Problem C3.1, page 110 in 5e Wooldridge

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:

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

(i) What is the most likely sign for β_2?

Sol. Probably $\beta_2 > 0$, as more income typically means better nutrition for the mother and better prenatal care.

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

Sol. Ex ante, it's not quite clear which direction they are correlated. One one hand, consumption generally increases with income, and $cigs$ and $faminc$ could be positively correlated. On the other hand, income is generally higher for more educated families, and education and cigarette smoking are negatively correlated. In this sample, the correlation between $cigs$ and $faminc$ is -0.1869.

(iii) Now, estimate the equation with and without $faminc$, using the data in BWGHT.dta. Report the results in equation form, including the sample size and R^2. Discuss your results, focusing on whether adding $faminc$ substantially changes the estimated effect of $cigs$ on $bwght$.

Sol. The equations estimated with Stata are:

$$\widehat{bwght} = 119.653 - .6025 cigs$$
$$R^2 = 0.034, \;\; N = 900$$

$$\widehat{bwght} = 117.075 - .5523 cigs + .0861 faminc$$
$$R^2 = 0.04, \;\; N = 900$$

In the first equation, if the average number of cigarettes smoked during pregnancy ($cigs$) increases by 1 cigarette, the predicted birth weight decreases by 0.6025 ounces. If $faminc$ is included, this predicted change in birth weight increases slightly to -0.5523 ounces.

The difference in $\hat{\beta}_1$ is due to the correlation between $cigs$ and $faminc$, so that it is downward biased (more negative) when income is not controlled for. However, the bias is small because $cigs$ and $faminc$ are not highly correlated, and $faminc$ has a small effect on birth weight.

Replicate Lecture 7 in R

See Lecture7.R and accompanying data file to replicate what we went over in Lecture 7.

```
#Lecture 7- Spring 2021


#-----------------------------------------
#when you run this for the first time uncomment and instal the packages below
#then for future runs comment them out
#-----------------------------------------
#... deleted to save space in this document
# Loading packages
library(dplyr)
library(haven)
library(readr)
library(knitr)
library(haven)
library(dplyr)
library(readxl)
library(psych)
library(ggplot2)
library(stats4)
library(lmSupport)
library(magrittr)
library(qwraps2)
library(stargazer)

#change directory
#-----------------------------------------
#set your working directory
#-----------------------------------------
#setwd("/Users/sberto/Desktop/")
setwd("/Users/sofiavillas-boas/Dropbox/EEP118_Spring2021/Lectures/Lecture7")


#-----------------------------------------
#1. Read in data
#-----------------------------------------
my_data <- read_dta("dataLecture7.dta")
head(my_data)

#generate variables Y and x1
my_data$Y<-my_data$trump16/my_data$clinton16
my_data$x1<-my_data$romney12/my_data$obama12

slide12<-plot(my_data$x1,my_data$Y)

#use non missing values only from now on
my_data2<-my_data[complete.cases(my_data),]
my_data3 <- my_data2[my_data2$Y !=Inf,]
```

```
#my_data3 <-my_data2[my_data2$white_pct!=NA]

summary(my_data3)
#regression full model
regfull<-lm(Y ~ x1 + white_pct, my_data3)
summary(regfull)

#regression no percent white as control
regsmall<-lm(Y ~ x1, my_data3)
summary(regsmall)

#collinearity Slides

#Baseline Model
regBase <- lm(Y~x1+female_pct+white_pct, my_data3)
summary(regBase)

#Alternative Model  Perfect collinearity
my_data3$male_pct=1-my_data3$female_pct
regPC <- lm(Y~x1+female_pct+male_pct+white_pct, my_data3)
summary(regPC)


#Alternative Model  Multi collinearity
regMC <- lm(Y~x1+female_pct+white_pct+bh_pct, my_data3)
#where corr(white_pct,bh_pct)=-0.92
summary(regMC)


#for graph in slide 28
#get the predicted Y hats
my_data3$Yhat<-regBase$fitted.values

#put Y and Y hat on same graph and x1 on horizontal axis
#make combined scatter plot of Y data and fitted values of Y (Yhat)
#given regression estimates usinf X1, white_pct and female_pct
scatter_data_fittedVals <- ggplot(data = my_data3) + geom_point(aes(x=x1, y=Y, color = "data")) +
  geom_point(aes(x=x1, y=Yhat, color = "fitted")) +
  xlab("x1=Ratio Rodney to Obama Votes in 2012") + ylab("Y=Ratio Trump to Clinton Votes in 2016 ") +
  ggtitle("Y (Red) and Predicted Y (Blue) and x1")
scatter_data_fittedVals
```