Homework 1

Amy Fan

9/10/23

(0) Instructions for installing tinytex for PDF rendering: **https://yihui.org/tinytex/**

```
install.packages('tinytex')
tinytex::install_tinytex()
```

I. Mathematical Derivations [20 pts]

(1) Show that Cov(e_i, e_j) = -σ^2/n]

Show: Cov(e_i, e_j) = -σ^2

Proof:

$$\begin{aligned}
&Cov(e_i, e_j) \\
&= E[(e_i - E(e_i))(e_j - E(e_j))] \\
&= E(e_i e_j) \\
&= E[(Y_i - \overline{Y})(Y_j - \overline{Y})] \\
&= E(Y_i)E(Y_j) - E(\overline{Y})E(Y_i) - E(\overline{Y})E(Y_j) + E(\overline{Y}^2) \\
&= \mu^2 - \mu^2 - \mu^2 + E(\overline{Y}^2) \\
&= -\mu^2 + E(\overline{Y}^2)
\end{aligned}$$

We know

$$Var(\overline{Y}) = \mu^2 - E(\overline{Y}^2) = \frac{\sigma^2}{n}$$

So

$$-\frac{\sigma^2}{n} = -\mu^2 + E(\overline{Y}^2) = cov(e_i, e_j)$$

(2) ALRM Exercise 1.5: When asked to state the simple linear regression model, a student wrote it as follows: $E(Y_i) = \beta_0 + \beta_1 * X_i + \epsilon_i$. Do you agree?

No, while E(Y_i) is equal to the linear regression Y value, the last term should be 0 since the E(ϵ_i} = 0. The correct statement would be: $E(Y_i) = \beta_0 + \beta_1 * X_i$

(3) ALRM Exercise 1.7: In the simulation exercise, regression model (1.1) applies with $\beta_0 = 100$, $\beta_1 = 20$, and $\sigma^2 = 25$. An observation on Y will be made for X = 5.

a. Can you state the exact probability that Y will fall between 195 and 205? Explain.

No, you do not know the distribution, so you cannot create a confidence interval with some pdf function.

b. If the normal error regression model (1.24) is applicable, can you now state the exact probability that Y will fall between 195 and 205? If so, state it.

Since we know the distribution of E{Y_i}, we can. E{Y_i} = 100 + 20*5 = 200 195 < 200 < 205 is exactly one standard deviatino away, so the probability is 68%.

(4) ALRM Exercise 1.12 (a)-(c): In a study of the relationship for senior citizens between physical activity and frequency of colds, participants were asked to monitor their weekly time spent in exercise over a five-year period and the frequency of colds. The study demonstrated that a negative statistical relation exists between time spent in exercise and frequency of colds. The investigator concluded that increasing the time spent in exercise is an effuctive strategy for reducing the frequency of colds for senior citizens.

a. Were the data obtained in the study observational or experimental data?

Observational data

b. Comment on the validity of the conclusions reached by the investigator.

Since this is not experimental data, there are no controls set in place to place this conclusion. There could be covariates that are not being considered in the study that are actually leading to less colds.

c. Identify two or three other explanatory variables that might affect both the time spent in exercise and the frequency of colds for senior citizens simultaneously.

1. Temperature could be a confounding variable. Lower temperatures can lead to less outdoor time and more colds.
2. Underlying autoimmune health conditions could decrease the exercise rates. It could also lead to more likely infections such as colds.

(5) ALRM Exercise 1.32: Derive the expression for b_1 in (1.10a) from the normal equations in (1.9).

1.9 equations:

$$\sum Y_i = nb_0 + b_1 \sum x_i$$
$$\sum x_i Y_i = b_0 \sum x_i + b_1 \sum x_i^2$$

Then it follows from the first equation,

$$b_0 = \frac{\sum Y_i - b_1 \sum x_i}{n}$$
$$b_0 = \overline{Y} - b_1 \overline{X}$$

And from the second equation,

$$b_0 = \frac{\sum X_i Y_i - b_1 \sum X_i^2}{\sum X_i}$$
$$b_0 \sum X_i = \sum X_i Y_i - b_1 \sum X_i^2$$

Now, substitute the left side b_0 with the first b_0 equation derived,

$$\begin{aligned}
\bar{Y}\sum X_i - b_1\bar{X}\sum X_i &= \sum X_iY_i - b_1\sum {X_i}^2 \\
\bar{Y}\bar{X} - b_1\bar{X}^2 &= \overline{XY} - b_1\overline{X^2} \\
b_1\overline{X^2} - b_1\bar{X}^2 &= \overline{XY} - \bar{X}\bar{Y} \\
b_1(\overline{X^2} - \bar{X}^2) &= \overline{XY} - \bar{X}\bar{Y} \\
b_1 &= \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2} \\
&^* = \frac{\frac{1}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum (X_i - \bar{X})^2} \\
&= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}
\end{aligned}$$

*= can be seen because,

$$\begin{aligned}
&E(XY) - E(X)E(Y) \\
&= cov(X, Y) \\
&= E[(X - \bar{X})(Y - \bar{Y})] \\
&= \frac{1}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})
\end{aligned}$$

(6) ALRM Exercise 1.36: Prove the result in (1.20) – that the sum of the residuals weighted by the fitted values is zero. Prove:

$$\sum \hat{Y}_i e_i = 0$$

Proof:

$$\begin{aligned}
&\sum \hat{Y}_i e_i \\
&= \sum (X_i\beta_0 + \beta_1)e_i \\
&= \beta_0 \sum X_i e_i + B_1 \sum e_i \\
&= 0
\end{aligned}$$

Since both terms equal zero due to the properties of simple linear regression.

II. Tree Cover in Tucson

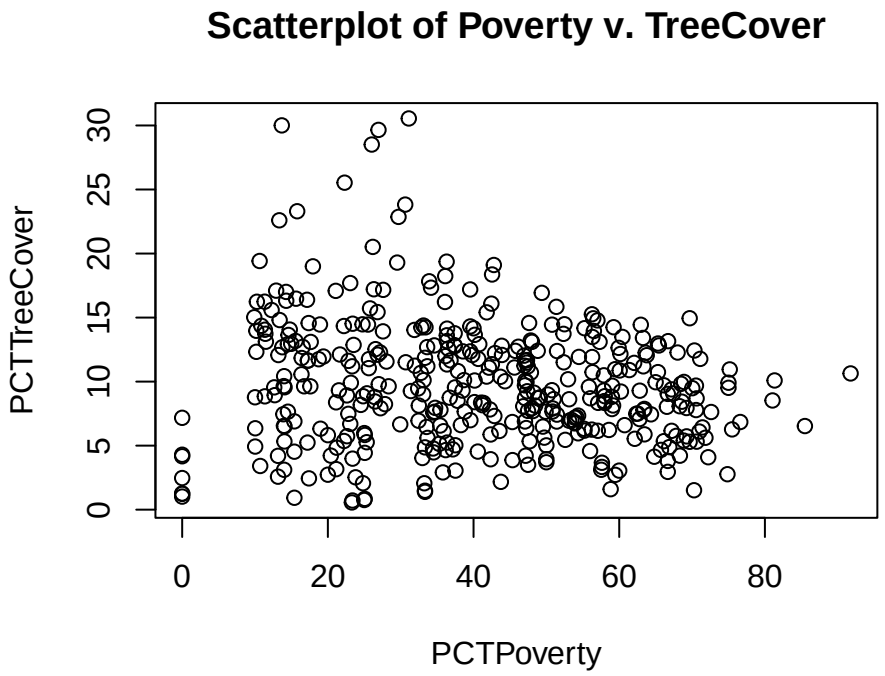
Download the Tree Equity Scores data set from Module 1 on D2L (Tree_Equity_Scores_Tucson_noNA.csv). Use the read.csv() function to import the comma spaced values as a dataframe in R. There are several variables recorded in this data set, but we will focus on two (read more here: https://gisdata.tucsonaz.gov/datasets/cotgis::tree-equity-scores-tucson-1/about). The first is PCTTreeCover, the percent of each corresponding census tract that is covered by tree canopy. The second is PCTPoverty, the percentage of households in each census tract with an income below the poverty line.

(7) Make a scatter plot of percent tree cover on the vertical axis and percent below the poverty line on the horizontal axis. Do you see any statistical relationship in the pattern of the points? If yes, what do you see? If no, why not? (Treat percent tree cover as the response variable and percent below poverty line as a predictor. Assume the simple linear regression model (eq. 1.1) is appropriate.)

There appears to be a negative association between the two variables. Tree Cover appears to decrease with increasing poverty.

```
rm(list = ls())
trees<-read.csv("./Datasets/Tree_Equity_Scores_Tucson_noNA.csv")
```

```
plot(x = trees$PCTPoverty,
     y = trees$PCTTreeCover,
     xlab = "PCTPoverty",
     ylab = "PCTTreeCover",
     main = "Scatterplot of Poverty v. TreeCover")
```



(8) Obtain the least-squares estimates of the intercept and slope, and state the estimated regression function.

PCTTreeCover = -0.03812227 * PCTPoverty + 11.29998467

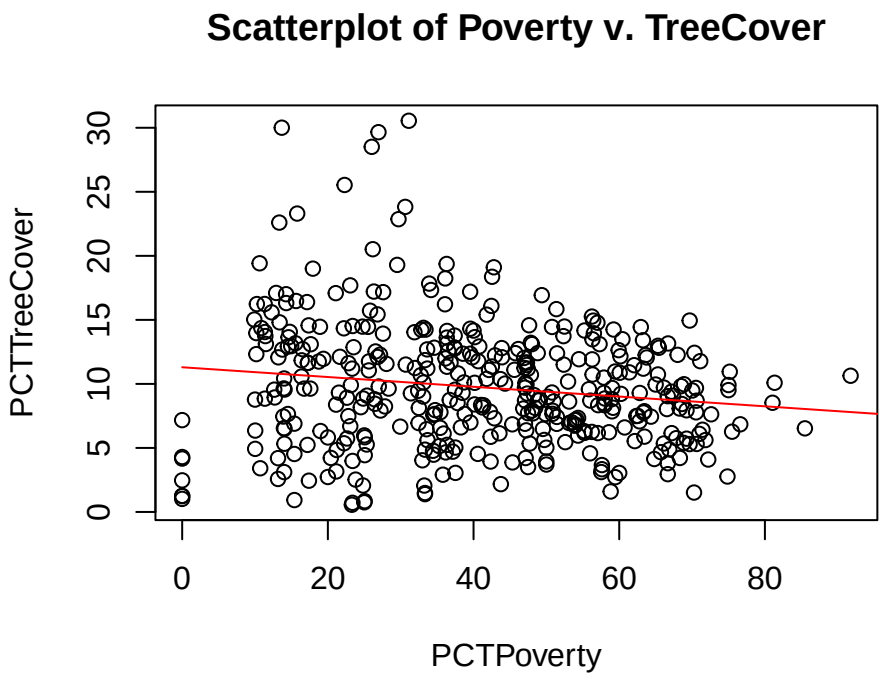
```
trees_fit <- lm(PCTTreeCover~PCTPoverty, data = trees)
coef(trees_fit)
```

```
## (Intercept)  PCTPoverty
## 11.29998467 -0.03812227
```

(9) Add a line. Does it fit the data well?

The estimated line fits the data decently, but there appears to be some heteroskedasticity. Lower percentages of poverty have higher variance.

```
plot(x = trees$PCTPoverty,
     y = trees$PCTTreeCover,
     xlab = "PCTPoverty",
     ylab = "PCTTreeCover",
     main = "Scatterplot of Poverty v. TreeCover")
X = seq(0,100,1)
regresline = predict(trees_fit, newdata = data.frame(PCTPoverty = X))
matlines(x = X, regresline, col = 'red')
```



(10) Give a point estimate of the mean percent tree cover for census tracts with 50% of households below the poverty line.

9.393871

```
predict(trees_fit, newdata = data.frame(PCTPoverty = 50))
```

```
##        1
## 9.393871
```

(11) How much is percent tree cover expected to change when the percent of households below the poverty line increased by 20%?

It decreases by 0.7624453%

```
slope = coef(trees_fit)[[2]]
20*slope
```

```
## [1] -0.7624453
```

(12) Use R to compute the MLE for σ^2. How does it compare to the MSE? In what units is σ^2 expressed?

The mle is slightly biased too high since the sum of squared errors is divided by n instead of with the n-1 correction. Since n is relatively large (405), the difference in estimates is pretty small. The units of sigma^2 is in percent squared.

```
mse = (summary(trees_fit)$sigma)^2

n = nrow(trees)
mle = sum((trees$PCTTreeCover-mean(trees$PCTTreeCover))^2)/n

mse
```

```
## [1] 22.76027
```

```
mle
```

```
## [1] 23.18175
```