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

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

Eq. (11.1.3):

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

Eq. (11.1.4):

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

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, -2, this is to say:

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

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

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

Combining the above relations, we obtain the following, leading us to Eq. (11.1.5):

$$\sum_{i=1}^{n} \beta_0 + \sum_{i=1}^{n} \beta_1 x_i = \sum_{i=1}^{n} y_i$$

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

$\sum_{i=1}^{n} \beta_0 x_i + \sum_{i=1}^{n} \beta_1 x_i^2 = \sum_{i=1}^{n} y_i$

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

We then note some shorthand, namely:

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

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

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, concluding that $\hat{\beta_0}$ and $\hat{\beta_1}$, are as given in Eq. (11.1.1), and are unique values of β_0 and β_1.

A Note on Notation

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

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

11.2 #2

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

Note:

Eq. (11.2.7)

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

A:

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

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

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

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

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

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

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

Thus, as:

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

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, 11.2.3, specifically:

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

Breaking out this equation gives us:

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

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

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

Note, as:

$Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

It then follows that:

$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) - \bar{x}_n \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)Y_i}{ss_x}$

After grouping terms of $\hat{\beta}_0$, we are left with:

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

Applying variance to the above relation, we have:

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

Under the assumption independence of the Y's,we may say:

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

Noting that 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:

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

Separating this relation, we have:

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

Looking at the latter-most term, we have:

$\sigma^2 \sum_{i=1}^{n} \frac{2\bar{x}_n}{nss_x} (x_i - \bar{x}_n) = \frac{2\sigma^2}{n} \sum_{i=1}^{n} \frac{\bar{x}_n}{x_i - \bar{x}_n}$

Note:

$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$

Such that we may say:

$\sigma^2 \sum_{i=1}^{n} \frac{2\bar{x}_n}{nss_x} (x_i - \bar{x}_n) = 0$ [This is hand-wavey and not proven, apologies]

Going back to relation (2), we may now say:

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

Thus, we may conclude:

$Var\left(\hat{\beta}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{ss_x} \right)$, 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): $Cov(\sum_{i=1}^{m} a_i X_i, \sum_{i=1}^{n} b_j Y_j) = \sum_{i=1}^{m} \sum_{i=1}^{n} a_i b_j Cov(X_i Y_j)$

A:

Warm-up Formulation

$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = Cov\left(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1\right) = Cov(\bar{y}, \hat{\beta}_1) - Cov(\bar{x}\hat{\beta}_1, \hat{\beta}_1)$

We assume independence of $\bar{y}, \hat{\beta}_1$, thus it follows that:

$Cov(\bar{y}, \hat{\beta}_1) = 0$

Thus:

$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -Cov(\bar{x}\hat{\beta}_1, \hat{\beta}_1) = -\bar{x}Cov(\hat{\beta}_1, \hat{\beta}_1) = -\bar{x}Var(\hat{\beta}_1)$

As we know $Var(\hat{\beta}_1)$, this relation simplifies to:

$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\bar{x}\sigma^2}{ss_x}$, as given in Eq. 11.2.6.

Alternative Formulation

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 $\bar{x}_n \neq 0$, we tend not to like dividing by zero.

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

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

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:

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

All Praise the Glorious Hints When They are Given

Tithe aside, in addition to taking advantage of relation (3) from Exercise 8, Sec. 4.6., we 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.

```
# A Absolute Deviation in the y
abs_dif <- function(par, x, y) {
  beta_0 <- par[1]
  beta_1 <- par[2]
  sum(abs(y - (beta_0 + beta_1 * x)))
}
opt_abs <- optim(par = c(1, 1), fn = abs_dif, x = x, y = y)

opt_abs$par
```

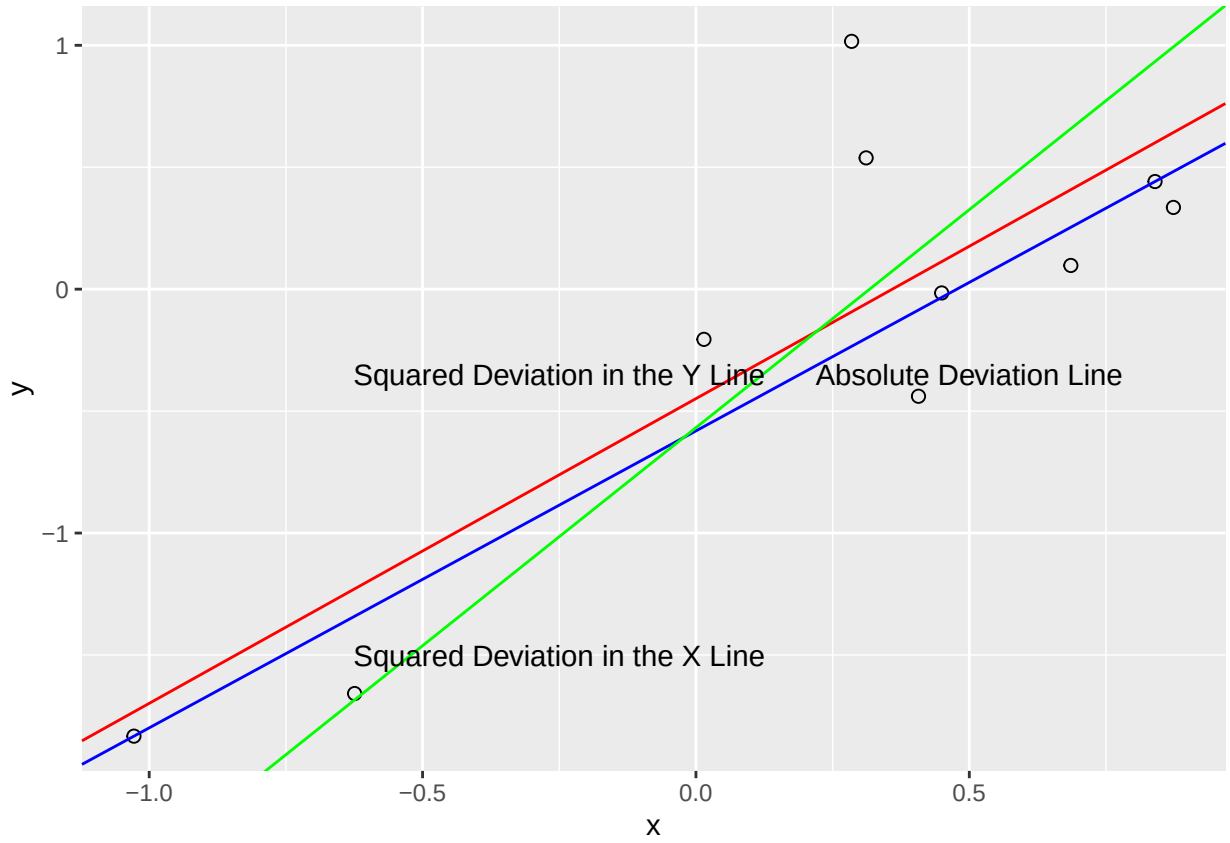
```
## [1] -0.5811111  1.2176767
```

```
# B Squared Deviation in the X
ss_x <- function(par, x, y) {
  beta_0 <- par[1]
  beta_1 <- par[2]
  sum((x - (y/beta_1) + (beta_0/beta_1) )^2)
}
opt_ss_x <- optim(par = c(1, 1), fn = ss_x, x = x, y = y)

opt_ss_x$par
```

```
## [1] -0.5679669  1.7879482
```

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.

```
ggplot(data=df, aes(x=x, y=y)) +
  geom_point(size=2, shape=1) +
  geom_abline(slope = 1.2495164, intercept = -0.4488259, colour="red") +
  geom_abline(slope = 1.2176767, intercept = -0.5811111, colour="blue") +
  geom_abline(slope = 1.7879482, intercept = -0.5679669, colour="green") +
  annotate("text", x = -.25, y = -.35, label = "Squared Deviation in the Y Line") +
  annotate("text", x = .5, y = -.35, label = "Absolute Deviation Line") +
  annotate("text", x = -.25, y = -1.5, label = "Squared Deviation in the X Line")
```



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 $[Y|X = x]$.

```
# compute 95% confidence interval for \beta_1
confint(m1)
```

```
##                 2.5 %      97.5 %
## (Intercept) -0.8680975 -0.02963914
## x            0.5808445  1.91830710
```

```
# compute 95% confidence interval for E(Y | X = x)
conf_values <-predict(m1, newdata = df, interval = 'confidence')
conf_data <- data.frame(conf_values)

df_conf <- cbind(df, conf_data)
df_conf
```

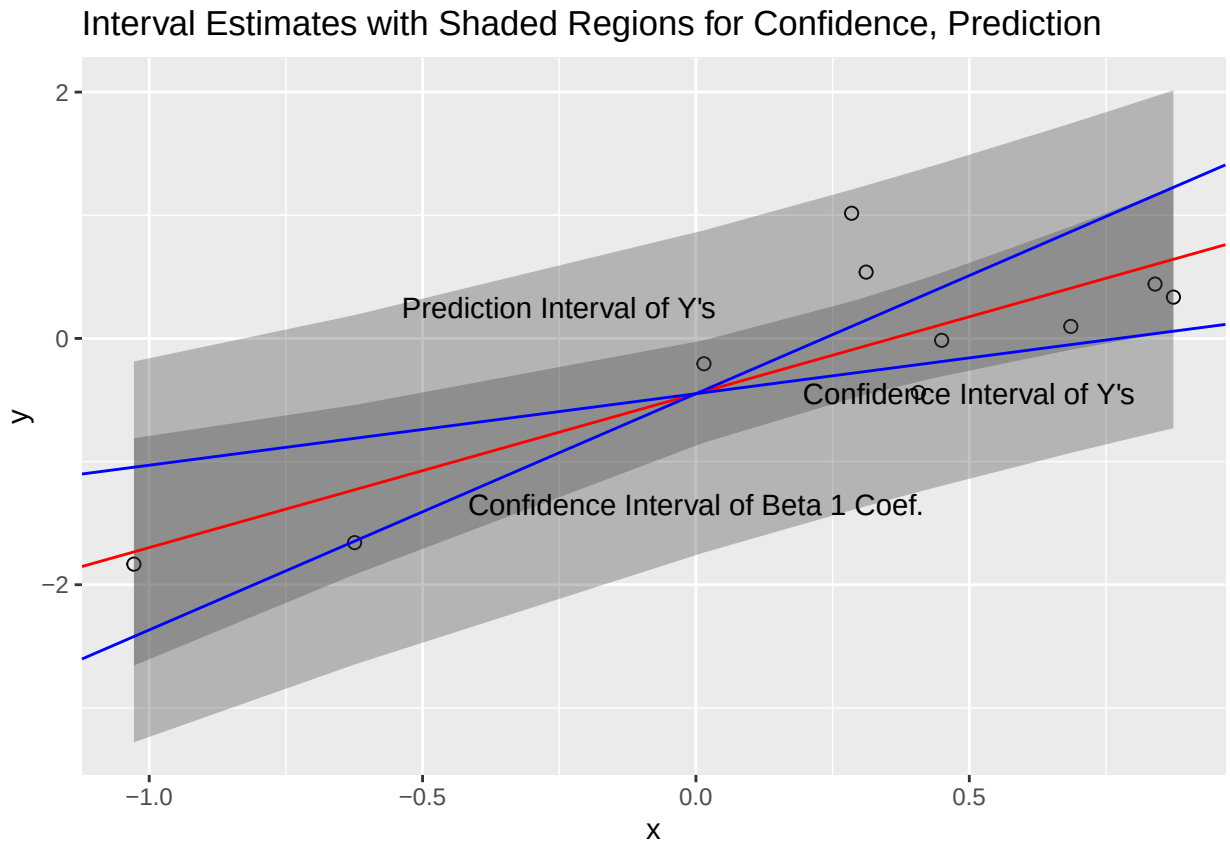
```
##              x           y         fit         lwr         upr
## 1   0.01464054 -0.20571383 -0.43057387 -0.84644711 -0.01470064
## 2   0.87328871  0.33518528  0.64237210  0.05592798  1.22881621
## 3  -1.02794620 -1.83281755 -1.73336503 -2.65629741 -0.81043265
```

```
## 4   0.68566463  0.09726232  0.40792158 -0.09233426  0.90817743
## 5   0.44943698 -0.01549706  0.11273724 -0.30810573  0.53358020
## 6   0.40701764 -0.43910434  0.05973105 -0.35168082  0.47114292
## 7   0.28473137  1.01620919 -0.09307491 -0.48758681  0.30143699
## 8  -0.62430939 -1.65803631 -1.22899024 -1.91721366 -0.54076682
## 9   0.83965601  0.44131851  0.60034549  0.03042326  1.17026772
## 10  0.31127919  0.53839578 -0.05990139 -0.45671353  0.33691074
```

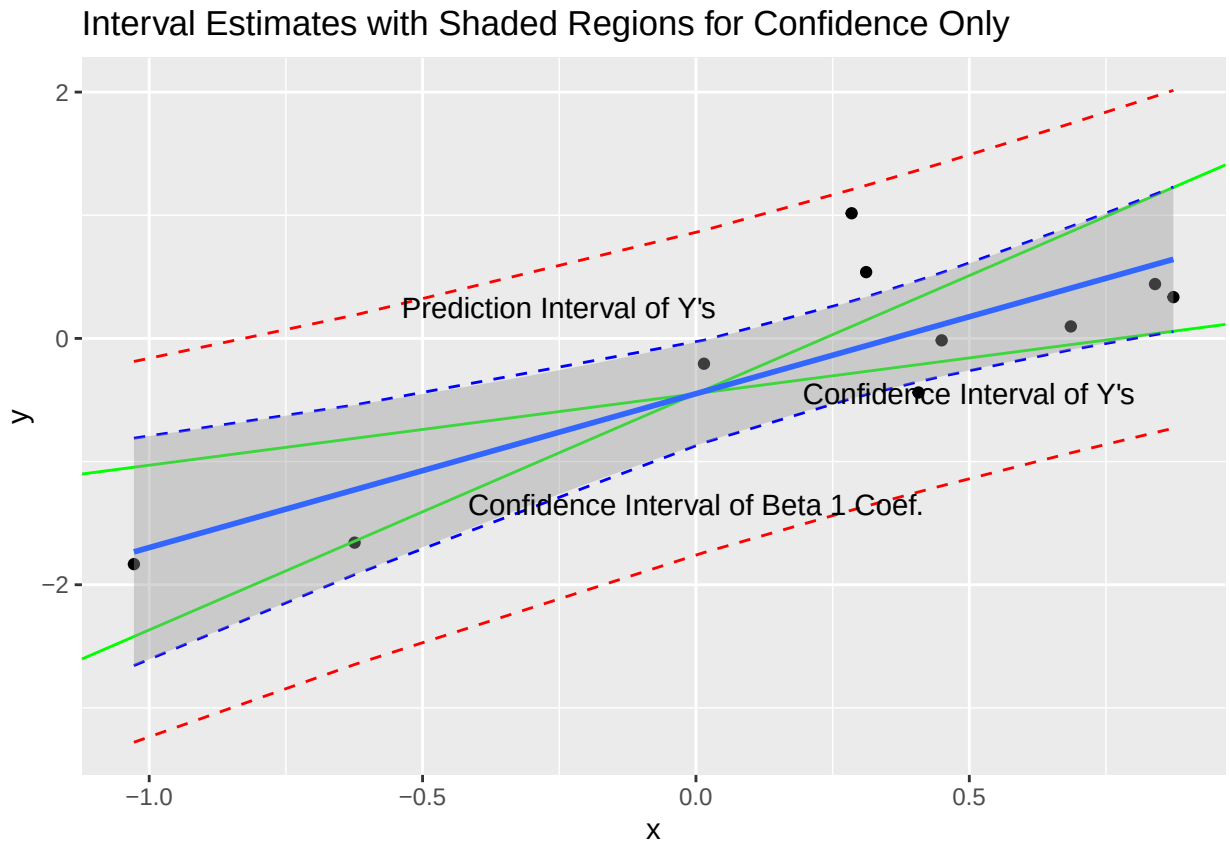
```r
# compute 95% prediction interval on [Y | X = x]
pred_values <- predict(m1, newdata = df, interval = 'prediction')
pred_data <- data.frame(pred_values)

df_pred <- cbind(df, pred_data)
df_pred
```

```
##              x           y         fit        lwr        upr
## 1   0.01464054 -0.20571383 -0.43057387 -1.7387710  0.8776232
## 2   0.87328871  0.33518528  0.64237210 -0.7296140  2.0143582
## 3  -1.02794620 -1.83281755 -1.73336503 -3.2794030 -0.1873271
## 4   0.68566463  0.09726232  0.40792158 -0.9294957  1.7453389
## 5   0.44943698 -0.01549706  0.11273724 -1.1970482  1.4225227
## 6   0.40701764 -0.43910434  0.05973105 -1.2470546  1.3665167
## 7   0.28473137  1.01620919 -0.09307491 -1.3946389  1.2084890
## 8  -0.62430939 -1.65803631 -1.22899024 -2.6474685  0.1894880
## 9   0.83965601  0.44131851  0.60034549 -0.7646602  1.9653511
## 10  0.31127919  0.53839578 -0.05990139 -1.3621644  1.2423616
```

```r
ggplot(data=df, aes(x=x, y=y)) +
  geom_point(size=2, shape=1) +
  geom_ribbon(data=df_conf,aes(ymin=lwr,ymax=upr),alpha=0.3) +
  geom_ribbon(data=df_pred,aes(ymin=lwr,ymax=upr),alpha=0.3) +
  geom_abline(slope = m1$coefficients[2], intercept = m1$coefficients[1], colour="red") +
  geom_abline(slope = confint(m1)[2,1], intercept = m1$coefficients[1], colour="blue") +
  geom_abline(slope = confint(m1)[2,2], intercept = m1$coefficients[1], colour="blue") +
  annotate("text", x = -.25, y = 0.25, label = "Prediction Interval of Y's") +
  annotate("text", x = .5, y = -.45, label = "Confidence Interval of Y's") +
  annotate("text", x = -0, y = -1.35, label = "Confidence Interval of Beta 1 Coef.") +
  ggtitle("Interval Estimates with Shaded Regions for Confidence, Prediction")
```




```
predict_int <- predict(m1, interval="prediction")
conf_int <- predict(m1, interval="confidence")
#Prediction Interval
new_df <- cbind(df,predict_int, conf_int)
colnames(new_df) <- make.unique(names(new_df))
ggplot(new_df, aes(x=x, y=y))+
  geom_point() +
  geom_abline(slope = confint(m1)[2,1], intercept = m1$coefficients[1], colour="green") +
  geom_abline(slope = confint(m1)[2,2], intercept = m1$coefficients[1], colour="green") +
    geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
    geom_line(aes(y=upr), color = "red", linetype = "dashed")+
    geom_line(aes(y=lwr.1), color = "blue", linetype = "dashed")+
    geom_line(aes(y=upr.1), color = "blue", linetype = "dashed")+
  geom_smooth(method="lm") +
annotate("text", x = -.25, y = 0.25, label = "Prediction Interval of Y's") +
  annotate("text", x = .5, y = -.45, label = "Confidence Interval of Y's") +
  annotate("text", x = -0, y = -1.35, label = "Confidence Interval of Beta 1 Coef.") +
  ggtitle("Interval Estimates with Shaded Regions for Confidence Only")
```




```
#95% Confidence Interval
```

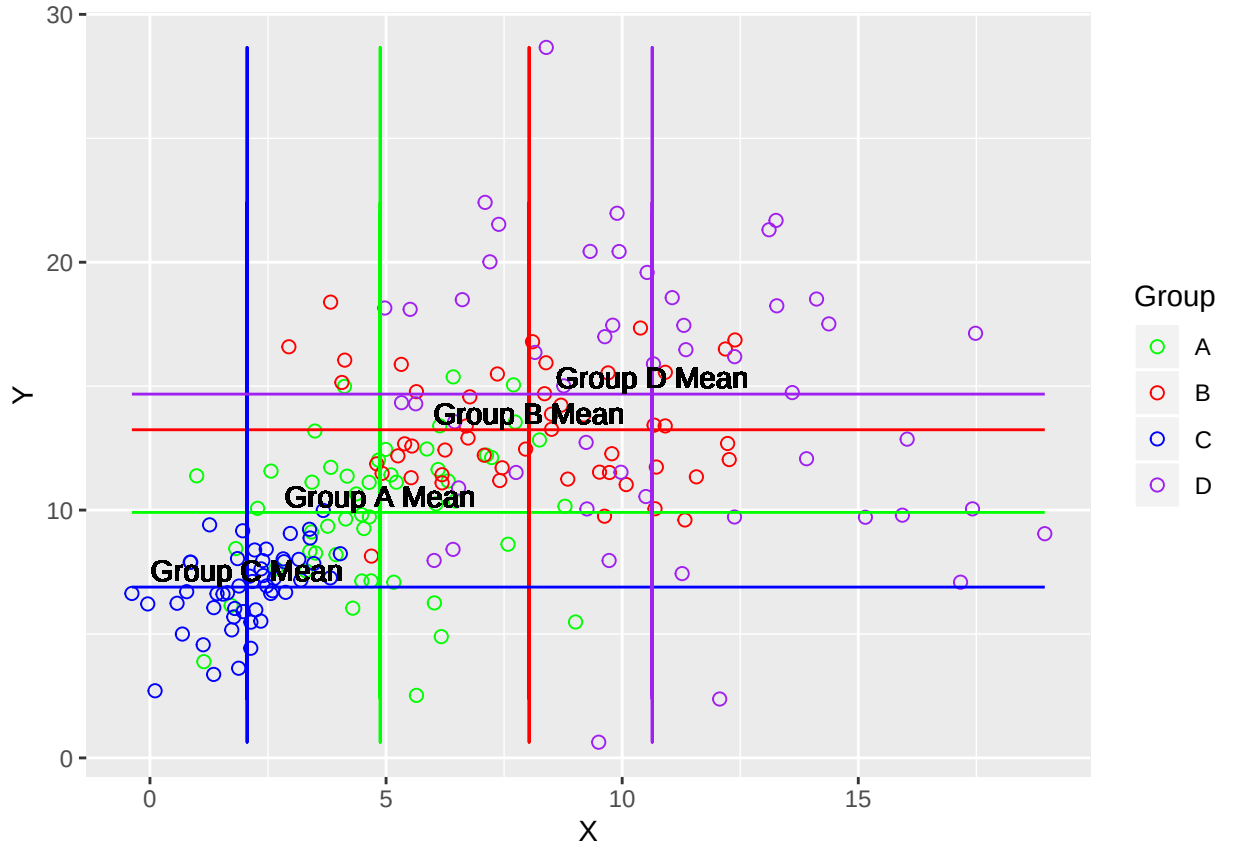
2. Ecological Fallacy refers to a situation where one draws inferences on the individual level from data that was collected at the group level.

```
# 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.

```
library(ggplot2)
library(plyr)

mean_table <- ddply(corrExerciseB, .(Z), summarize,  X_mean =mean(X), Y_mean =mean(Y))

a_x_mean <- mean_table$X_mean[1]
a_y_mean <- mean_table$Y_mean[1]
b_x_mean <- mean_table$X_mean[2]
b_y_mean <- mean_table$Y_mean[2]
c_x_mean <- mean_table$X_mean[3]
c_y_mean <- mean_table$Y_mean[3]
d_x_mean <- mean_table$X_mean[4]
```

```
d_y_mean <- mean_table$Y_mean[4]

ggplot(data=corrExerciseB, aes(x=X, y=Y, color=Z)) +
  geom_point(size=2, shape=1) +
  scale_color_manual(values=c("green", "red", "blue", "purple")) +
  geom_line(aes(x=a_x_mean), colour="green") +
  geom_line(aes(y=a_y_mean), colour="green") +
  geom_line(aes(x=b_x_mean), colour="red") +
  geom_line(aes(y=b_y_mean), colour="red") +
  geom_line(aes(x=c_x_mean), colour="blue") +
  geom_line(aes(y=c_y_mean), colour="blue") +
  geom_line(aes(x=d_x_mean), colour="purple") +
  geom_line(aes(y=d_y_mean), colour="purple") +
  geom_text(x=a_x_mean, y=a_y_mean+.66, label="Group A Mean", colour="black") +
  geom_text(x=b_x_mean, y=b_y_mean+.66, label="Group B Mean", colour="black") +
  geom_text(x=c_x_mean, y=c_y_mean+.66, label="Group C Mean", colour="black") +
  geom_text(x=d_x_mean, y=d_y_mean+.66, label="Group D Mean", colour="black") +
  guides(color=guide_legend(title="Group"))
```



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?

```r
library(plyr)

# Group Level Correlations

cor_func <- function(corrExerciseB)
{
return(data.frame(COR = cor(corrExerciseB$X, corrExerciseB$Y)))
}

ddply(corrExerciseB, .(Z), cor_func)
```

```
##   Z         COR
## 1 A  0.24205633
## 2 B -0.08753011
## 3 C  0.49426902
## 4 D -0.19326301
```

```r
# Validate Group Level Correlations

group_a_cor <- corrExerciseB[which(corrExerciseB$Z=='A'),]
cor(group_a_cor$X, group_a_cor$Y)
```

```
## [1] 0.2420563
```

```r
group_b_cor <- corrExerciseB[which(corrExerciseB$Z=='B'),]
cor(group_b_cor$X, group_b_cor$Y)
```

```
## [1] -0.08753011
```

```r
group_c_cor <- corrExerciseB[which(corrExerciseB$Z=='C'),]
cor(group_c_cor$X, group_c_cor$Y)
```

```
## [1] 0.494269
```

```r
group_d_cor <- corrExerciseB[which(corrExerciseB$Z=='D'),]
cor(group_d_cor$X, group_d_cor$Y)
```

```
## [1] -0.193263
```

```r
# Total Data Correlation Using Means
cor(mean_table$X, mean_table$Y)
```

```
## [1] 0.9921153
```

This appears to be an example of Simpson's paradox, also known as the Yule-Simpson effect, which *'occurs when the marginal association between two categorical variables is qualitatively different from the partial association between the same two variables after controlling for one or more other variables.'*

To directly answer the question at hand, and to provide personal perspective beyond regurgitating a definition—as relevant as it may be—the correlations at the group level may exceed those at an individual

level when there are large differences between the number of observations available across the groups being considered, in addition to differences between observations within the same group.

Though this is common, I would caution against interpreting this as 'Defcon 1' for statisticans, because there are a host of factors to consider. Before dampening the argument though, I will point out that improving data literacy (including knowledge of fallacies, paradoxes, and other phenomenon) would go a long way to mitigating risks, and improving data literacy in targetted efforts (or perhaps to say group-level educational efforts) would certainly improve things.

Nonetheless, I would dampen the heart-grabbing, table-clearing reaction one may have had by noting an ecological fallacy ultimately causes harm when it isn't considered, or when a decision is made without vetting the potential for an ecological fallacy to have occured. Additional dimensions to this point may be noted, particularly the application and field of study for the data in question.

That being said, this is one of many concepts and observations to consider–such as correlation vs. causation, extrapolation, robustness of results, and countably many others—when handling data. At the very least, it would do good to prioritize this on the list of things to know. I'd recommend putting it above the Borel–Kolmogorov Paradox and the Banach–Tarski Paradox.

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

The consequence of committing an ecological fallacy is having inappropriate conclusions.

To expand on this point: If one uses the total correlation of X on Y (positive) to infer what happens in a particular group's value of Y when X increases, one would be led to inappropriate conclusions. For one: Inferences of groups 'B' and 'D', whose correlation value was negative, would likely assume the incorrect relationship by signage. By contrast, an inference of groups 'A' and 'C' would also be incorrect, though due to magnitude instead of signage.

Akin to the observation of 'statistical significance' and 'magnitude significance', committing an ecological fallacy makes for inappropriate conclusions, both in signage and magnitude.