Hayden C. Hancock
Dr. Hwu
GBA:6220 Advanced Stats
Due: 2/11/2022

Assignment 1

1. Let kids denote the number of children ever born to a woman and let educ denote years of education for the woman. A simple model relating fertility to years of education is

$$kids = \beta 0 + \beta 1 educ + u$$

   a) The error term "u" represents all other unobserved factors that might affect fertility other than education. We hold "u" as independent of education and affecting fertility. Some examples I would think of holding significance in place of "u" is genetics, age, partner fertility and income level. This is under the assumption that if "u" has correlation with education but holding the equation assumption $E(U|x)=0$.
   b) "ceteris paribus" means "other things being equal or held constant". In this example there will not be a ceteris paribus effect on education and fertility. Since we do not have any other control variables there is not a way to uncover the ceteris paribus effect of education on fertility. A simple regression would tell you the overall effect of education on kids while not controlling anything else.

2. Suppose that average worker productivity at manufacturing firms (avgprod) depends on two factors, average hours of training (avgtrain) and average worker ability(avgabil):

$$avgprod = \beta 0 + \beta 1 avgtrain + \beta 2 avgabil + u$$

   a) Positive b2 and a negative correlation between avgtrain and avgabil would give us a negative bias for b1.

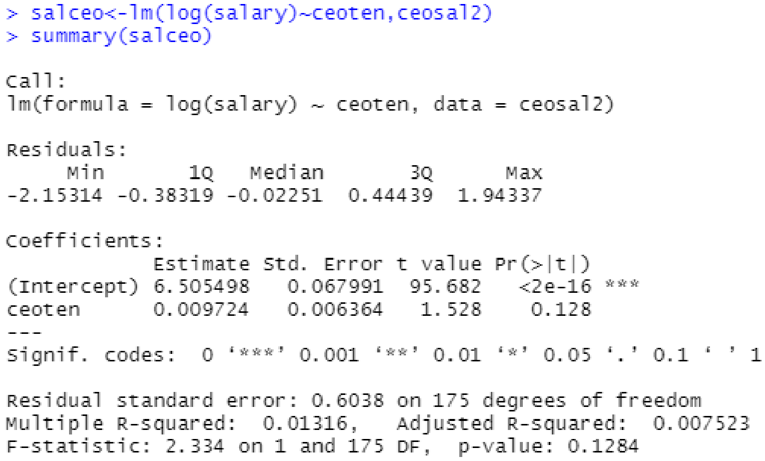
3. The following equation describes the median housing price in a community in terms of amount of pollution(nox for nitrous oxide) and the average number of rooms in houses in the community(rooms):

$$\log(price) = \beta 0 + \beta 1 \log(nox) + \beta 2 rooms + u$$

   a) We expect B1 < 0 because more pollution can be expected to lower housing values. B1 is the elasticity of price with respect to nitrous oxide. B2 is most likely positive because the amount of rooms most likely has a positive slope as more rooms would increase the price of a house.
   b) This example can be characterized as negative bias. If the assumption is that that independently quantity of rooms increases the price of the house & increase of nox decreases the price of a house. This would indicate rooms and nox are negatively correlated. For example, when poorer neighborhoods have more pollution but many rooms, the price would still go down.
   c) As your increase rooms by 1 your would expect an increase in price, and we can infer the correlation between rooms and nox would be negative. A negative times a positive would indicate that nox has a negative bias.

The negative relationship between nox and price is much higher than expected in the simple linear regression, potentially overestimated. In the Second example with nox and rooms, this is more of what I would have expected, balancing the rooms and the presence of nitrous oxide in a neighborhoods affect on price. ‘

4. The data set in CEOSAL2 contains information on chief executive officers for U.S corporations. The variable salary is annual compensation, in thousands of dollars, and ceoten is prior number of years as company CEO



```
> salceo<-lm(log(salary)~ceoten,ceosal2)
> summary(salceo)

Call:
lm(formula = log(salary) ~ ceoten, data = ceosal2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.15314 -0.38319 -0.02251  0.44439  1.94337 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.505498   0.067991  95.682   <2e-16 ***
ceoten      0.009724   0.006364   1.528    0.128    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6038 on 175 degrees of freedom
Multiple R-squared:  0.01316,	Adjusted R-squared:  0.007523 
F-statistic: 2.334 on 1 and 175 DF,  p-value: 0.1284
```

Log(sarlary)= 6.505498 + .009724(CEOTEN)

a) .009724 is the predicted increase in salary given one more year as CEO.

5. A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth weight that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equations that recognizes this is

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u$$

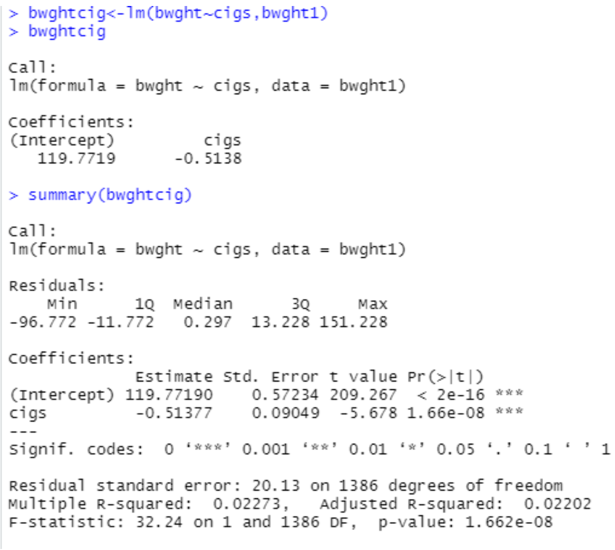
a) What is the likely sign for B2?
    a. The likely sign for b2(family income) is positive, as stated in the question, higher income generally results in access to better prenatal care and mother nutrition

b) Do you think cigs and faminc are likely to be correlated, Explain why the correlation might be positive or negative.
    a. I think family income would be negatively correlated with cigarette consumption. Typically lower income groups such as manual labor jobs have a higher consumption of cigarettes. This is not always the case, there are many people in high income groups that do consume cigarettes but it would be assumed less likely.

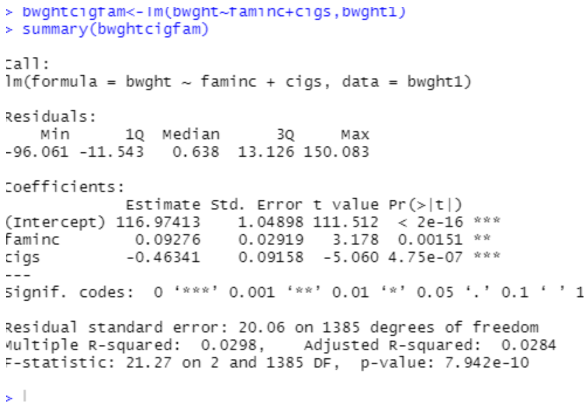
c) Now, estimate the equation with and without faminc, using the data in BWGHT1. Report the results in equation form, including the sample size and R-squared. Discuss your results, focusing on whether adding faminc substantially changes the estimated effect of cigs on bwght.

BWGHT1=119.779 + cigs(-0.51377)X + u
BWGHT1= 116.97413 +faminc(.09276)X +cigs(-.46341)X +u

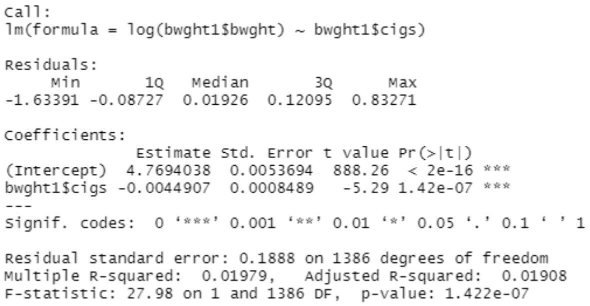
a. Shown in the clippings and equations below the effect of cigarettes on birthweight was reduced when adding in family income. The intercept lowered as well as the slope of cigarettes in the regression equation. I believe this was the case because there actually is correlation between cigarettes and family income.



```
> bwghtcig<-lm(bwght~cigs,bwght1)
> bwghtcig

Call:
lm(formula = bwght ~ cigs, data = bwght1)

Coefficients:
(Intercept)         cigs
   119.7719      -0.5138

> summary(bwghtcig)

Call:
lm(formula = bwght ~ cigs, data = bwght1)

Residuals:
    Min      1Q  Median      3Q     Max
-96.772 -11.772   0.297  13.228 151.228

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 119.77190    0.57234 209.267  < 2e-16 ***
cigs         -0.51377    0.09049  -5.678 1.66e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.13 on 1386 degrees of freedom
Multiple R-squared:  0.02273,	Adjusted R-squared:  0.02202
F-statistic: 32.24 on 1 and 1386 DF,  p-value: 1.662e-08
```

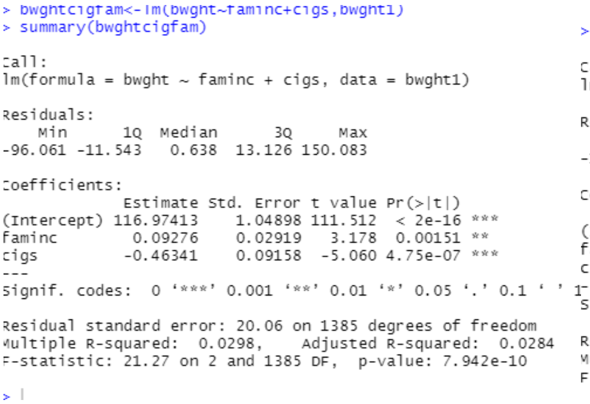


```
> bwghtcigfam<-lm(bwght~faminc+cigs,bwght1)
> summary(bwghtcigfam)

Call:
lm(formula = bwght ~ faminc + cigs, data = bwght1)

Residuals:
    Min      1Q  Median      3Q     Max
-96.061 -11.543   0.638  13.126 150.083

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.97413    1.04898 111.512  < 2e-16 ***
faminc        0.09276    0.02919   3.178  0.00151 **
cigs         -0.46341    0.09158  -5.060 4.75e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.06 on 1385 degrees of freedom
Multiple R-squared:  0.0298,	Adjusted R-squared:  0.0284
F-statistic: 21.27 on 2 and 1385 DF,  p-value: 7.942e-10

>
```

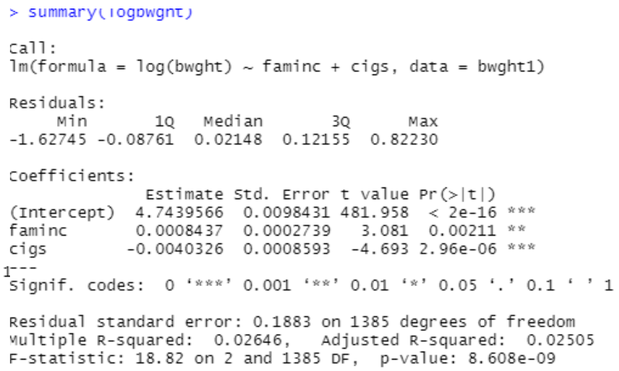


```
Call:
lm(formula = log(bwght1$bwght) ~ bwght1$cigs)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63391 -0.08727  0.01926  0.12095  0.83271

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.7694038  0.0053694  888.26  < 2e-16 ***
bwght1$cigs -0.0044907  0.0008489   -5.29 1.42e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1888 on 1386 degrees of freedom
Multiple R-squared:  0.01979,	Adjusted R-squared:  0.01908
F-statistic: 27.98 on 1 and 1386 DF,  p-value: 1.422e-07
```

If you take the log of BWGHT when considering cigarettes and famine vs the log of BWGHT considering only cigarettes. The difference of R2 from with and without shows that the R2 is higher when considering family income.



```
> bwghtcigfam<-lm(bwght~faminc+cigs,bwght1)
> summary(bwghtcigfam)

Call:
lm(formula = bwght ~ faminc + cigs, data = bwght1)

Residuals:
    Min      1Q  Median      3Q     Max 
-96.061 -11.543   0.638  13.126 150.083 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 116.97413    1.04898 111.512  < 2e-16 ***
faminc        0.09276    0.02919   3.178  0.00151 ** 
cigs         -0.46341    0.09158  -5.060 4.75e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.06 on 1385 degrees of freedom
Multiple R-squared:  0.0298,	Adjusted R-squared:  0.0284 
F-statistic: 21.27 on 2 and 1385 DF,  p-value: 7.942e-10

> |
```



```
> summary(logbwght)

Call:
lm(formula = log(bwght) ~ faminc + cigs, data = bwght1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.62745 -0.08761  0.02148  0.12155  0.82230 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.7439566  0.0098431 481.958  < 2e-16 ***
faminc       0.0008437  0.0002739   3.081  0.00211 ** 
cigs        -0.0040326  0.0008593  -4.693 2.96e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1883 on 1385 degrees of freedom
Multiple R-squared:  0.02646,	Adjusted R-squared:  0.02505 
F-statistic: 18.82 on 2 and 1385 DF,  p-value: 8.608e-09
```

6. Use the data in HPRICE1 to estimate the model
   price = β0 + β1sqrft + β2bdrms + u

   a. Price=-19.31500 +sqrft(0.12844)X+bdrms(15.19819)X

   b. 15.19819 more, using the avg sqrtft of the dataset 2013.693, and the average bedrooms of the dataset 3.568182, rounding to 4

      i. =-19.31500 + 261.119548 +15.19819(4)

         1. -19.31500 +261.119548(for 2013.693sqrft) +60.792(for 4 bedrooms) =302.596548

         2. -19.31500+261.119548+15.19819(5)=317.794738

            a. The remainder is the slope of bdrms being (15.19819)

   C. The estimated increase in the price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part B.

      i. =-19.31500+261.119548(@avgsqrft 2013.6913)+avgbdrm(4)(60.792)=302.596548

      ii= -19.31500+(2033.008+140)+a15.19819(5)= 335.777098

         a. 335.777098-302.596548=33.18055

This is an increase of price, due to the square footage increase on top of the increase in bedrooms.

   d) The R-squared is .6319 so 63.19% of the variation in price is explained by square footage and number of bedrooms

   e) The predicted selling price for a house with 2,438 sqrt and 4 bedrooms

      a. -19.31500 + 313.13672 + 60.79276=354.61448

   f) Residual = observed value – predicted value

      a. 300-354.61448=-54.61338