Applied Micro PSET 2

Matthew Zhao working with Cedric Elkouh

2023-01-16

```
library(tidyverse)
library(haven)
library(stargazer)
library(xtable)
options(scipen=999)
```

Question 1

```
df <- read_stata('data/gpa2.DTA')
```

a)

$$colgpa_i = \beta_0 + \beta_1 female_i + U_i \tag{1}$$

```
model_a <- lm(colgpa ~ female, data = df)
stargazer(model_a,type='text',digits=3,
          title='Table 1 - OLS Estimates of (1)')
```

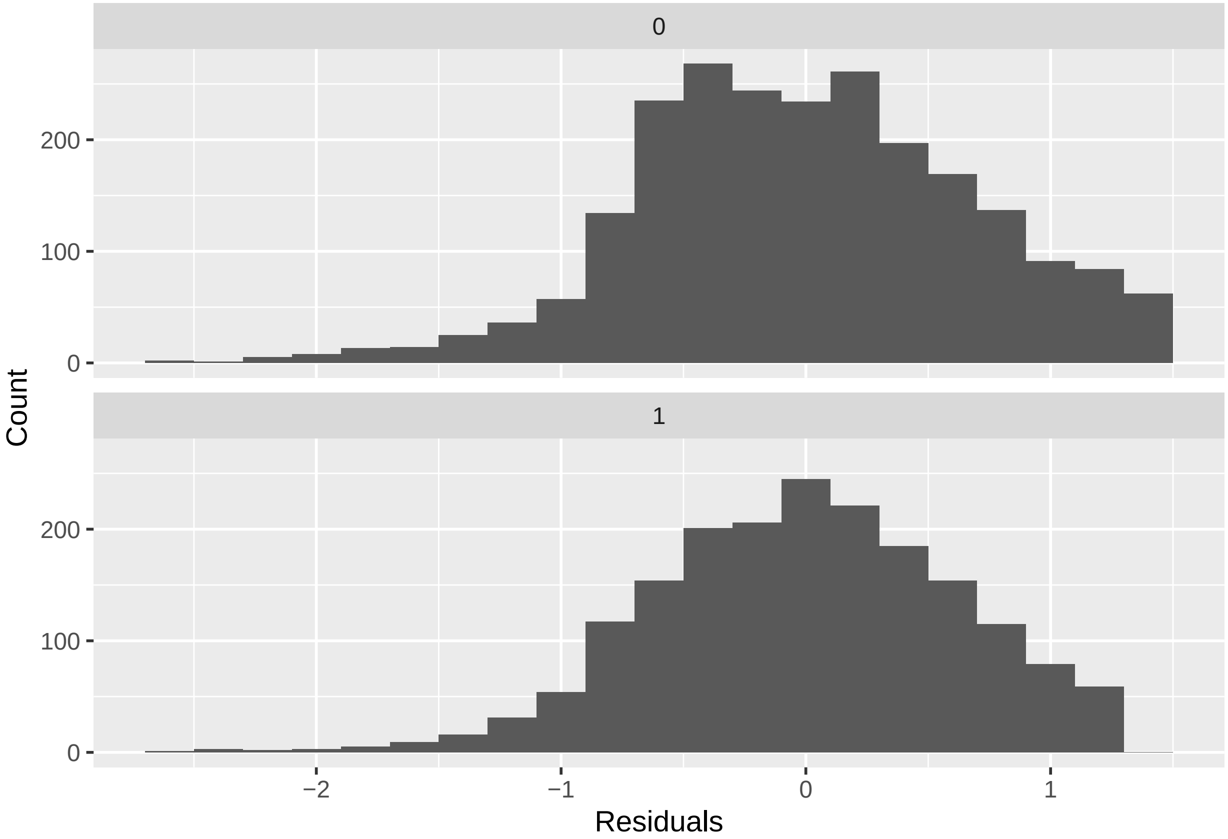
```
##
## Table 1 - OLS Estimates of (1)
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                               colgpa
## -----------------------------------------------
## female                       0.142***
##                               (0.020)
##
## Constant                     2.589***
##                               (0.014)
##
## -----------------------------------------------
## Observations                   4,137
## R2                             0.012
## Adjusted R2                    0.011
## Residual Std. Error      0.655 (df = 4135)
## F Statistic          48.157*** (df = 1; 4135)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

From Table 1, we see that the OLS estimate for β_1 in (1) is $\hat{\beta}_1 = 0.142$ and is significant at the 1% significance level. We interpret this coefficient as an individual who is female has a college gpa that is on average approximately 0.142 higher than someone who is not. We can interpret the OLS estimate of β_0 which is $\hat{\beta}_0 = 2.589$ as the approximate average college gpa of someone who is not female.

b)

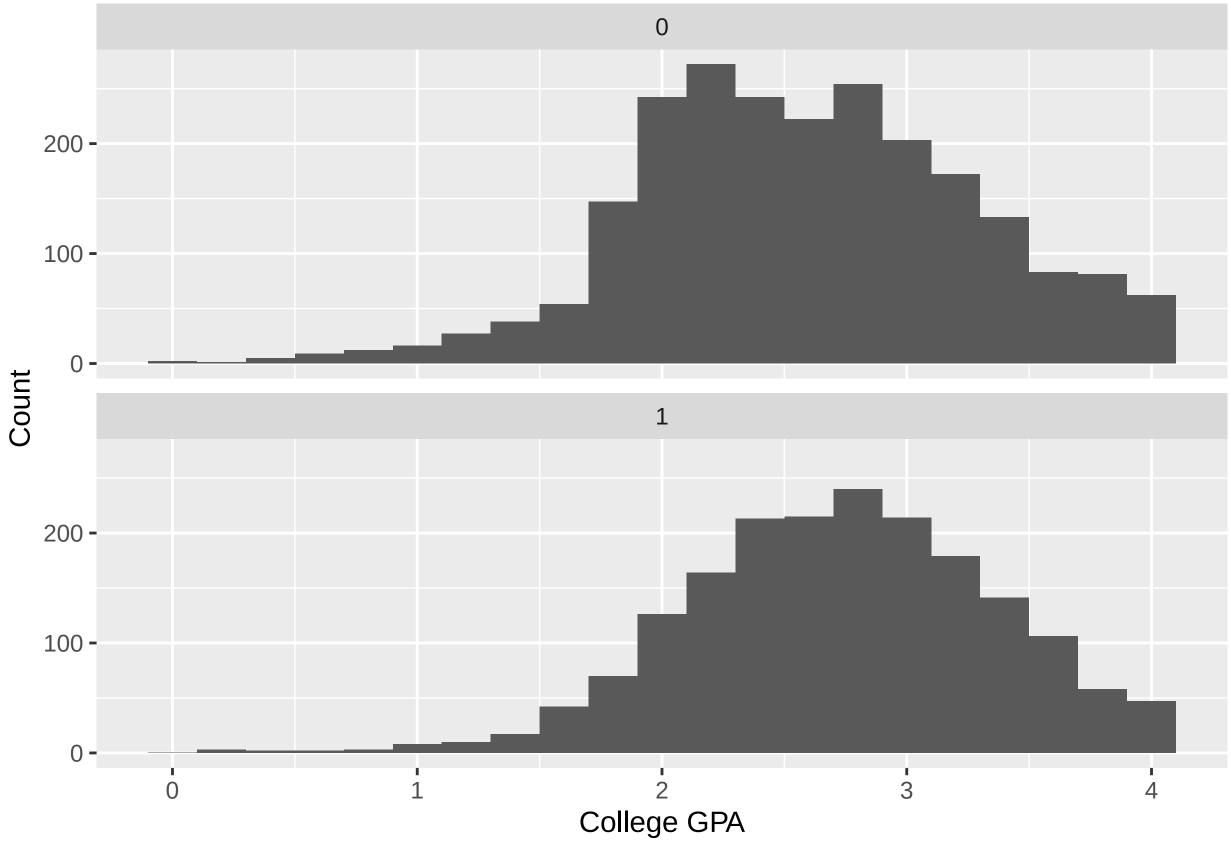
```
temp <- df %>%
  mutate(res_b = lm(colgpa ~ female, data = df)$residuals)

ggplot(data = temp, aes(x = res_b)) +
  geom_histogram(binwidth = 0.2) +
  facet_wrap(~female,nrow=2) +
  labs(x = 'Residuals', y = 'Count')
```



Homoskedasticity is defined as $Var(U_i|X_i) = \sigma^2$ for all $i = 1, .., n$. From the distributions of residuals above, we see that the spread in residuals is slightly smaller for observations with female = 1 i.e. individuals who are female than for those who are not. This could potentially be evidence of heteroskedasticity.

```
ggplot(data = df, aes(x = colgpa)) +
  geom_histogram(binwidth = 0.2) +
  facet_wrap(~female,nrow=2) +
  labs(x = 'College GPA', y = 'Count')
```



The distribution of residuals conditional on female is a reflection of the distribution of colgpa conditional on female because in this regression the only explanatory variable we are using is female. As a result, the spread of college gpa conditional on female directly determines the spread of residuals conditional on female e.g. individuals who are not female tend to have a wider range of gpas means that the residuals for individuals who are not female will be more spread out since they will be further from the fitted values from the regression and hence have a wider range of residuals.

c)

$$colgpa_i = \beta_0 + \beta_1 female_i + \beta_2 sat_i + V_i \tag{2}$$

```
model_c <- lm(colgpa ~ female + sat, data = df)
stargazer(model_c,type='text',digits=3,
          title='Table 2 - OLS Estimates of (2)')
```

```
##
## Table 2 - OLS Estimates of (2)
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                               colgpa
## -----------------------------------------------
## female                       0.231***
##                               (0.019)
##
```

```
## sat                           0.002***
##                              (0.0001)
##
## Constant                      0.429***
##                               (0.071)
##
## -----------------------------------------------
## Observations                    4,137
## R2                              0.197
## Adjusted R2                     0.196
## Residual Std. Error       0.590 (df = 4134)
## F Statistic           506.099*** (df = 2; 4134)
## ===============================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

From Table 2, we see that the OLS estimate for β_1 in (2) is $\hat{\beta}_1 = 0.231$ and that β_2 in (2) is $\hat{\beta}_2 = 0.002$, with both significant at the 1% significance level. We interpret the first coefficient as an individual who is female has a college gpa that is on average approximately 0.231 higher than someone who is not with the same sat score. We interpret the second coefficient as an individual would have a college gpa that is approximately 0.002 higher on average if they had an sat score than is 1 point higher and are the same gender. The OLS estimate of β_0 which is $\hat{\beta}_0 = 2.589$ represents the approximate average college gpa of someone who is not female and has an sat score of 0. However, since the sample does not contain anyone with an sat score of 0, this is an extrapolation and hence may not be an accurate interpretation.

Since the partial effect of female increased after controlling for sat score, the change suggests that female is negatively correlated with sat score.

In comparing models (1) and (2), we are looking at potential OVB of SLR vs MLR, where we can decompose the OLS estimate of β_1 from the SLR (from here on referred to as $\tilde{\beta}_1$) as $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$ where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the OLS estimates of the parameters of (2) and $\tilde{\delta}_1$ is from the OLS estimate of $sat = \delta_0 + \delta_1 female + U$. Specifically, we can derive the implied OVB of leaving out sat using the first equation from earlier, where $E[\tilde{\beta}_1] - \hat{\beta}_1 = \hat{\beta}_2\tilde{\delta}_1$. Using the values from part (b) and below, we find that the implied OVB of leaving out sat in the regression in (a) is around $0.002 * -43.07331 = -0.086$, which is roughly the difference between $\hat{\beta}_1$ and $\tilde{\beta}_1$ ($\tilde{\beta}_1 - \hat{\beta}_1 = 0.142 - 0.231 = -0.089$).

```
temp <- lm(sat ~ female, data = df)
temp$coefficients
```

```
## (Intercept)      female
##  1049.69697   -43.07331
```

d)

While the OLS estimator $\hat{\beta}$ of the parameter vector β in (2) remains unbiased and consistent, the variance of this estimator $Var(\hat{\beta})$ is now biased.

e)

Breuch-Pagan

```r
U_hat <- model_c$residuals
U_reg <- lm(U_hat^2 ~ df$female + df$sat)
U_rsq <- summary(U_reg)$r.squared
n <- nrow(df)
k <- 2

F_stat <- (U_rsq/k) / ((1-U_rsq)/(n-k-1))
pf(F_stat,k,n-k-1,lower.tail = F)
```

```
## [1] 0.004509745
```

White Test

```r
white_reg <- lm(U_hat^2 ~ df$colgpa + I(df$colgpa)^2)

stargazer(white_reg,type='text',digits=3,
          title='Table 4 - White Test for Heteroskedasticity')
```

```
##
## Table 4 - White Test for Heteroskedasticity
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                U_hat2
## -----------------------------------------------
## colgpa                        -0.186***
##                                (0.013)
##
## colgpa)
##
##
## Constant                      0.842***
##                                (0.035)
##
## -----------------------------------------------
## Observations                    4,137
## R2                              0.050
## Adjusted R2                     0.049
## Residual Std. Error       0.537 (df = 4135)
## F Statistic           216.105*** (df = 1; 4135)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

We do find some evidence for heteroskedasticity since we see that the BP test rejects the null of homoskedasticity at the 1% significance level. However the F-stat for the White Test is large and significant, meaning that the coefficients are likely to be non-zero and the test fails to reject.

f)

```
library(lmtest)
library(sandwich)

summary(model_c)
```

```
##
## Call:
## lm(formula = colgpa ~ female + sat, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.98049 -0.37896  0.02343  0.41931  1.77611
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 0.42895422 0.07105247   6.037        0.00000000171 ***
## female      0.23067073 0.01867624  12.351 < 0.0000000000000002 ***
## sat         0.00205761 0.00006665  30.870 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5905 on 4134 degrees of freedom
## Multiple R-squared:  0.1967, Adjusted R-squared:  0.1963
## F-statistic: 506.1 on 2 and 4134 DF,  p-value: < 0.00000000000000022
```

```
coeftest(model_c, vcov = vcovHC(model_c, type = 'HC1'))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value              Pr(>|t|)
## (Intercept) 0.428954224 0.069069447  6.2105       0.0000000005803 ***
## female      0.230670735 0.018451723 12.5013 < 0.00000000000000022 ***
## sat         0.002057612 0.000065125 31.5947 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The standard errors are lower while the t-statistics are larger. We expect this since we saw earlier that the variance in residuals is lower for females (female = 1) than for males (female = 0). This means that the residuals are negatively correlated with σ_i^2. As a result, since the formula for robust standard errors is $\hat{\sigma}(\hat{\beta}_j) = \sqrt{\frac{\sum_{i=1}^{n} \hat{\epsilon}_{ij}^2 \hat{U}_i^2}{SSR_j^2}}$ and for non robust is $\hat{\sigma}(\hat{\beta}_j) = \frac{\hat{\sigma}_U}{\sqrt{SST_j(1-R_j^2)}}$, we see that the robust standard errors are smaller than nonrobust when the variance and residuals are negatively correlated as we see here.

g)

```
m1 <- lm(colgpa ~ female + sat, data = df)
m2 <- lm(colgpa ~ female + sat + I(sat^2), data = df)
```

```r
m3 <- lm(colgpa ~ female + sat + I(sat^2) + I(sat^3), data = df)
models <- list(m1,m2,m3)

stargazer(models,type='text',digits=3,
          title='Table 5 - Different Functional Forms Variants of (2)')
```

```
##
## Table 5 - Different Functional Forms Variants of (2)
## ==================================================================================================
##                                                   Dependent variable:
##                      ------------------------------------------------------------------------------
##                                                         colgpa
##                                (1)                       (2)                       (3)
## --------------------------------------------------------------------------------------------------
## female                      0.231***                  0.239***                  0.239***
##                              (0.019)                   (0.019)                   (0.019)
##
## sat                         0.002***                  -0.002***                  -0.003
##                             (0.0001)                   (0.001)                   (0.004)
##
## I(sat2)                                              0.00000***                  0.00000
##                                                       (0.00000)                 (0.00000)
##
## I(sat3)                                                                          -0.000
##                                                                                  (0.000)
##
## Constant                    0.429***                  2.479***                  2.846**
##                              (0.071)                   (0.342)                   (1.300)
##
## --------------------------------------------------------------------------------------------------
## Observations                  4,137                     4,137                     4,137
## R2                            0.197                     0.204                     0.204
## Adjusted R2                   0.196                     0.203                     0.203
## Residual Std. Error     0.590 (df = 4134)         0.588 (df = 4133)         0.588 (df = 4132)
## F Statistic         506.099*** (df = 2; 4134) 352.873*** (df = 3; 4133) 264.618*** (df = 4; 4132)
## ==================================================================================================
## Note:                                                                 *p<0.1; **p<0.05; ***p<0.01
```

We find that adding polynomial terms for sat works up to the second power, with the coefficient being statistically significant at the 1% level. Subsequent terms add little to the regression and are not significant.

Non-random selection of students into college may explain this weakly negative relationship between college gpa and sat score since generally, students go to the best school they are able to. As a result, they typically will be taking classes with students who are at similar levels of intelligence/work ethic. With classes being curved, there is a limit to how well you can do based on how well other students in your class do. This has the largest effect on students at the higher end of the sat distribution, since the sat is an absolute measure of high school performance, while college gpa is, to an extent, a relative measure of performance in college. Hence, the quadratic form best fits the functional form relationship between college gpa and sat.

h)

```
with_hsperc <- lm(colgpa ~ female + sat + hsperc, data = df)
stargazer(list(with_hsperc,model_c),type='text',digits=4,
          title='Table 6 - Comparison of Model with hsperc')
```

```
## 
## Table 6 - Comparison of Model with hsperc
## =========================================================================
##                                       Dependent variable:                
##                      -----------------------------------------------------
##                                             colgpa                       
##                                 (1)                        (2)           
## -------------------------------------------------------------------------
## female                      0.1489***                  0.2307***         
##                              (0.0180)                   (0.0187)         
##                                                                          
## sat                         0.0016***                  0.0021***         
##                              (0.0001)                   (0.0001)         
##                                                                          
## hsperc                      -0.0126***                                   
##                              (0.0006)                                    
##                                                                          
## Constant                    1.1908***                  0.4290***         
##                              (0.0750)                   (0.0711)         
##                                                                          
## -------------------------------------------------------------------------
## Observations                  4,137                      4,137           
## R2                            0.2853                     0.1967          
## Adjusted R2                   0.2848                     0.1963          
## Residual Std. Error     0.5570 (df = 4133)         0.5905 (df = 4134)    
## F Statistic         549.9337*** (df = 3; 4133) 506.0991*** (df = 2; 4134)
## =========================================================================
## Note:                                         *p<0.1; **p<0.05; ***p<0.01
```

```
sat_nohsperc <- lm(sat ~ female, data = df)
sat_whsperc <- lm(sat ~ female + hsperc, data = df)

stargazer(list(sat_whsperc,sat_nohsperc),type='text',digits=4,
          title='Table 7 - sat on explanatory vars')
```

```
## 
## Table 7 - sat on explanatory vars
## =========================================================================
##                                       Dependent variable:                
##                      -----------------------------------------------------
##                                               sat                        
##                                 (1)                        (2)           
## -------------------------------------------------------------------------
## female                     -55.9253***                -43.0733***        
##                              (4.1350)                   (4.3056)         
##                                                                          
## hsperc                      -2.6281***                                   
##                              (0.1242)                                    
```

```
##
## Constant                 1,106.0330***            1,049.6970***
##                            (3.8217)                 (2.8870)
##
## ----------------------------------------------------------------------------
## Observations                4,137                    4,137
## R2                          0.1191                   0.0236
## Adjusted R2                 0.1187                   0.0234
## Residual Std. Error    130.8685 (df = 4134)      137.7611 (df = 4135)
## F Statistic         279.4685*** (df = 2; 4134) 100.0816*** (df = 1; 4135)
## ============================================================================
## Note:                                         *p<0.1; **p<0.05; ***p<0.01
```

hsperc represents the class percentile a student was in in high school starting from the top. Omitting hsperc causes $\hat{\beta}_2$ to be larger because a student's class percentile is most likely correlated with sat score. As a result of this relationship, including hsperc would reduce the unique variation in sat and cause this coefficient to fall. Additionally, omitting hsperc should reduce the variance of $\hat{\beta}_2$ since $\hat{\sigma}(\hat{\beta}_j) = \frac{\hat{\sigma}_U}{\sqrt{SST_j(1-R_j^2)}}$. Omitting hsperc would reduce R_j^2 since the other explanatory variables account for less of the variation in sat. As a result, because the denominator is larger, the overall variance/se is smaller.

i)

The idea randomized experiment would be to take a large group of individuals right before they enter college and clone a female version of them with the same background, experiences, etc or turn them female so that we can completely isolate the causal effect of gender on college gpa.

j)

Besides the practical difficulties, empirically, it is difficult to implement the ideal randomized experiment we described in (i) because an individual's life is largely shaped by their assigned sex at birth, meaning that it is very unlikely for MLR.4 to hold since there are innumerable unobservables that affect an individual. An example in this case would be that even if we could change some individual's genders, their past experience in their initial gender could have an impact on their college gpa. More concretely, there are numerous papers which show differential education and educational opportunities between men and women starting from a young age. As a result, switching from male to female is simply not comparable to being female your whole life, making it empirically challenging to learn the true effect of being female on college gpa.

Conceptually, it does not make sense to estimate a causal effect of female. Firstly, this is because, while quantifying the gender gap in educational achievement is of interest, practically it is difficult to define what being female means. Would we define being female as biological in terms of horomones and physical traits, or focus on identity and perceived gender? These questions make it conceptually difficult to understand what a causal effect of female would even mean, hence the question not making sense.

k)

I do not find MLR.4 credible in the population model. This is because it is very likely that any number of unobservables could be correlated with female and affect college gpa e.g. high school performance, work ethic, geography, cultural values, number/gender/order of siblings, etc. I would want to collect information on the already mentioned variables in addition to family wealth, parental education, schools attended/attending, etc.

I believe that including the most important omitted variable in (1) (work ethic) would reduce $\hat{\beta}_1$ since the estimator is initially positive in the OLS estimate of (1) and work ethic is likely to be positively correlated with being female. Additionally, since work ethic is positively correlated with being female, it is likely to increase the variance of $\hat{\beta}_1$ since it reduces the unique variance of female in explaining college gpa.

Problem 2

a)

Given X_2 is independent of X_3 and that X_1 is correlated with X_2 and X_3:

$X_2 = \delta_0^2 + \delta_1^2 X_1 + \epsilon_2 \Rightarrow \epsilon_2 = X_2 - \delta_0^2 - \delta_1^2 X_1$ and $X_3 = \delta_0^3 + \delta_1^3 X_1 + \epsilon_3 \Rightarrow \epsilon_3 = X_3 - \delta_0^3 - \delta_1^3 X_1$.

$Cov(\epsilon_2, \epsilon_3) = Cov(X_2 - \delta_0^2 - \delta_1^2 X_1, X_3 - \delta_0^3 - \delta_1^3 X_1) = -\delta_1^3 Cov(X_2, X_1) - \delta_1^2 Cov(X_1, X_3) + \delta_1^2 \delta_1^3 Var(X_1)$

Since we know that X_1 is correlated with X_2 and X_3 and variance is non-negative, $Cov(\epsilon_2, \epsilon_3) \neq 0$ and so they are correlated.

From there it is sufficient to show that if α_2 in the equation $X_3 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + U$ is nonzero, the OLS estimate for β_2 in the equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ is biased. Since $\gamma_2 = \alpha_2$ where γ_2 is from the equation $\epsilon_3 = \gamma_0 + \gamma_2 \epsilon_2 + V$, and we previously showed that $\rho_{\epsilon_2, \epsilon_3} \neq 0$, $\gamma_2 = \alpha_2 \neq 0$. Hence, the OLS estimate for β_2 will be biased.

b)

Since X_2 is now independent of X_1, it follows from part (a) that $\hat{\beta}_2$ is now unbiased.

For $\hat{\beta}_1$:

From 3.23: $E[\hat{\beta}_1] = \beta_1 + \beta_3 \tilde{\delta}_1 \Rightarrow OVB = E[\hat{\beta}_1] - \beta_1 = \beta_3 \tilde{\delta}_1$

$$\tilde{\delta}_1 = \frac{Cov(X_1, X_3)}{Var(X_1)} = \frac{\frac{1}{n}\sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{3i} - \bar{X}_3)}{\frac{1}{n}\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2}$$

$\sum (X_{1i} - \bar{X}_1)(X_{3i} - \bar{X}_3) = \sum X_{1i} X_{3i} - \sum X_{1i} \bar{X}_3 - \sum \bar{X}_1 X_{3i} + \sum \bar{X}_1 \bar{X}_3$

$= \sum X_{1i} X_{3i} - \bar{X}_1 X_{3i} + \sum \bar{X}_1 \bar{X}_3 - X_{1i} \bar{X}_3$

$= \sum (X_{1i} - \bar{X}_1) X_{3i} + n\bar{X}_3 \sum \bar{X}_1 - X_{1i}$

Since $\sum \bar{X}_1 - X_{1i} = 0$, the prior expression is equal to $\sum (X_{1i} - \bar{X}_1) X_{3i}$

Therefore, our approximate OVB for β_1 is $\dfrac{\sum_{i=1}^n (X_{1i} - \bar{X}_1) X_{3i}}{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2}$

Problem 3

To preface my answer to this section, I have decided to change my paper to replicate from my answer in the previous week since I was not able to work with the data for that paper. As a result, I have now chosen a different paper for which I have the data files (linked **here**)

a)

The main dependent variable of interest is stock portfolio allocation. The main explanatory variables are retirement, marital status, family labor income, net worth, pension income, number of children, age, health care expenditures. Since this is panel data, the author also includes household and time fixed effects.

b)

I have been able to replicate these in stata but I aim to translate these into R for my final project.