Problem Set 3

Ayush Pundyavana

```r
knitr::opts_chunk$set(
echo = TRUE,
results = 'markup',
tidy = TRUE,
comment = NA,
width = 60, # wrap R output lines
max.print = 100 # limit huge outputs
)
options(width = 60) # ensures printed output wraps too
```

```r
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.4.3
```

```r
library(dplyr)
```

```
Warning: package 'dplyr' was built under R version 4.4.3


Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

Question 1

```r
data1.1 <- read.csv("grades_and_temps.csv")
```

#1.1 Construct a variable that you will use throughout this question: the average yearly temperature in a given year, expressed in Farenheit degrees. Call it avg temp f. Generate a scatter plot with the average yearly temperature (avg temp f) on the x-axis and the average math score (math score) on the y-axis. Using visual inspection, do these variables seem to be positively correlated, negatively correlated, or not correlated at all?

```r
# Adding variable to df
data1.2 <- data1.1 %>% mutate(avg_temp_f = (avg_temp * 9/5) + 32)

#Getting rid of missing values
colSums(is.na(data1.2))                    #Finding where missing values lie
```

```
      cname         year   read_score   math_score
          0            0            1            0
  sci_score     avg_temp        gdppc income_group
          0            2            7            0
 avg_temp_f
          2
```

```r
data1.2 %>% filter(!complete.cases(.))      #show observations w/ missing values
```

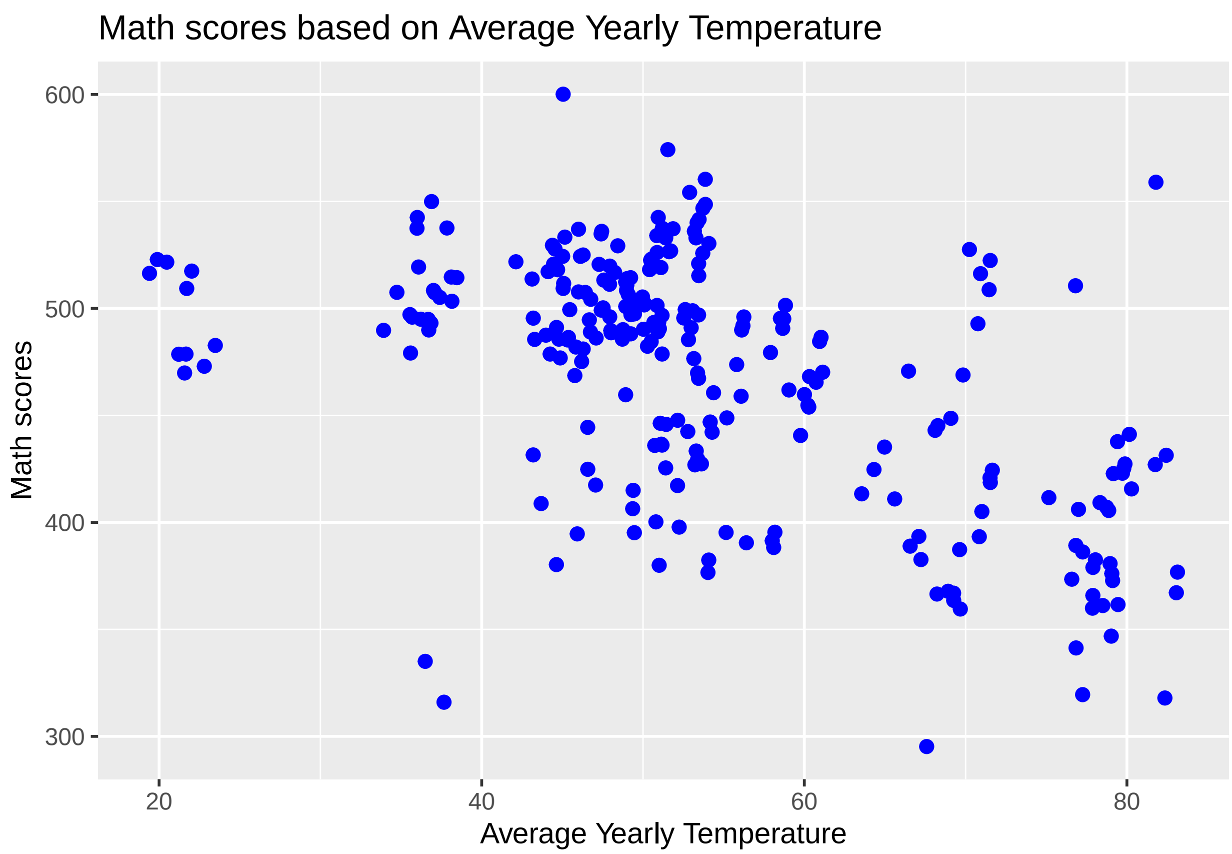
```
          cname year read_score math_score sci_score
1 Liechtenstein 2012   518.3541   537.8998  525.5432
2   New Zealand 2000   527.9097   537.4211  530.5949
3   New Zealand 2003   523.7341   526.1274  523.8999
4   New Zealand 2006   523.3495   523.0433  532.3081
5   New Zealand 2009   523.2422   522.5285  534.5719
6   New Zealand 2012   514.0916   501.3907  517.7480
7     Singapore 2012   537.3722   568.3597  546.6971
8 United States 2006         NA   475.1775  488.2919
   avg_temp    gdppc income_group avg_temp_f
1        NA       NA  high_income         NA
2 10.666180       NA  high_income   51.19912
3 10.484893       NA  high_income   50.87281
4 10.289181       NA  high_income   50.52053
5 10.262914       NA  high_income   50.47325
6 10.483628       NA  high_income   50.87053
7        NA       NA  high_income         NA
8  7.886481 45052.92  high_income   46.19567
```

```r
#Get rid of rows missing avg temp
data1.3 <- data1.2 %>% filter(!is.na(data1.2$avg_temp))
prev_rows <- nrow(data1.2)
curr_rows <- nrow(data1.3)

cat("\n\nNumber of rows excluded (due to missing values): ", prev_rows-curr_rows)
```

```
Number of rows excluded (due to missing values):  2
```

```r
#Display the scatterplot
ggplot(data = data1.3, aes(x = avg_temp_f, y = math_score)) +
  geom_point(color = "blue", size = 2) +
  labs(
    title = "Math scores based on Average Yearly Temperature",
    x = "Average Yearly Temperature",
    y = "Math scores"
  )
```




```r
#Conclusion
cat('Using visual inspection of the scatterplot, there is a roughly downward-sloping
↪ trend,\nmeaning variables seem to be negatively correlated.')
```

```
Using visual inspection of the scatterplot, there is a roughly downward-sloping trend,
meaning variables seem to be negatively correlated.
```

#1.2 Regress the average math score on the average yearly temperature. Report the estimated intercept ($\hat{\alpha}$) and the estimated slope ($\hat{\beta}$). Interpret both coefficients. Does it make sense to interpret $\hat{\alpha}$?

```r
reg_model <- lm(math_score ~ avg_temp_f, data = data1.3)
summary(reg_model)
```

```
Call:
lm(formula = math_score ~ avg_temp_f, data = data1.3)

Residuals:
     Min       1Q   Median       3Q      Max 
-188.712  -30.505    6.058   30.289  153.653 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 589.4595    11.9071   49.51   <2e-16 ***
avg_temp_f   -2.2511     0.2133  -10.55   <2e-16 ***
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.93 on 261 degrees of freedom
Multiple R-squared:  0.2992,    Adjusted R-squared:  0.2965 
F-statistic: 111.4 on 1 and 261 DF,  p-value: < 2.2e-16
```

```r
alpha_hat <- summary(reg_model)$coefficients[1]
beta_hat <- summary(reg_model)$coefficients[2]

alpha_hat_SE <- round(summary(reg_model)$coefficients[3], 3)
beta_hat_SE <- round(summary(reg_model)$coefficients[4], 3)


cat("----------------------------------------------------------------\nAlpha hat:",
↪  alpha_hat," --> According to the regression model, this is the math score when
↪  \ntemperature is 0, or in other words, the y-intercept.\n
Beta hat:", beta_hat," --> According to the model, for every 1 degree Fahrenheit
↪  increase in \ntemperature, the math score decreases by this much\n
It does not make sense to interpret alpha hat because the temperature of 0 is outside of
↪  the \nrange of observed temperatures.")
```

```
----------------------------------------------------------------
Alpha hat: 589.4595  --> According to the regression model, this is the math score when 
temperature is 0, or in other words, the y-intercept.

Beta hat: -2.251118  --> According to the model, for every 1 degree Fahrenheit increase in 
temperature, the math score decreases by this much

It does not make sense to interpret alpha hat because the temperature of 0 is outside of the 
range of observed temperatures.
```

#1.3 Based on visual inspection of the plot in question 1, do you think the errors are homoskedastic or heteroskedastic? Why? Compare the standard errors for $\hat{\alpha}$ and $\hat{\beta}$ both under the homoskedasticity and the heteroskedasticty asssumptions.

```r
cat("I think the errors are heteroskedastic because from the visualization, it seems
↪  that the \nvariance of the error seems to change (increase) as x increases\n\n")
```

```
I think the errors are heteroskedastic because from the visualization, it seems that the
variance of the error seems to change (increase) as x increases
```

```r
#install.packages("sandwich")
library(sandwich)
```

```
Warning: package 'sandwich' was built under R version 4.4.3
```

```r
#install.packages("lmtest")
library(lmtest)
```

```
Warning: package 'lmtest' was built under R version 4.4.3

Loading required package: zoo

Warning: package 'zoo' was built under R version 4.4.3


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```r
vcov <- vcovHC(reg_model, type = "HC3")
robust_se <- sqrt(diag(vcov))
robust_se
```

```
(Intercept)  avg_temp_f
 12.6112888   0.2304818
```

```r
coeftest(reg_model, vcov. = vcov)
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 589.45951   12.61129  46.741 < 2.2e-16 ***
avg_temp_f   -2.25112    0.23048  -9.767 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("\n\nErrors under homoskedastic assumptions-   hat:", round(alpha_hat_SE,3),  " beta
 ↪  hat:", round(beta_hat_SE, 3),
"\nErrors under heteroskedastic assumptions- alpha hat:", round(robust_se[1],3), " beta
 ↪  hat:", round(robust_se[2],3),
"\n\nThe SE's of alpha hat and beta hat are both larger under heteroskedastic
 ↪  assumptions-\n alpha hat:", robust_se[1]-alpha_hat_SE, "greater    beta hat:",
 ↪  robust_se[2]-beta_hat_SE, "greater")
```

```
Errors under homoskedastic assumptions-   hat: 11.907  beta hat: 0.213
Errors under heteroskedastic assumptions- alpha hat: 12.611  beta hat: 0.23

The SE's of alpha hat and beta hat are both larger under heteroskedastic assumptions-
 alpha hat: 0.7042888 greater    beta hat: 0.01748176 greater
```

#1.4 Using the heteroskedasticity-robust standard errors: (a) test the null hypothesis that H0 : $\beta = 0$ at the 5% significance level, and (b) test the null hypothesis that H0 : $\beta = -1.85$ at both the 5% and the 10% significance levels. Write out the t-statistic formulas to perform these two-sided tests.

```r
#Part A

#H0 :   = 0  (5% significance level)
cat("Formula for t-stat: ", "t = (Betahat - Beta_null) / SE(Betahat)\n\n")
```

```
Formula for t-stat:  t = (Betahat - Beta_null) / SE(Betahat)
```

```r
t_statA <- (beta_hat - 0) / robust_se[2] #Finding t-statistic
df <- nrow(data1.3) - length(coef(reg_model))
p_valueA <- 2*pt(-abs(t_statA), df = df)

cat("Since the p-value (", p_valueA, ") is less than alpha (0.05), we reject the null
 ↪  hypothesis \nthat the true coefficient is 0 with 95% confidence.")
```

```
Since the p-value ( 2.087355e-19 ) is less than alpha (0.05), we reject the null hypothesis
that the true coefficient is 0 with 95% confidence.
```

```r
#Part B

#H0: Beta = -1.85  (5%, 10% significance level)
cat("Formula for t-stat: ", "t = (Betahat - Beta_null) / SE(Betahat)\n\n")
```

```
Formula for t-stat:  t = (Betahat - Beta_null) / SE(Betahat)
```

```r
t_statB <- (beta_hat + (1.85)) / robust_se[2]
p_valueB <- 2*pt(-abs(t_statB), df = df)


cat("\n05% significance level:  Since the p-value (", round(p_valueB, 3), ") is greater
↪ than alpha (0.05), we fail \nto reject the null hypothesis that the true coefficient
↪ is -1.85 with 95% confidence",
    "\n\n10% significance level:  Since the p-value (", round(p_valueB,3), ") is less
     ↪ than alpha (0.10), we reject \nthe null hypothesis that the true coefficient
     ↪ \nis -1.85 with 90% confidence")
```

```
05% significance level:  Since the p-value ( 0.083 ) is greater than alpha (0.05), we fail
to reject the null hypothesis that the true coefficient is -1.85 with 95% confidence

10% significance level:  Since the p-value ( 0.083 ) is less than alpha (0.10), we reject
the null hypothesis that the true coefficient
is -1.85 with 90% confidence
```

#1.5 Suppose the OECD is seriously considering implementing stricter climate change agreements that are projected to decrease global average yearly temperatures by 2.35◦F. Using the econometric model in Equation 1 and your estimated coefficients, compute by how much you would expect the average math score to change as a result of this projected temperature decrease.

```r
change <- beta_hat * -2.35

cat("Using the econometric model in Equation 1 and estimated coefficients, I would
↪ expect an \nincrease in average math score by", change, "as a \nresult of the
↪ projected temperature decrease.")
```

```
Using the econometric model in Equation 1 and estimated coefficients, I would expect an
increase in average math score by 5.290126 as a
result of the projected temperature decrease.
```

#1.6 Do you think your estimate $\hat{\beta}$ is causal (e.g., does the answer in the previous part make sense to you)? Explain your answer.

```r
cat("No, I don't think beta hat is not causal because there are likely omitted variables
↪  \nsuch as GDP per capita, education quality, and regional development \nthat affect
↪  both temperature and math scores.

Therefore, I believe the observed relationship reflects correlation, not causation.")
```

```
No, I don't think beta hat is not causal because there are likely omitted variables
such as GDP per capita, education quality, and regional development
that affect both temperature and math scores.

Therefore, I believe the observed relationship reflects correlation, not causation.
```

##Question 2

```r
data2.1 <- read.csv("middlesex_permits.csv")

head(data2.1)
```

```
  record_id municipality_name construction_cost units  fees
1  89042692          CARTERET           3563225    54 87800
2  10000018          CRANBURY            150500     1  4407
3  10000019          CRANBURY            150500     1  4332
4  10000046          CRANBURY            150500     1  5329
5  10000049          CRANBURY            150500     1  3231
6  10000053          CRANBURY            150501     1  5429
  square_feet volume
1       15117 755503
2        3293  62859
3        3293  62859
4        3106  88521
5        2272  42789
6        3106  88521
```

```r
colSums(is.na(data2.1))
```

```
        record_id municipality_name construction_cost
                0                 0                 0
            units              fees       square_feet
                0                 0                 0
           volume
                0
```

#2.1 Estimate Equation 2. Report your estimate for ˆβ1 and its respective heteroskedasticity- robust standard error. Interpret the coefficient.

```r
reg_model_cc_f <- lm(construction_cost ~ fees, data = data2.1)
vcov_cc_f <- vcovHC(reg_model_cc_f, type = "HC3")
coeftest(reg_model_cc_f, vcov. = vcov_cc_f)
```

```
t test of coefficients:

              Estimate  Std. Error t value Pr(>|t|)
(Intercept) -140521.927   61734.642 -2.2762   0.0229 *
fees             73.025      10.141  7.2008 7.46e-13 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
robust_se_2 <- sqrt(diag(vcov_cc_f))

beta1hat <- (reg_model_cc_f$coefficients[2])
beta1hatSE <- robust_se_2[2]

cat("\nBeta1 hat is estimated to be", round(beta1hat,3), "and its respective
 ↪  heteroskedasticity-robust \nstandard error is estimated to be",
 ↪  round(beta1hatSE,3),"\n
Interpret the coefficient: For every 1 dollar increase in the total sum charged for
 ↪  \nthe building, the construction cost is expected to go up by\n", round(beta1hat,3),
 ↪  "dollars")
```

```
Beta1 hat is estimated to be 73.025 and its respective heteroskedasticity-robust
standard error is estimated to be 10.141

Interpret the coefficient: For every 1 dollar increase in the total sum charged for
the building, the construction cost is expected to go up by
 73.025 dollars
```

#2.2 Since you took Econ 322, you suspect that Equation 2 may suffer from omitted vari- able bias (OVB). But you also have data on other permit characteristics, so you can investigate whether OVB is a concern! You start by exploring whether the number of units in the building can be a source of OVB. Compute the correlation between units and fees. Compute the corre- lation between units and construction cost. Based on these results, do you think your estimate of ˆβ1 is biased? If so, is it upward or downward biased? Provide an intuitive explanation.

```r
#Correlation between units and fees
corr_u_f <- cor(data2.1$units, data2.1$fees, use = "complete")

#Correlation between units and construction
corr_u_cc <- cor(data2.1$units, data2.1$construction_cost, use = "complete")
```

```
cat("Based on these results, I do think my estimate of Beta1 hat is biased since \nboth
↪  correlation between units and fees(X) (", round(corr_u_f,3), ") AND \ncorrelation
↪  between units and construction(Y) (", round(corr_u_cc,3), ") are strongly
↪  \ncorrelated. Since both correlations are positive, the \nbias must be upward
↪  biased.

Intuitive Explanation: The estimate of the coeff. on fees is upward biased because
↪  \nlarger buildings usually have more unit, which leads to both higher fees and
↪  \nhigher construction costs. Since the number of units isn't included in the model,
↪  \npart of the effect of building size is being incorrectly attributed to fees,
↪  \nmaking the estimated impact of fees on construction cost appear larger \nthan it
↪  truly is.")
```

```
Based on these results, I do think my estimate of Beta1 hat is biased since
both correlation between units and fees(X) ( 0.711 ) AND
correlation between units and construction(Y) ( 0.675 ) are strongly
correlated. Since both correlations are positive, the
bias must be upward biased.

Intuitive Explanation: The estimate of the coeff. on fees is upward biased because
larger buildings usually have more unit, which leads to both higher fees and
higher construction costs. Since the number of units isn't included in the model,
part of the effect of building size is being incorrectly attributed to fees,
making the estimated impact of fees on construction cost appear larger
than it truly is.
```

#2.3 Report the estimated ˆβ2 and ˆθ2 coefficients, along with their respective heteroskedasticity- robust standard errors. Interpret the coefficients. How does ˆβ2 compare with your estimate ˆβ1 from Equation 2? Relate the answer to this question to your answer in the previous part.

```
reg_model_cc_f_u <- lm(construction_cost ~ fees + units, data = data2.1)
vcov_cc_f_u <- vcovHC(reg_model_cc_f_u, type = "HC3")
coeftest(reg_model_cc_f_u, vcov. = vcov_cc_f_u)
```

```
t test of coefficients:

              Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -1.2876e+05  5.4434e+04 -2.3654   0.01807 *
fees         6.3098e+01  9.5513e+00  6.6062 4.615e-11 ***
units        2.0261e+04  2.0175e+04  1.0043   0.31533
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
bet_hat <- reg_model_cc_f_u$coefficients[2]
thet_hat <- reg_model_cc_f_u$coefficients[3]

cat("\nBeta2 hat: For every $1 increase in fees, construction cost increases by  $",
 ↪  round(bet_hat,3), "
Theta2 hat: For every increase in units in the builiding by 1, construction cost
 ↪  \nincreases by  $", round(thet_hat,3), "

Beta2 hat is less than Beta1 hat which makes sense because as I stated in the previous
 ↪  \nproblem, part of the effect of the building was incorrectly being attributed \nto
 ↪  fees, making Beta1 hat appear larger than the true value. \nSo, now accounting
 ↪  number of units, the new effect of the fees charged for \nthe building (Beta2 hat)
 ↪  is lower, and closer to the true value population value.")
```

```
Beta2 hat: For every $1 increase in fees, construction cost increases by  $ 63.098
Theta2 hat: For every increase in units in the builiding by 1, construction cost
increases by  $ 20260.85

Beta2 hat is less than Beta1 hat which makes sense because as I stated in the previous
problem, part of the effect of the building was incorrectly being attributed
to fees, making Beta1 hat appear larger than the true value.
So, now accounting number of units, the new effect of the fees charged for
the building (Beta2 hat) is lower, and closer to the true value population value.
```

#2.4 Note that you will need to construct two new variables from the variable municipality name: new brunswick is a dummy variable that takes value 1 if the permit was issued in New Brunswick, 0 otherwise; and edison is a dummy variable that takes value 1 if the per- mit was issued in Edison, 0 otherwise. Report all the estimated coefficients in this regres- sion, along with their heteroskedastic-robust standard errors. How does ˆβ3 compare with your estimate ˆβ2 from Equation 3? Do you think that your estimate of β3 is causal? Ex- plain.

```r
#Making necessary additions to the data
table(data2.1$municipality_name)
```

```
       CARTERET         CRANBURY         DUNELLEN
              1              182               92
  EAST BRUNSWICK           EDISON         HELMETTA
            116              287                4
   HIGHLAND PARK        JAMESBURG         METUCHEN
             15                4               54
      MIDDLESEX         MILLTOWN       MONROE TWP
             16                1             1047
   NEW BRUNSWICK  NORTH BRUNSWICK       OLD BRIDGE
             56               43              478
```

```
      PERTH AMBOY       PISCATAWAY       PLAINSBORO
               52              118               67
       SAYREVILLE      SOUTH AMBOY   SOUTH BRUNSWICK
               98               27              165
  SOUTH PLAINFIELD     SOUTH RIVER        SPOTSWOOD
               40               18                5
       WOODBRIDGE
              150
```

```r
data2.2 <- data2.1 %>% mutate(new_brunswick = as.numeric(municipality_name=="NEW
↪  BRUNSWICK")) %>% mutate(edison = as.numeric(municipality_name=="EDISON"))

head(data2.2)
```

```
  record_id municipality_name construction_cost units  fees
1  89042692          CARTERET           3563225    54 87800
2  10000018          CRANBURY            150500     1  4407
3  10000019          CRANBURY            150500     1  4332
4  10000046          CRANBURY            150500     1  5329
5  10000049          CRANBURY            150500     1  3231
6  10000053          CRANBURY            150501     1  5429
  square_feet volume new_brunswick edison
1       15117 755503             0      0
2        3293  62859             0      0
3        3293  62859             0      0
4        3106  88521             0      0
5        2272  42789             0      0
6        3106  88521             0      0
```

```r
#Check accuracy
nrow(data2.2[data2.2$new_brunswick==1,])
```

```
[1] 56
```

```r
nrow(data2.2[data2.2$edison==1,])
```

```
[1] 287
```

```r
#Running the regression
reg_model_cc_many <- lm(construction_cost ~ fees + units + square_feet + volume +
↪  new_brunswick + edison, data = data2.2)
vcov_cc_many <- vcovHC(reg_model_cc_many, type = "HC3")
coeftest(reg_model_cc_many, vcov. = vcov_cc_many)
```

```
t test of coefficients:

                 Estimate  Std. Error t value Pr(>|t|)
(Intercept)   -1.3756e+05  6.7164e+04 -2.0482  0.04063 *
fees           4.8150e+01  2.5510e+01  1.8875  0.05918 .
units          1.7471e+04  1.9791e+04  0.8828  0.37741
square_feet    7.1103e-01  7.2355e-01  0.9827  0.32584
volume         1.4664e+00  2.3410e+00  0.6264  0.53111
new_brunswick  4.7630e+04  2.0438e+05  0.2331  0.81574
edison         3.2444e+03  5.4092e+04  0.0600  0.95218
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
fee_coeff <- reg_model_cc_many$coefficients[2]
uni_coeff <- reg_model_cc_many$coefficients[3]
sqf_coeff <- reg_model_cc_many$coefficients[4]
vol_ceoff <- reg_model_cc_many$coefficients[5]
nb_coeff <-  reg_model_cc_many$coefficients[6]
edi_coeff <- reg_model_cc_many$coefficients[7]

robust_se_many <- sqrt(diag(vcov_cc_many))
fee_SE <- robust_se_many[2]
uni_SE <- robust_se_many[3]
sqf_SE <- robust_se_many[4]
vol_SE <- robust_se_many[5]
nb_SE  <- robust_se_many[6]
edi_SE <- robust_se_many[7]

#Comparing
diff <- round(bet_hat,3) - round(fee_coeff,3)
cat("Beta3 hat is less than Beta2 hat by", diff, "which can be attributed to the
↪  inclusion of \nkappa3, upsilon3, and phi3\n\n")
```

```
Beta3 hat is less than Beta2 hat by 14.948 which can be attributed to the inclusion of
kappa3, upsilon3, and phi3
```

```r
#Checking for causality - Beta3 hat > |1.96*SE|
signif <- fee_coeff > abs(1.96 * round(fee_SE,3))

if (signif) {
  cat("Beta3 hat is significant because Beta3 hat > |1.96*SE|")
} else {
  cat("Beta3 hat is not significant because Beta3 hat < |1.96*SE|")
  cat("\nEven though Beta3 captures a conditional relationship after controlling for
  ↪  size and \nmunicipality, it is not necessarily causal because unobserved \nfactors
  ↪  (e.g., project complexity, land cost) may still bias the estimate.")

}
```

```
Beta3 hat is not significant because Beta3 hat < |1.96*SE|
Even though Beta3 captures a conditional relationship after controlling for size and
municipality, it is not necessarily causal because unobserved
factors (e.g., project complexity, land cost) may still bias the estimate.
```

#2.5 In the previous regression, (i) How do you interpret α3? Does this interpretation make sense? and (ii) How do you interpret the coefficient on new brunswick? Use your estimates to answer these questions.

```r
cat("alpha3 Interpretation:  The construction cost when fees, units, square feet,
 ↪  \nvolume are 0 and the permit is not issued in New Brunswick or Edison.
\nDoes this make sense:  No, the interpretation does not make sense because at least 1
 ↪  \nof these variables, such as square feet, can not realistically be zero.
\nnew_brunswick coeff. Interpretation: When the permit is issued in New Brunswick, the
 ↪  \nestimated construction cost increases by  $", 4.7630e+04)
```

```
alpha3 Interpretation:  The construction cost when fees, units, square feet,
volume are 0 and the permit is not issued in New Brunswick or Edison.

Does this make sense:  No, the interpretation does not make sense because at least 1
of these variables, such as square feet, can not realistically be zero.

new_brunswick coeff. Interpretation: When the permit is issued in New Brunswick, the
estimated construction cost increases by  $ 47630
```