Midterm

Brian Gilmore

2023-02-20

First I load my packages and save our data as m_df:

```
library(pacman)
pacman::p_load(Ecdat, magrittr, dplyr, here, ggplot2, fixest)
m_df <- read.csv("midterm-data.csv")
```

[1]

```
workhrs <- lm(hrs_work_perwk ~ i_female, data = m_df)
summary(workhrs)
```

```
##
## Call:
## lm(formula = hrs_work_perwk ~ i_female, data = m_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.237  -4.237   0.763   4.329  63.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.2373     0.1995  196.71   <2e-16 ***
## i_female     -3.5665     0.2941  -12.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.69 on 7498 degrees of freedom
## Multiple R-squared:  0.01924,    Adjusted R-squared:  0.01911
## F-statistic: 147.1 on 1 and 7498 DF,  p-value: < 2.2e-16
```

[2] Our regression shows that we can expect non-females to work an average of 39.2373 hours each week. Our p-value indicates our output to be statistically significant as it is less than .05. We can also expect females to work 3.5665 less on average than non-females.The p-value is also less than .05, which shows statistical significance. Therefore we can expect non-females to work significantly less than 40 hours a week on average.

[3]

```
robustwrk <- feols(hrs_work_perwk ~ i_female, data = m_df, vcov = "hetero")
summary(robustwrk)
```

```
## OLS estimation, Dep. Var.: hrs_work_perwk
## Observations: 7,500
## Standard-errors: Heteroskedasticity-robust
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 39.23734   0.194714 201.5129 < 2.2e-16 ***
## i_female    -3.56652   0.295262 -12.0792 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 12.7   Adj. R2: 0.01911
```

No, utilizing heteroskedasticity robust standard errors doesn't affect my answers to the last question. However, standard error for the intercept is slightly smaller while the standard error for i_female has slightly increased.

[4]

```
kids_wrk <- lm(hrs_work_perwk ~ i_female + i_kids, data = m_df)
summary(kids_wrk)
```

```
##
## Call:
## lm(formula = hrs_work_perwk ~ i_female + i_kids, data = m_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.444  -4.105   0.895   4.455  63.455
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.1055     0.2316 168.875   <2e-16 ***
## i_female     -3.5604     0.2941 -12.106   <2e-16 ***
## i_kids        0.3384     0.3018   1.121    0.262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.69 on 7497 degrees of freedom
## Multiple R-squared:  0.01941,    Adjusted R-squared:  0.01914
## F-statistic: 74.18 on 2 and 7497 DF,  p-value: < 2.2e-16
```

[5] From our regression, our intercept is statistically different from 0 with a p-value<2e-16, which is less than .05. Our intercept tells us that we can expect a non-female with no kids to work an average of 39.1055 hours a week. Our coefficient estimate for i_female tells us that we can expect a female to work 3.5604 hours less than non-female counterparts. Our coefficient estimator for i_kids tells us that if an individual has kids, we can expect a .3384 average increase in the number of work hours. Our coefficients for both i_female and i_kids are statistically significant.

[6]

```
kids_fem_wrk <- lm(hrs_work_perwk ~ i_female + i_kids + i_female:i_kids, data = m_df)
summary(kids_fem_wrk)
```

```
##
## Call:
```

```
## lm(formula = hrs_work_perwk ~ i_female + i_kids + i_female:i_kids,
##     data = m_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.781  -3.890   1.110   4.531  63.210
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      38.8903     0.2553 152.340   <2e-16 ***
## i_female         -3.1001     0.3734  -8.303   <2e-16 ***
## i_kids            0.8904     0.4089   2.177   0.0295 *
## i_female:i_kids  -1.2118     0.6059  -2.000   0.0455 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.69 on 7496 degrees of freedom
## Multiple R-squared:  0.01993,    Adjusted R-squared:  0.01954
## F-statistic: 50.81 on 3 and 7496 DF,  p-value: < 2.2e-16
```

[7] Our intercept tells us that we can expect a non-female with 0 kids to work an average of 38.8903 hours. Our coefficient for i_female tells us that we can expect a female to work 3.1001 less than non-females on average. Our estimate on i_kids expects that for an individual with kids we can expect average work hours to increase by .8904. Our interaction coefficient tells us that the expected return for females with kids will work 1.2118 hours less than non-females without kids. Our coefficients and intercept are statistically significant at the 5% level since no p-value exceeds .05.

[8] Hours worked by females with children: (38.8903-1.2118) = 37.6785 Hours worked by females without children: (38.8903-3.1001) = 35.7902 Numerical difference between the number of hours: (37.6785-35.7902) = 1.8883

[9] For non-females with children: (38.8903 + 0.8904) = 39.7807 average hours worked.

[10] I think that omitting education level could cause bias since education level will have an effect on hours worked. Education level and the indicator for i_female could be correlated since returns to education could differ between females and non-females. Education level and the indicator for i_kids could also be correlated since the decision to have kids might rely on whether or not someone has finished school, for example. This fulfills the requirements for OVB since yrs_education could have an effect on hrs_work_perwk. Additionally yrs_education and i_female, as well as yrs_education and i_female, could have covariance not equal to 0.

[11]

```
ed_reg <- lm(hrs_work_perwk ~ i_female + i_kids + i_female:i_kids + yrs_education, data = m_df)
summary(ed_reg)
```

```
##
## Call:
## lm(formula = hrs_work_perwk ~ i_female + i_kids + i_female:i_kids +
##     yrs_education, data = m_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39.71  -3.71   1.11   4.75  67.98
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     34.25727    0.61841  55.396  < 2e-16 ***
## i_female        -3.23339    0.37209  -8.690  < 2e-16 ***
## i_kids           1.15457    0.40839   2.827  0.00471 **
## yrs_education    0.33087    0.04026   8.218 2.42e-16 ***
## i_female:i_kids -1.33416    0.60338  -2.211  0.02706 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.63 on 7495 degrees of freedom
## Multiple R-squared:  0.02868,    Adjusted R-squared:  0.02816
## F-statistic: 55.33 on 4 and 7495 DF,  p-value: < 2.2e-16
```

[12] It's possible that yrs_education was causing OVB since the coefficient for yrs_education is statistically different from 0, with a p-value at 2.42e-16, having an effect of .33087 on the expected average hours worked per week. This indicates that yrs_education may have an effect on our outcome variable. Additionally, implementing yrs_education changes our coefficient estimates in the model, implying non-zero covariance between yrs_education and other explanatory variables.

[13] The intercept changed since yrs_education had an effect on the outcome variable that was correlated with our explanatory variable coefficient estimates. Implementing yrs_education means our intercept returns the expected hours worked per week for non-females with no kids and zero years of education.

[14] In this context, there is potential for measurement error in our regression. For example, yrs_education could contain some noise in our estimate, biasing our estimate towards zero. This could be the result of misreporting actual years of education within our sample, which depends on the validity of the data collection methods.

[15] Given a random sample of 7,500 employed individuals in California with income less than one million dollars, we ran various regression models to learn more about our data. Throughout our regression analysis, we found a negative statistically significant relationship between average hours worked per week (hrs_work_perwk) and whether or not an individual is female (i_female). Our first regression indicated that we could expect females to work 3.5665 less hours than non-females, where non-females are expected to work 39.2373 hours on average. Adjusting our model for standard errors to be robust to heteroskedasticity, we ran the regression again and found no change in our estimates, but with slight changes in our standard errors. We then adjusted our model by adding i_kids to indicate whether or not an individual has children. This model showed a positive relationship between work hours and children, indicating that we can expect individuals with kids to work 0.3384 more hours, females to work 3.5604 less hours than non-females, and non-females without children to work 39.1055 hours. We then added an interaction term between i_female and i_kids, which indicated a negative relationship of -1.2118 on the returns on work hours for females with children. From this model we expected non-females with no children to work an average of 38.8903 hours a week, females to work 3.1001 less hours than non-females, and individuals with children to work an additional 0.8904 average hours per week. Adding education level to the model showed a positive relationship between years of education and work hours, where each additional year of education would result in 0.33087 more work hours on average. Females would work 3.23339 less hours than non-females, individuals with children would work 1.15457 more hours than with no children, returns/effects of work hours for females with children to be -1.33416, and non-female individuals with no children and no education would work 34.25727 hours a week on average. All of our models were statistically significant with p-values $< 2.2e\text{-}16$, although the adjusted R-squared values were low which indicates that only a small amount of variance is explained in each model.