Problem Set Assignment No. 3

Aria Muchhal

04 March, 2023

```
## Keep this line always
knitr::opts_chunk$set(echo = TRUE,
                      collapse = TRUE,
                      warning = FALSE, message = FALSE,
                      fig.align = 'center')
```

0.

Task A.

```
library(knitr)
library(tinytex)
library(wooldridge)
library(lmtest)
library(sandwich)
library(AER)
library(fixest)
library(modelsummary)
```

1.

Task A.

```
data(CASchools)
table(CASchools$county)
##
##         Alameda          Butte       Calaveras    Contra Costa       El Dorado
##               1              6              1              7             10
##          Fresno          Glenn       Humboldt        Imperial            Inyo
##              12              3             17              6              1
##            Kern          Kings           Lake         Lassen     Los Angeles
##              27              9              2              5             27
##          Madera         Marin       Mendocino         Merced        Monterey
##               5              8              1             11              7
##          Nevada         Orange         Placer      Riverside      Sacramento
##               9             11             11              4              7
##      San Benito  San Bernardino      San Diego     San Joaquin San Luis Obispo
```

```
##            3           10           21            6            2
##    San Mateo Santa Barbara  Santa Clara   Santa Cruz       Shasta
##           17           11           20            7           13
##     Siskiyou       Sonoma   Stanislaus       Sutter       Tehama
##            9           29            7            6            8
##     Trinity       Tulare    Tuolumne       Ventura         Yuba
##            2           24            6            9            2
```

Question A.

a. How many observations are in the data?

420.

b. We are interested in test scores. Which variable(s) in CASchools would be our outcome of interest?

read and math

c. Using the output from the table(. . .) command, what county has the most observations in the data?

Sonoma.

Task B. Data Cleaning

(a)

```
studentTeacherRatio=CASchools$students/CASchools$teachers
CASchools$studentTeacherRatio<-studentTeacherRatio
```

(b)

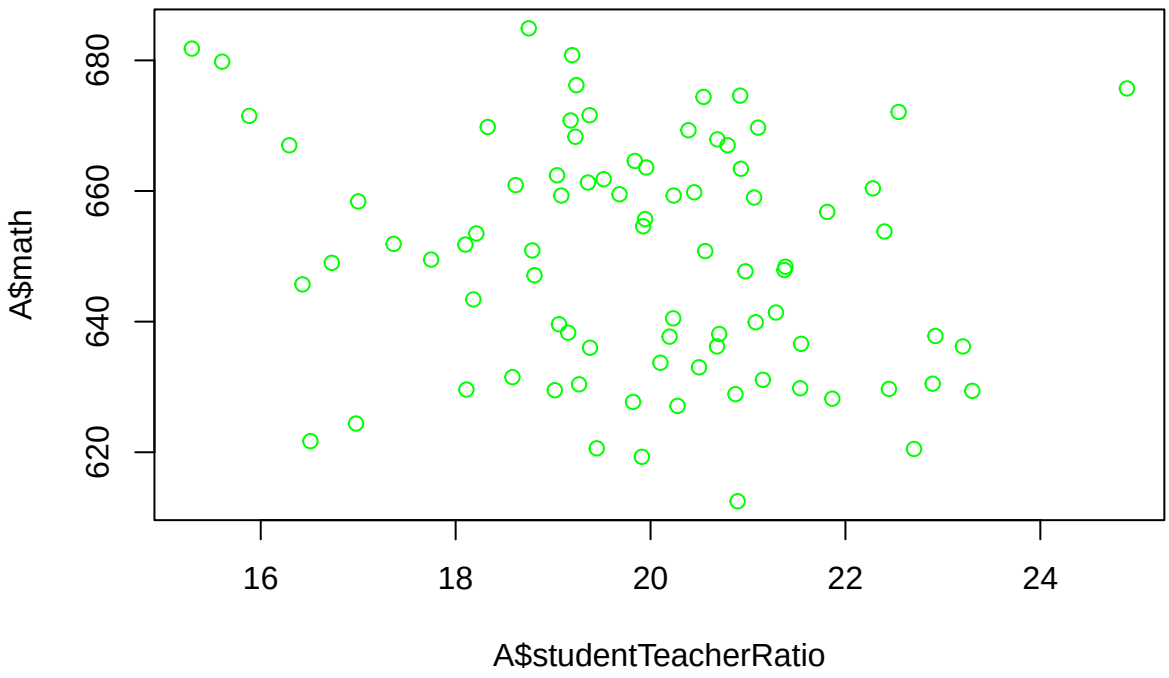
```
bigCounties=CASchools[CASchools$county=='Sonoma'|CASchools$county=='Los Angeles'|CASchools$county=='Ker
```

(c)

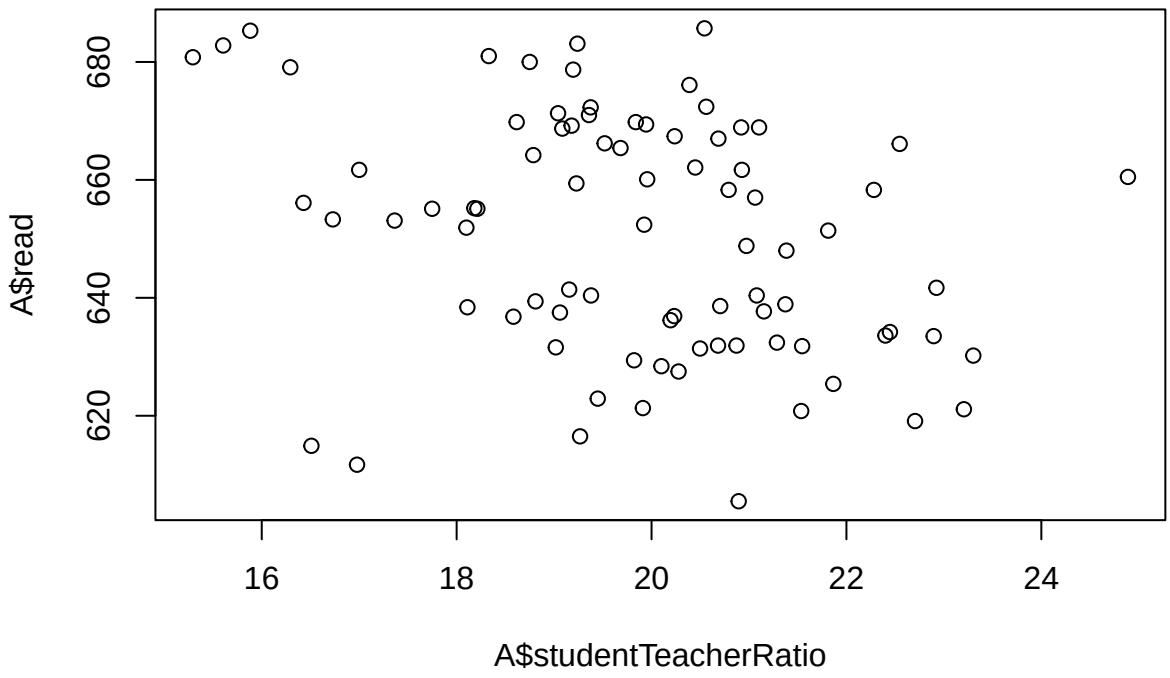
```
A=CASchools[CASchools$county=='Sonoma'|CASchools$county=='Los Angeles'|CASchools$county=='Kern', c('dis
```

(d)

```
plot(A$studentTeacherRatio, A$math, col="green")
```

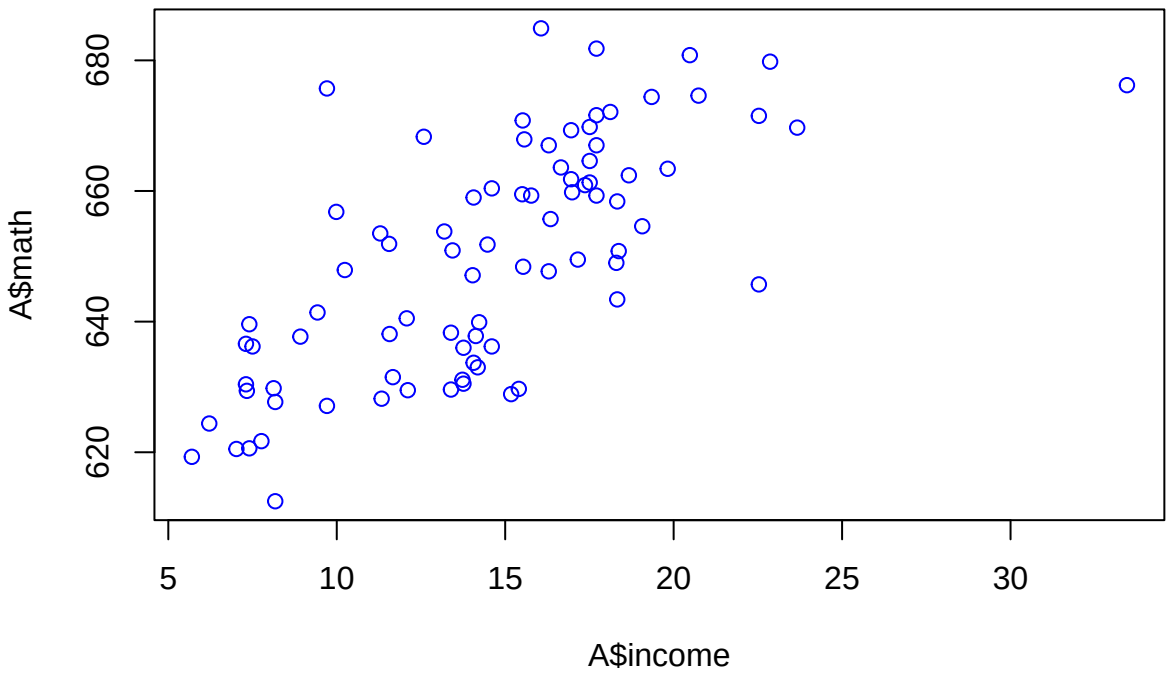



```
plot(A$studentTeacherRatio, A$read)
```

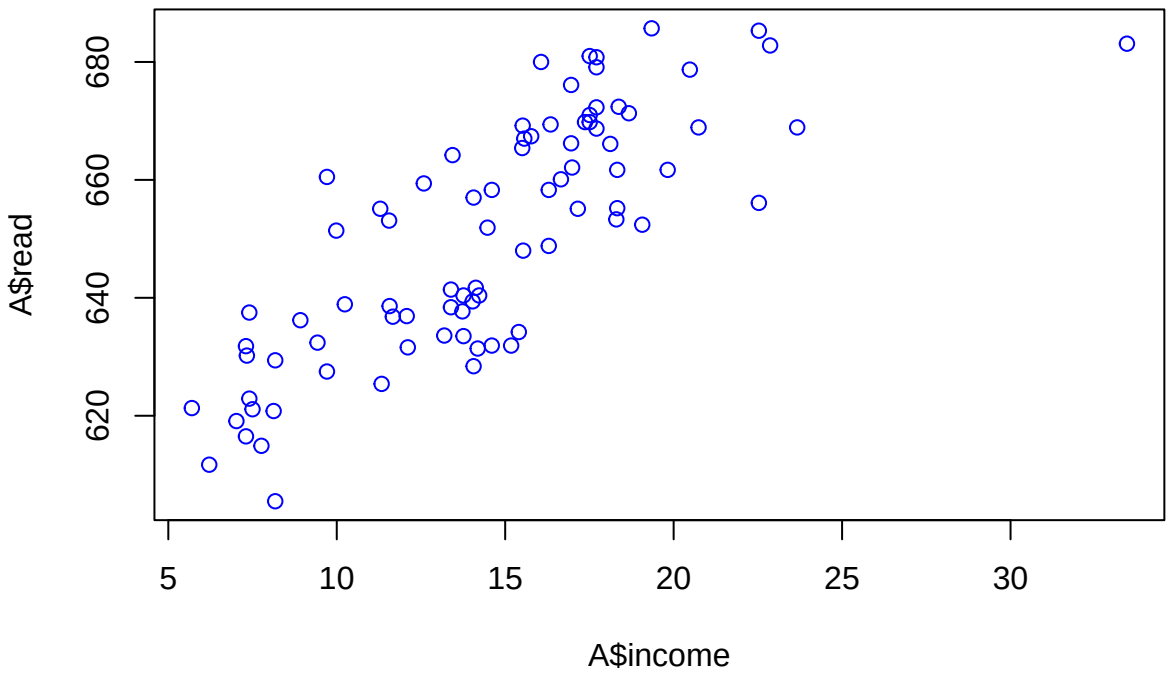


(e)

```
plot(A$income, A$math, col="blue")
```

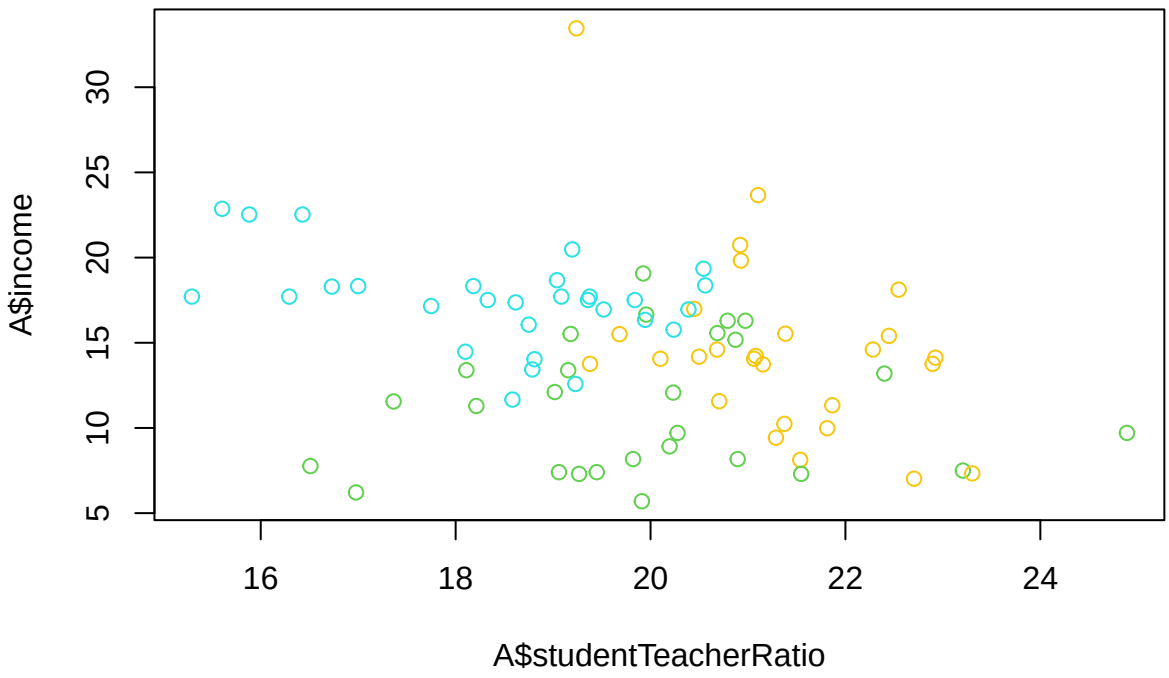



```
plot(A$income, A$read, col="blue")
```



(f)

```
plot(A$studentTeacherRatio, A$income, col=as.factor(A$county))
```



Question B. Plot the relationship

a. Using your first two plots, does there appear to be a relationship between higher student:teacher ratios and math scores? What about reading scores?

One can see a slight negative correlation between math and reading scores and student:teacher ratios. Essentially, higher ratios appear to be associated with lower scores.

b. Using your last two plots, does there appear to be a relationship between higher income and math scores? What about reading scores?

One can see a very strong positive correlation between math and reading scores and income.

c. Using the final plot, does it appear that some counties have higher income or higher student:teacher ratios (or both)?

While some counties are more dispersed than others, it does seem like 1 county has a higher income and lower student to teacher ratio, another has a higher income and higher student to teacher ratio, and the last has a lower income and dispersed student to teacher ratio.

Task C

(a)

```
coeftest(lm(read ~ calworks, A), vcov = vcovHC, "HC1")
##
## z test of coefficients:
```

```
##
##             Estimate Std. Error  z value  Pr(>|z|)
## (Intercept) 667.3168     2.6766 249.3168 < 2.2e-16 ***
## calworks     -1.1971     0.1680  -7.1253 1.038e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question C

a. What is the coefficient on calworks and what does it mean? Note that calworks is in percentage points (you can see the range using range(CASchools$calworks))

-1.1971, so one additional point in a school's reading scores is associated with a 1.1971% decrease in calworks.

b. What potential omitted variables might bias this coefficient? That is, is there something unobserved correlated with calworks that might also be correlated with read?

Calworks would definitely be associated with income and lunch, and it is likely that both would be correlated with read.

Task D

(a)

```
coeftest(lm(read ~ calworks+as.factor(county), A), vcov = vcovHC, "HC1")
##
## z test of coefficients:
##
##                               Estimate Std. Error  z value  Pr(>|z|)
## (Intercept)                  653.11817    4.57937 142.6218 < 2.2e-16 ***
## calworks                      -0.84613    0.18518  -4.5692 4.895e-06 ***
## as.factor(county)Los Angeles   5.99662    4.18989   1.4312    0.1524
## as.factor(county)Sonoma       20.94302    4.31021   4.8589 1.180e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question D

a. What is the new coefficient on calworks? Is it larger or smaller?

4.895e-06+2.2e-16=4.895e-6, which is smaller.

b. What is the base county level?

653.11817

c. What is the expected reading score for an observation in Sonoma County, at a schools with a calworks value of 25%?

653.11817+20.94302+(2.2e-16+1.180e-06)*25=674.061219

Task E

(a)

```
uhat1=residuals(lm(read ~ calworks+as.factor(county), A), na.rm=FALSE)
A$uhat1<-uhat1
uhat2=uhat1^2
A$uhat2<-uhat2
```

(b) Run the appropriate regression (see slide 108 of 02-Multivariate Regression) and show the results

```
appropReg<-lm(uhat2 ~ calworks+as.factor(county), A)
summary(appropReg)
##
## Call:
## lm(formula = uhat2 ~ calworks + as.factor(county), data = A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -295.32 -126.37  -54.88   45.30  728.67
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    286.2575    63.9841   4.474 2.55e-05 ***
## calworks                         0.4682     2.5380   0.184   0.8541
## as.factor(county)Los Angeles  -155.7419    62.1401  -2.506   0.0143 *
## as.factor(county)Sonoma       -203.6461    66.7400  -3.051   0.0031 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 227.8 on 79 degrees of freedom
## Multiple R-squared:  0.1375, Adjusted R-squared:  0.1048
## F-statistic: 4.199 on 3 and 79 DF,  p-value: 0.008246
```

Question E

a. What is the relevant output of the regression summary for our Breusch-Pagan Test H0: "No Heteroskedasticity present"

The p-value of the F statistic from this test. If it's small, we reject the H0–homoskedasticity–which would indicate heteroskedasticity.

b. What is your interpretation of the results? Should we be using heteroskedastic errors?

The p-value is small: .008246, so we can reject the H0 and feel comfortable using heteroskedastic errors.

Task F

```
bptest(read ~ calworks+as.factor(county), data=A)
##
##  studentized Breusch-Pagan test
##
```

```
## data:  read ~ calworks + as.factor(county)
## BP = 11.416, df = 3, p-value = 0.009679
```

Question F

a. What is the interpretation of the result from this version of the Breusch-Pagan test? The pvalue is quite low, 0.009679, so we should reject the H0 of "No Heteroskedasticity present" and use HC robust errors.

b. Is it the same (or very close) to our results from Task 1.E? It is not the same, but it is quite close.

Task G

```
coeftest(lm(read ~ calworks+english+calworks*english+as.factor(county), A), vcov = vcovHC, "HC1")
##
## z test of coefficients:
##
##                                Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept)                  663.6423532   3.0262200 219.2975 < 2.2e-16 ***
## calworks                      -0.7711882   0.1786349  -4.3171 1.581e-05 ***
## english                       -0.8269718   0.1362847  -6.0680 1.295e-09 ***
## as.factor(county)Los Angeles   6.3300599   2.3866774   2.6522  0.007996 **
## as.factor(county)Sonoma       17.6514008   2.4963541   7.0709 1.540e-12 ***
## calworks:english               0.0158432   0.0064111   2.4712  0.013466 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question G

a. What is the effect of an increase in calworks by one unit for a school with 0% english learners (english=0)?

663.6423532-0.7711882=662.871165 increase in reading score, with all else held equal.

b.What is the effect of an increase in calworks by one unit for a school with 40% english learners?

663.6423532-0.7711882-(0.8269718*.4)+(0.0158432*.4)=662.546714 increase in reading score, with all else

c.What is the formula you used to determine dRead/dCalworks ? Hint: it includes the variable english. Write it using LaTeX.

$\frac{dRead}{dCalworks} = \beta_0 + \beta_{calworks} * calworks + \beta_{english} * english + \beta_{calworks-and-english} * (calworks - and - english)$

2. Last Question

I spent 9 continuous hours on this problem set, but it took around 2 days with breaks.