\sum_{i=1}^{N} X_i] = \frac{1}{N}\sum_{i=1}^{N} E[X_i] = \frac{1}{N} \cdot N \cdot \mu_X = \mu_X$

(b) Yes, $E[\bar{Y}] = E[\frac{1}{N}\sum_{i=1}^{N} Y_i] = \frac{1}{N}\sum_{i=1}^{N} E[Y_i] = \frac{1}{N} \cdot N \cdot \mu_Y = \mu_Y$. And $\lim_{N\to\infty} Var(\bar{Y}) = \lim_{N\to\infty} Var(\frac{1}{N}\sum_{i=1}^{N} Y_i) = \lim_{N\to\infty} \frac{1}{N^2} \cdot N \cdot \sigma_Y^2 = \lim_{N\to\infty} \frac{4\sigma^2}{N} = 0$.

(c) $\bar{X}$ is a more efficient estimator for its respective population mean because $Var(\bar{X}) = \frac{\sigma^2}{N} \leq Var(\bar{Y})$.

(d) $E[\frac{\bar{X}+\bar{Y}}{2}] = E[\frac{\bar{X}}{2}] + E[\frac{\bar{Y}}{2}] = \frac{1}{2}E[\bar{X}] + \frac{1}{2}E[\bar{Y}] = \frac{\mu}{2} + \frac{\mu}{2} = \mu$
$\therefore$ the average of the sample means is an unbiased estimator of the population mean.

$$\min Var(a\bar{X} + b\bar{Y}) \; s.t. a + b = 1$$

Since X_i and Y_i are i.i.d.

$$Var(a\bar{X} + b\bar{Y}) = a^2 Var(\bar{X}) + b^2 Var(\bar{Y}) = a^2\sigma^2 + 4b^2\sigma^2$$

So, the minimization problem becomes $\min a^2\sigma^2 + 4b^2\sigma^2 \; s.t. a + b = 1$

$$\mathscr{L} = a^2\sigma^2 + 4b^2\sigma^2 - \lambda(a + b - 1)$$

$$\frac{\partial \mathscr{L}}{\partial a} = 2\sigma^2 a - \lambda = 0$$

$$\rightarrow a = \frac{\lambda}{2\sigma^2}$$

$$\frac{\partial \mathscr{L}}{\partial b} = 8\sigma^2 b - \lambda = 0$$

$$\rightarrow b = \frac{\lambda}{8\sigma^2} = \frac{a}{4}$$

$$\frac{\partial \mathscr{L}}{\partial \lambda} = a + b - 1 = 0$$

$$\rightarrow a + \frac{a}{4} = 1$$

$$\rightarrow a = \frac{4}{5},\; b = \frac{1}{5}$$

$\therefore$ No, the average of the sample means is not the best estimator one can construct.

(e) Looking at the previous part, the best linear unbiased estimator (BLUE) of the population mean is $\hat{\mu} = \frac{4}{5}\bar{X} + \frac{1}{5}\bar{Y}$