MATH 392 Problem Set 7

Exercises from the book

11.1: #4

Q: For $i = 1, ..., n$, let $\hat{y}_i = \beta_o + \beta_1 x_i$.

Show that $\hat{\beta}_0$ and $\hat{\beta}_1$, as given in Eq. (11.1.1), are unique values of β_0 and β_1 such that:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

And:

$$\sum_{i=1}^{n} x_i (y_i - \hat{y}_i) = 0$$

Note:

Eq. (11.1.1):

(1): $\hat{\beta}_1 = \dfrac{\sum\limits_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$

(2): $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Eq. (11.1.3):

(3): $\frac{\partial Q}{\partial \beta_0} = -2 \sum\limits_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$

Eq. (11.1.4):

(4): $\frac{\partial Q}{\partial \beta_1} = -2 \sum\limits_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i$

A:

Note: The values of β_0 and β_1 must satisfy relation (3) as above, such that $\frac{\partial Q}{\partial \beta_0} = 0$.

Thus, noting that $\hat{y}_i = \beta_0 + \beta_1 x_i$, and taking out the constant, this is to say:

$\frac{\partial Q}{\partial \beta_0} = 0 = \sum\limits_{i=1}^{n} (y_i - \hat{y}_i)$.

However, the above relations must also satisfy relation (4) given above, leading us to a similar statement:

$\frac{\partial Q}{\partial \beta_1} = 0 = \sum\limits_{i=1}^{n} (y_i - \hat{y}_i) x_i$

Combining the above relations together, we obtain the following, as noted in Eq. (11.1.5):

(5): $\beta_0 n + \beta_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i$

(6): $\beta_0 \sum\limits_{i=1}^{n} x_i + \beta_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i$

We then note some shorthand, namely:

(7): $\bar{x} = \frac{1}{n} \sum\limits_{i=1}^{n} x_i$

(8): $\bar{y} = \frac{1}{n} \sum\limits_{i=1}^{n} y_i$

Bearing in mind relations (7) and (8), we may then note relations (5) and (6) are the same as those given in (1) and (2) respectively, such that we conclude that $\hat{\beta}_0$ and $\hat{\beta}_1$, as given in Eq. (11.1.1), are unique values of β_0 and β_1.

Notational Note

In the exercises that follow, we use the following shorthand:

(*): $ss_x = \sum\limits_{i=1}^{n} (x_i - \bar{x}_n)^2$

11.2 #2

Q: Show that $E\left(\hat{\beta}_1\right) = \beta_1$

Note:

Eq. (11.2.7)

(1): $\hat{\beta}_1 = \dfrac{\sum\limits_{i=1}^{n}(x_i - \bar{x})Y_i}{ss_x}$

A:

Note: $E(Y_i) = \beta_0 + \beta_1 x_i$. Combining this with the above relation (1) gives us:

$E\left(\hat{\beta}_1\right) = \dfrac{\sum\limits_{i=1}^{n}(x_i - \bar{x}_n)(\beta_0 + \beta_1 x_i)}{ss_x}$

Then, note: $\sum\limits_{i=1}^{n} (x_i - \bar{x}_n) = 0$.

Furthermore, we note: $\sum\limits_{i=1}^{n} (x_i - \bar{x}_n)x_i = \sum\limits_{i=1}^{n} x_i(x_i - \bar{x}_n) - \bar{x}_n \sum\limits_{i=1}^{n} (x_i - \bar{x}_n) = \sum\limits_{i=1}^{n} (x_i - \bar{x}_n)^2 = ss_x$

Thus, we're left with:

$E\left(\hat{\beta}_1\right) = \frac{ss_x \beta_1}{ss_x} = \beta_1$

11.2 #3

Q: Show that $E\left(\hat{\beta}_0\right) = \beta_0$

A:

Note the following relations:

(1): $E(\bar{Y}_n) = \frac{1}{n} \sum\limits_{i=1}^{n} E(Y_i)$

(2): $E(\bar{Y}_n) = \frac{1}{n} \sum\limits_{i=1}^{n} \beta_0 + \beta_1 x_i$

(3): $E(\bar{Y}_n) = \beta_0 + \beta_1 \bar{x}_n$

Thus, as:

$$\bar{Y}_n = \beta_0 + \beta_1 \bar{x}_n \to \beta_0 = \bar{Y}_n - \beta_1 \bar{x}_n$$

We may note:

$$E\left(\hat{\beta}_0\right) = E(\bar{Y}_n - \beta_1 \bar{x}_n)$$

Taking advantage of linearity, we have:

$$E\left(\hat{\beta}_0\right) = E(\bar{Y}_n) - E(\beta_1 \bar{x}_n)$$

Using relation (3) given above, and the results of the prior exercise, we have:

$$E\left(\hat{\beta}_0\right) = \beta_0 + \beta_1 \bar{x}_n - \bar{\beta}_1 \bar{x}_n = \beta_0$$

11.2 #4

Q: Show that $Var\left(\hat{\beta}_0\right)$ is as given in Eq. (11.2.5).

Note:

Eq. (11.2.5)

(1): $\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{ss_x}\right)$

A:

Note the beginning of the prior exercise, specifically:

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n$$

Breaking out this equation gives us:

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \bar{x}_n$$

Using the equation for $\hat{\beta}_1$ gives us:

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i - \bar{x}_n \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n) Y_i}{ss_x}$$

We can simplify this equation as follows:

Double Check this part

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}_n}{ss_x} (x_i - \bar{x}_n) Y_i\right)$$

Using the above equation for variance, we have:

$$Var(\hat{\beta}_0) = Var(\sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}_n}{ss_x} (x_i - \bar{x}_n) Y_i\right))$$

We assume independence of the Y's, such that we may say:

$$Var(\hat{\beta}_0) = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}_n}{ss_x} (x_i - \bar{x}_n)\right)^2 Var(Y_i)$$

Since each Y has variance σ^2, we have:

$$Var(\hat{\beta}_0) = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}_n}{ss_x} (x_i - \bar{x}_n)\right)^2$$

Expanding this equation gives us:

$$Var(\hat{\beta}_0) = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} - \frac{2\bar{x}_n}{nss_x}(x_i - \bar{x}_n) + \frac{\bar{x}_n^2}{ss_x^2}(x_i - \bar{x}_n)^2 \right)$$

Additionally, we note an alternative form of $\hat{\beta}_1$ as follows:

Check this part too

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}_n \bar{y}_n}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}_n^2}$$

Giving us:

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} - 0 + \frac{\bar{x}_n^2}{ss_x} \right)$$

Thus, we may conclude:

$$Var\left(\hat{\beta}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{ss_x} \right), \text{ as given in Eq. (11.2.5).}$$

11.2 #5

Q: Show that $Cov\left(\hat{\beta}_0, \hat{\beta}_1\right)$ is as given in Eq. (11.2.6). *Hint:* Use the result in Exercise 8 in Sec. 4.6.

Notes:

Eq. (11.2.6):

(1): $Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\bar{x}\sigma^2}{s_x^2} \equiv -\frac{\bar{x}\sigma^2}{ss_x}$

The equivalence relation is to solidfy the notation being used in this problemset, as noted in relation (*), and is repeated due to the reference to an equation used in the book.

Exercise 8, Sec. 4.6.:

(2): For $X_1, ..., X_m$ and $Y_1, ..., Y_n$ random variables, let $i = 1, ..., m$ and $j = 1, ..., n$. Suppose $Cov(X_i, Y_j)$ exists and let $a_1, ..., a_m$ and $b_1, ..., b_n$ be constants. Then the following holds:

(3): $Cov(\sum\limits_{i=1}^{m} a_i X_i, \sum\limits_{i=1}^{n} b_j Y_j) = \sum\limits_{i=1}^{m}\sum\limits_{i=1}^{n} a_i b_j Cov(X_i Y_j)$

A:

Note:

$$Var(\bar{Y}_n) = Var(\hat{\beta}_0) + Var(\hat{\beta}_1 \bar{x}_n) + Cov(\hat{\beta}_0, \hat{\beta}_1 \bar{x}_n)$$

Taking out values of $\bar{x}_n$ gives us:

$$Var(\bar{Y}_n) = Var(\hat{\beta}_0) + \bar{x}_n^2 Var(\hat{\beta}_1) + 2\bar{x}_n Cov(\hat{\beta}_0, \hat{\beta}_1)$$

Isolating $Cov\left(\hat{\beta}_0, \hat{\beta}_1\right)$ gives us:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{1}{2\bar{x}_n}\left(Var(\bar{Y}_n) - Var(\hat{\beta}_0) - \bar{x}_n^2 Var(\hat{\beta}_1)\right)$$

Note: The above relation holds for $\bar{x}_n \neq 0$.

Using the equations for the Variance of $\bar{Y}_n, \hat{\beta}_0,$ and $\hat{\beta}_1$ gives us:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{1}{2\bar{x}_n}\left(\frac{\sigma^2}{n} - \frac{\sum\limits_{i=1}^{n} x_i^2}{nss_x}\sigma^2 - \frac{\bar{x}_n^2}{ss_x}\sigma^2 \right)$$

Bringing out like terms gives us:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{\sigma^2}{2\bar{x}_n}\left(\frac{ss_x - \sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2}{nss_x}\right)$$

Simplifying this equation gives us:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{\sigma^2}{2\bar{x}_n}\left(\frac{-2n\bar{x}_n^2}{nss_x}\right) = -\frac{\bar{x}_n^2\sigma^2}{ss_x}$$

Another Condition

Let us then determine $Cov\left(\hat{\beta}_0, \hat{\beta}_1\right)$ when $\bar{x}_n = 0$.

Under this assumption, we have:

$$\hat{\beta}_0 = \bar{Y}_n.$$

Noting $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$, we have:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = Cov\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{1}{ss_x}\sum_{j=1}^{n} x_j Y_j\right)$$

Taking advantage of relation (3) given above gives us:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{1}{nss_x}\sum_{i=1}^{n}\sum_{j=1}^{n} x_j Cov(Y_i, Y_j)$$

In addition to taking advantage of relation (3) from Exercise 8, Sec. 4.6., note the Y's are independent and each has variance σ^2. Importantly, this means that for $i = j$, $Cov(Y_i, Y_j) = \sigma^2$, and for $i \neq j$, $Cov(Y_i, Y_j) = 0$.

Thus, we are left with the following relation:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{1}{nss_x}\sum_{i=1}^{n}\sum_{j=1}^{n} x_j Cov(Y_i, Y_j) = \frac{\sigma^2}{nss_x}\sum_{i=1}^{n} x_i = \frac{\bar{x}_n\sigma^2}{ss_x} = 0$$

However, under the condition of interest, $\bar{x}_n = 0$, we have: $Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{\sigma^2}{nss_x}\sum_{i=1}^{n} x_i = \frac{\bar{x}_n\sigma^2}{ss_x} = 0 = -\frac{\bar{x}_n^2\sigma^2}{ss_x}$,

and it holds that $Cov\left(\hat{\beta}_0, \hat{\beta}_1\right)$ is as given in Eq. (11.2.6) $\forall \bar{x}_n$.

Additional Exercises

1. Consider the following dataset:

```
set.seed(32)
n <- 10
x <- rnorm(n)
y <- -1 + 1.3 * x + rnorm(n, .3)
df <- data.frame(x, y)
```

We can find the least squares regression line by running `lm()` (which uses the normal equations), and extract the coefficient estimates.

```
m1 <- lm(y ~ x, data = df)
coef(m1)
```

```
## (Intercept)            x
##  -0.4488683    1.2495758
```

A more general approach to finding the estimates that optimize a loss function is to use a numerical optimization technique. Here we use `optim()` to minimize the RSS. By default this function uses the Nelder-Mead algorithm), but you can also toggle to another algorithm such as BFGS or select an entirely different optimization function/package.

```
RSS <- function(par, x, y) {
  beta_0 <- par[1]
  beta_1 <- par[2]
  sum((y - (beta_0 + beta_1 * x))^2)
}
opt <- optim(par = c(0, 0), fn = RSS, x = x, y = y)
```

The `par` argument is the set of values of the two parameters that you want to initialize the algorithm at. You can try several different values and see if the final estimates agree. The final estimates are found in the `opt` object.

```
opt$par
```

```
## [1] -0.4488259  1.2495164
```

Which agree very closely with the analytical solutions from the normal equations.

a. Using numerical optimization, find the estimates that minimize two additional loss functions: a) the absolute deviation in the `y` and b) the squared deviation in the x.

b. Plot all three lines on top of a scatterplot of the data. Add an `annotate()` layer or legend to make it clear which line is which.

c. Create a second scatterplot that again shows the least squares regression line. Add to this plot pairs of lines that represent each of the following intervals:

    - A confidence interval on β_1.
    - A confidence interval on $E(Y|X = x)$.
    - A prediction interval on $[Y|X = x]$.

2. Ecological Fallacy refers to a situation where one draws inferences on the individual level from data that was collected at the group level.

```
# install.packages("resampledata")
library(resampledata)
data(corrExerciseB)
```

a. Create a scatter plot of all of the data, which each group plotted in a different color. Add in the group means for each.

b. Compute two sample correlations: one for the group means, the other for all of the data. Under which conditions, stated informally, will the correlation at the group level exceed that at the individual level? Do you expect that this is the a more common or less common feature of aggregated data in the real world?

c. In a setting such as this, what is the consequence for your data analysis of committing an ecological fallacy?