Chapter 3 Homework

HW Ch 3

Packages used in this key

```
library(wooldridge)
library(tidyverse)
library(pander)
library(mosaic)
library(foreign)
library(car)
```

1 c model_2, 2 a doing two variables, figuring out a certain value, and percentages of the model explained by _____

Question 1

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 equation that recognizes this is

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

(a) What is the most likely sign for $\beta 2$?

Positive because it's likely to increase the birth weight.

(b) Do you think *cigs* and $faminc$ are likely to be correlated? Explain why the correlation might be positive or negative.

Well, I think that's a hard question, and it depends on the years. I think culturally the use of cigarettes has shifted from upper class to mainstream to more lower class over time. So, I suppose nowdays there is likely to be a negative correlation, meaning that increased cigarette use is more likely among lower-earning people.

(c) How, estimate the equation with and without $faminc$, using the data in "bwght." 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*

```
data(bwght, package='wooldridge')
faminc <- bwght$faminc
cigs <- bwght$cigs
bwght <- bwght$bwght

m1 <- lm(bwght~cigs)
summary(m1)
```

```
##
## Call:
## lm(formula = bwght ~ cigs)
##
## 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
```

```
m2 <- lm(bwght~cigs+faminc)
summary(m2)
```

```
##
## Call:
## lm(formula = bwght ~ cigs + faminc)
##
## 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 ***
## cigs         -0.46341    0.09158  -5.060 4.75e-07 ***
## faminc        0.09276    0.02919   3.178  0.00151 **
## ---
## 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
```

```
nrow(model.frame(m1))
```

```
## [1] 1388
```

```
nrow(model.frame(m2))
```

```
## [1] 1388
```

The first equation is estimated to be y = 119.77 + -0.51 cigs. It has a sample size of 1388 observations and an R-squared of 0.022. The second equation is y = 116.97 + -0.46 cigs + 0.09 faminc. It still has a sample size of 1388 observations, it just includes more variables, and has an R-squared of 0.028. However, I do not consider the addition of family income to have a substantial influence on the estimated effect of cigarettes on birth weight. I would base my analysis on the new standard error, not the R-squared. The new standard error is 0.092, and the old standard error was 0.090. So, we haven't removed much of the variance in cigarettes, and I would not consider us to changed the model very much.

Question 2

Use the data in "hprice1" to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u$$

where *price* is the household price measured in thousands of dollars.

```
data(hprice1, package='wooldridge')
sqrft <- hprice1$sqrft
bdrms <- hprice1$bdrms
price <- hprice1$price


mp <- lm(price~sqrft+bdrms)
mp
```

```
##
## Call:
## lm(formula = price ~ sqrft + bdrms)
##
## Coefficients:
## (Intercept)        sqrft        bdrms
##    -19.3150       0.1284      15.1982
```

(a) Write out the results in equation form.

y = -19.315 + .1284 sqrft + 1.1982 bdrms

(b) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

There is an estimated increase of 15.20 in our model, which translates to an increase of $15,200 (rounded).

(c) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answers in part (ii)

```
y = (.1284*140) + 1.1982
y
```

```
## [1] 19.1742
```

```
summary(mp)
```

```
##
## Call:
## lm(formula = price ~ sqrft + bdrms)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -127.627  -42.876   -7.051   32.589  229.003
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.31500   31.04662  -0.622    0.536
## sqrft         0.12844    0.01382   9.291 1.39e-14 ***
## bdrms        15.19819    9.48352   1.603    0.113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.04 on 85 degrees of freedom
## Multiple R-squared:  0.6319, Adjusted R-squared:  0.6233
## F-statistic: 72.96 on 2 and 85 DF,  p-value: < 2.2e-16
```

When we consider the addition of square feet to our model as well as listing 1 more bedroom the model estimates an increase of \$19,200 (rounded). This reveals a flaw in our previous analysis, because the previous price increase would only be true for a bedroom of 0 sq ft, which of course is impossible.

(d) What percentage of the variation in price is explained by square footage and number of bedrooms?

If I understand the question correctly, the R-squared value is .62, so our model explains approximately 62% of the variation in our model. Admittedly our intercept helps the model to account for more variation as well.

(e) The first house in the sample has $sqrft = 2438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

```
y1 = -19.315 + (.1284*2438) + (1.1982*4)
y1
```

```
## [1] 298.517
```

The predicted selling price for this house is \$298,500 (rounded).

(f) The actual selling price of the first house in the sample was $\$300,000$ (so $price = 300$). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

```
# The residual for that price would be 1.5 or $1,500 (rounded). That suggests that the buyer overpaid f
```

Question 3

Use the data in "meap93" to answer this question.

(a) Estimate the model

$$math10 = \beta_0 + \beta_1 \log(expend) + \beta_2 lnchprg + u$$

and report the results in the usual form, including the sample size and R-squared. Are the signs of the slope coefficients what you expected? Explain.

```
data(meap93, package='wooldridge')
math10 <- meap93$math10
lnchprg <- meap93$lnchprg
lexpend <- meap93$lexpend

m10 <- lm(math10~log(lexpend)+lnchprg)
summary(m10)
```

```
##
## Call:
## lm(formula = math10 ~ log(lexpend) + lnchprg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.305  -6.173  -1.302   4.860  43.266
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -78.40335   53.58329  -1.463   0.1442
## log(lexpend)  51.86694   25.13634   2.063   0.0397 *
## lnchprg       -0.30477    0.03537  -8.618   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.528 on 405 degrees of freedom
## Multiple R-squared:  0.1797, Adjusted R-squared:  0.1756
## F-statistic: 44.35 on 2 and 405 DF,  p-value: < 2.2e-16
```

```
nrow(model.frame(m10))
```

```
## [1] 408
```

The sample size is 408 and the R-squared is .18 (rounded). The sign for expenditures is positive, and the sign for lnchprg is negative. I actually would have expected those signs. It makes sense that schools that spend more money have schools who do better on tests because they have better curriculum, or because they come from better socio-economic background. In contrast, I would expect that students enrolled in the lunch program come from a lower socioeconomic status and score lower on the test.

(b) What do you make of the intercept you estimated in part (i)? In particular, does it make sense to set the two explanatory variables to zero? [*Hint*: Recall that $\log(1) = 0$]

In part a I estimated the intercept at 117.0 (rounded). I think it could make sense to set the explanatory variables at 0. It is possible for a school to have a 0% change in expenditures, which is what log(lexpend) would measure. I also think it's possible though to have lunch program enrollment at 0, in a wealthy school where no students need lunches.

(c) Now run the simple regression of $math10$ on $\log(expend)$, and compare the slope coefficients with the estimate obtained in part (i). Is the estimated spending effect now larger or smaller than in part (i)?

```
mlog <- lm(math10~log(lexpend))
summary(mlog)
```

```
##
## Call:
## lm(formula = math10 ~ log(lexpend))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.308  -7.097  -0.919   6.153  39.081
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -174.97      56.93  -3.073 0.002259 **
## log(lexpend)    93.70      26.80   3.497 0.000523 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.35 on 406 degrees of freedom
## Multiple R-squared:  0.02924,    Adjusted R-squared:  0.02685
## F-statistic: 12.23 on 1 and 406 DF,  p-value: 0.0005226
```

The slope co-efficient is 93.70 now, whereas it was 51.86 before. So, the estimated spending effect is now larger.

(d) Find the correlation between

$$lexpend = \log(expend)$$

and

$$lnchprg$$

. Does its sign make sense to you?

```
mlog2 <- lm(lexpend~log(lexpend)+lnchprg)
mlog2
```

```
##
## Call:
## lm(formula = lexpend ~ log(lexpend) + lnchprg)
##
## Coefficients:
##  (Intercept)  log(lexpend)       lnchprg
##   -9.590e+00     8.454e+00     5.282e-06
```

The sign that I got for log(expend) is positive, as is lnchprg. It does make sense to me that expenditures increase as the percentage of expenditures increases, and as more students are enrolled in the lunch program.

(e) Use part (iv) to explain your findings in part (iii)

The coefficient log(expend) really just tells me how much expend increases in dollars for every % change in expend. Basically, each percentage increase corresponds with $8.45e per student. In addition, I can find how much of an increase, on average, there is in expenditures for another 1% of the student body that is enrolled in the lunch program. For every 1% enrollment increase there is an increase of $5.28e per student.

Question 4

Use the data "meapsingle" to study the effects of single-parent households on student math performance. These data are for a subset of schools in southeast Michigan for the year 2000. The socio-economic variables are obtained at the ZIP code level (where ZIP code is assigned to schools based on their mailing address)

(a) Run the simple regression of

$$math4$$

on

$$pctsgle$$

and report the results in the usual format. Interpret the slope coefficient. Does the effect of single parenthood seem large or small?

```
data(meapsingle, package='wooldridge')
math4 <- meapsingle$math4
pctsgle <- meapsingle$pctsgle

ms <- lm(math4~pctsgle)
summary(ms)
```

```
##
## Call:
## lm(formula = math4 ~ pctsgle)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.791  -8.310   1.600   8.092  50.317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 96.77043    1.59680   60.60   <2e-16 ***
## pctsgle     -0.83288    0.07068  -11.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.48 on 227 degrees of freedom
## Multiple R-squared:  0.3795, Adjusted R-squared:  0.3768
## F-statistic: 138.9 on 1 and 227 DF,  p-value: < 2.2e-16
```

The R-squared is .3768 and the sample size is 229 observations. The slope coefficient is -0.83288 which means that single parents do have a negative effect on math scores. However, it makes a difference of less than 1%, which is probably not a very large effect.

(b) Add the variables

$$lmedinc$$

and

$$free$$

to the equation. What happens to the coefficient on

$$pctsgle$$

? Explain what is happening.

```
lmedinc <- meapsingle$lmedinc
free <- meapsingle$free

ms2 <- lm(math4~pctsgle+lmedinc+free)
summary(ms2)
```

```
##
## Call:
## lm(formula = math4 ~ pctsgle + lmedinc + free)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.919  -7.195   0.931   7.313  50.152
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.72322   58.47814   0.884    0.377
## pctsgle     -0.19965    0.15872  -1.258    0.210
## lmedinc      3.56013    5.04170   0.706    0.481
## free        -0.39642    0.07035  -5.635  5.2e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.7 on 225 degrees of freedom
## Multiple R-squared:  0.4598, Adjusted R-squared:  0.4526
## F-statistic: 63.85 on 3 and 225 DF,  p-value: < 2.2e-16
```

After I add free and median income to the equation the coefficient on pctsgle decreases down to -0.200 (rounded). That means that factoring in the other variables suggests that the effect of a single parent is smaller than we previously thought.

(c) Find the sample correlation between $lmedinc$ and $free$. Does it have the sign you expect?

```
ms3 <- lm(lmedinc~free)

summary(ms3)
```

```
##
## Call:
## lm(formula = lmedinc ~ free)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52751 -0.13281 -0.01978  0.11409  0.96440
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.3671789  0.0183937  617.99   <2e-16 ***
## free        -0.0118464  0.0006998  -16.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2078 on 227 degrees of freedom
## Multiple R-squared:  0.558,  Adjusted R-squared:  0.556
## F-statistic: 286.5 on 1 and 227 DF,  p-value: < 2.2e-16
```

Free has a negative sign for median income. That suggests that the percentage of people eligible for free lunches is correlated with a lower median income in a geographic area.

(d) Does the substantial correlation between *lmedinc* and $free$ mean that you should drop one from the regression to better estimate the causal effect of single parenthood on student performance? Explain.

It doesn't look like medinc and free have particularly significant correlations. However, we should never drop a variable because it is similar to another variable. We always base our inclusion on theoretical reasons, such as variables that we believe explain one another.

(e) Find the variance inflation factors (VIFs) for each of the explanatory variables appearing in the regression in part (ii). Which variable has the largest VIF? Does this knowledge affect the model you would use to study the causal effect of single parenthood on math performance?

```
vif(ms2)
```

```
##  pctsgle  lmedinc     free
## 5.740981 4.118812 3.188079
```

pctsgle has the largest VIF. I suppose it depends on whether I think single parenthood is significant. If I think that single parenthood can in essense be distilled down to income and likelyhood of participation in free lunches, I can remove it. However, I believe single parenthood likely has an effect outside of income and social benefits, and so I'm obligated to keep it in the model.