ECON 121 FA25 Problem Set 2

Solution

Question 2

If education and experience are exogenous, then β_1 represents the causal effect of education on log wages. The quantitative interpretation is that each additional year of education raises wages by $100 * \beta_1$ percent.

The squared term in experience allows for wages to vary non-linearly with experience. For instance, we might expect wages to rise with experience at a decreasing rate. In this case, β_3 would be positive and beta4 would be negative.

Question 3

The code cleans and summarizes the variables. The mean log wage is 3.13, or approximately $23/hour. Education averages 14.5 years, with a standard deviation of 2.8. Experience averages 24 years, with a standard deviation of 11. 12 percent of the sample is black.

```
# Drop observations with <50 weeks or <35 hours or missing hours
# or 0 income (which will drop when we take logs anyway)
cps18 <-
  cps18 %>%
  filter(hrs_per_wk >= 35 & hrs_per_wk <= 170 & incwage>0 & wkswork>=50)

# Generate new variables
cps18 <- cps18 %>% mutate(lnw = log(incwage/(wkswork*hrs_per_wk)),
                          black = ifelse(race == "black", 1, 0),
                          asian = ifelse(race == "asian/pacific", 1, 0),
                          native = ifelse(race == "native", 1, 0),
                          other = ifelse(race == "multiple/other", 1, 0),
                          edyrs = case_when((ed_lt_hs==1) ~ 6,
                                            (ed_some_hs==1) ~ 10,
                                            (ed_hs_degree==1) ~ 12,
                                            (ed_some_col==1) ~ 14,
                                            (ed_ba_degree==1) ~ 16,
                                            (ed_post_degree==1) ~ 19),
                          exper = age - edyrs - 5,
                          exper2 = exper^2)

# Summarize the new variables
cps18 %>%
  summarise(mean_lnw = mean(lnw,na.rm=TRUE),
            mean_edyrs = mean(edyrs,na.rm=TRUE),
            mean_exper = mean(exper,na.rm=TRUE),
            mean_black = mean(black,na.rm=TRUE),
            mean_asian = mean(asian,na.rm=TRUE),
            mean_native = mean(native,na.rm=TRUE),
            mean_other = mean(other,na.rm=TRUE),
```

```
            mean_male = mean(male,na.rm=TRUE)
  )
```

```
## # A tibble: 1 x 8
##   mean_lnw mean_edyrs mean_exper mean_black mean_asian mean_native mean_other
##      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>       <dbl>      <dbl>
## 1     3.13       14.5       23.8      0.118     0.0780      0.0123     0.0179
## # i 1 more variable: mean_male <dbl>
```

```
cps18 %>%
  summarise(sd_lnw = sd(lnw,na.rm=TRUE),
            sd_edyrs = sd(edyrs,na.rm=TRUE),
            sd_exper = sd(exper,na.rm=TRUE),
            sd_black = sd(black,na.rm=TRUE),
            sd_asian = sd(asian,na.rm=TRUE),
            sd_native = sd(native,na.rm=TRUE),
            sd_other = sd(other,na.rm=TRUE),
            sd_male = sd(male,na.rm=TRUE)
  )
```

```
## # A tibble: 1 x 8
##   sd_lnw sd_edyrs sd_exper sd_black sd_asian sd_native sd_other sd_male
##    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>     <dbl>    <dbl>   <dbl>
## 1  0.732     2.82     11.1    0.322    0.268     0.110    0.133   0.497
```

Question 4

The code estimates the Mincerian regression. The estimated return is 0.110, or an 11 percent wage increase per year of education. It is highly statistically significant, with a t-statistic of nearly 100.

```
feols(lnw ~ edyrs + exper + exper2, data = cps18, vcov = 'hetero')
```

```
## OLS estimation, Dep. Var.: lnw
## Observations: 49,153
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  1.164737   0.022075  52.7624 < 2.2e-16 ***
## edyrs        0.110355   0.001136  97.1727 < 2.2e-16 ***
## exper        0.025744   0.001175  21.9149 < 2.2e-16 ***
## exper2      -0.000362   0.000024 -15.0595 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.665639   Adj. R2: 0.1723
```

Question 5

The code estimates the extended Mincerian regression. The extended Mincerian regression yields an estimated return of 0.114, which is similar to but slightly larger than the original estimate.

```
feols(lnw ~ edyrs + exper + exper2 + male + black + asian + native + other,
      data = cps18,
      vcov = 'hetero')
```

```
## OLS estimation, Dep. Var.: lnw
## Observations: 49,153
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error   t value   Pr(>|t|)
```

```
## (Intercept)  0.996619   0.022100  45.09639  < 2.2e-16 ***
## edyrs        0.114043   0.001124 101.49545  < 2.2e-16 ***
## exper        0.023768   0.001148  20.69725  < 2.2e-16 ***
## exper2      -0.000321   0.000024 -13.65482  < 2.2e-16 ***
## male         0.274881   0.005917  46.45646  < 2.2e-16 ***
## black       -0.163516   0.009336 -17.51380  < 2.2e-16 ***
## asian        0.035559   0.011099   3.20376 1.3573e-03 **
## native      -0.113356   0.026437  -4.28784 1.8076e-05 ***
## other       -0.056254   0.024147  -2.32969 1.9827e-02 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.648408   Adj. R2: 0.214517
```

Question 6

The code saves the extended model using the name 'extended' and then takes a linear combination of the coefficients on `black` and `male`. The female-male wage gap is 0.274 log points, while the black white gap is -0.164 log points, leading to a difference of 0.111, which is significant at less than the 0.1% level. Note that I wrote "black + male" because the coefficient on "male" is the male - female wage gap, and I want the female-male wage gap.

```
extended <- feols(lnw ~ edyrs + exper + exper2 + male + black +
                  asian + native + other,data = cps18, vcov = 'hetero')
hypotheses(extended,"black + male=0")
```

```
##
##            Term Estimate Std. Error    z Pr(>|z|)    S  2.5 % 97.5 %
##  black + male=0    0.111     0.0113 9.82   <0.001 73.3 0.0891  0.134
##
## Columns: term, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
```

Question 7

The code runs separate regressions for men and women using the `split` option. We find a return of 0.110 for men and 0.120 for women, implying a difference in returns of 0.010 log points or 1.0 percent.

To assess whether the difference is significant, we can compute the t-statistic using the coefficients and standard errors. That's because the male and female samples are independent, so the coefficients have no covariance. The code finds a t-statistic of 4.45, so the difference in coefficients is significant at the 5% level.

```
feols(lnw ~ edyrs + exper + exper2 + black + asian + native + other,
      data = cps18,
      vcov = 'hetero',
      split = ~male)
```

```
## Standard-errors: Heteroskedasticity-robust
## Sample: 0
##              Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)  0.978669   0.031749 30.825352  < 2.2e-16 ***
## edyrs        0.120032   0.001692 70.937615  < 2.2e-16 ***
## exper        0.016441   0.001576 10.431250  < 2.2e-16 ***
## exper2      -0.000183   0.000032 -5.648021 1.6432e-08 ***
## black       -0.126304   0.012729 -9.922570  < 2.2e-16 ***
## asian        0.062352   0.015922  3.916096 9.0267e-05 ***
## native      -0.113835   0.036581 -3.111870 1.8615e-03 **
## other       -0.039018   0.040647 -0.959938 3.3710e-01
```

```
## ---
## Sample: 1
##               Estimate Std. Error   t value  Pr(>|t|)
## (Intercept)  1.258817   0.029484  42.69537 < 2.2e-16 ***
## edyrs        0.109961   0.001500  73.30849 < 2.2e-16 ***
## exper        0.030495   0.001646  18.52449 < 2.2e-16 ***
## exper2      -0.000444   0.000034 -13.23807 < 2.2e-16 ***
## black       -0.201566   0.013725 -14.68558 < 2.2e-16 ***
## asian        0.018559   0.015429   1.20283 0.2290528
## native      -0.112180   0.037894  -2.96039 0.0030751 **
## other       -0.069008   0.028264  -2.44152 0.0146321 *
```

```
(0.120032-0.109961)/sqrt(0.001692^2+0.001500^2)
```

```
## [1] 4.4539
```

Question 8

To match the approach in Problem 7, the code allows ALL of the coefficients to vary by gender, so we need many interaction terms. The coefficient on the interaction between education and female is -0.010, with a t-statistic of 4.45, just as in Problem 7!

The code uses the delta method to estimate the ratio of returns, finding that the female/male ratio is .09 above the null hypothesis of 1, i.e. the ratio is 1.09. The p-value on the test is <0.001, so the difference from 1 is statistically significant.

```
cps18 <- cps18 %>% mutate(edyrs_m = edyrs*male,
                          exper_m = exper*male,
                          exper2_m = exper2*male,
                          black_m = black*male,
                          asian_m = asian*male,
                          native_m = native*male,
                          other_m = other*male)
interacted <- feols(lnw ~ edyrs + edyrs_m + exper + exper_m + exper2 + exper2_m +
                    black + black_m + asian + asian_m + native + native_m +
                    other + other_m + male,
                    data = cps18,
                    vcov = 'hetero')
summary(interacted)
```

```
## OLS estimation, Dep. Var.: lnw
## Observations: 49,153
## Standard-errors: Heteroskedasticity-robust
##               Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)  0.978669   0.031748 30.825971  < 2.2e-16 ***
## edyrs        0.120032   0.001692 70.939040  < 2.2e-16 ***
## edyrs_m     -0.010071   0.002261 -4.454021 8.4463e-06 ***
## exper        0.016441   0.001576 10.431459  < 2.2e-16 ***
## exper_m      0.014054   0.002279  6.166370 7.0418e-10 ***
## exper2      -0.000183   0.000032 -5.648134 1.6309e-08 ***
## exper2_m    -0.000261   0.000047 -5.592643 2.2483e-08 ***
## black       -0.126304   0.012729 -9.922770  < 2.2e-16 ***
## black_m     -0.075263   0.018719 -4.020578 5.8142e-05 ***
## asian        0.062352   0.015922  3.916175 9.0086e-05 ***
## asian_m     -0.043793   0.022171 -1.975219 4.8249e-02 *
## native      -0.113835   0.036580 -3.111932 1.8597e-03 **
```

```
## native_m     0.001655   0.052670  0.031431 9.7493e-01
## other       -0.039018   0.040646 -0.959958 3.3708e-01
## other_m     -0.029990   0.049507 -0.605761 5.4468e-01
## male         0.280149   0.043327  6.465861 1.0167e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.647862   Adj. R2: 0.215729
```

```
hypotheses(interacted, "edyrs/(edyrs+edyrs_m)=1")
```

```
##
##                     Term Estimate Std. Error    z Pr(>|z|)    S  2.5 % 97.5 %
##  edyrs/(edyrs+edyrs_m)=1   0.0916     0.0214 4.28   <0.001 15.7 0.0496  0.134
##
## Columns: term, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
```

Question 9

The code summarizes the NLSY dataset. The sample is 25% black and 14% Hispanic, but when we use sampling weights, those shares fall to 14% black and 6% Hispanic. Because the sampling weights undo the NLSY's over-sampling, the latter estimates are representatives of US adults who were teenagers residing in the United States in 1979. The weighted statistics provide unbiased estimates of the population racial composition, since they restore representativeness in the sample.

```
nlsy79 %>% summarise(mean_black = mean(black),
                     wtmean_black = weighted.mean(black, w=perweight),
                     mean_hisp = mean(hisp),
                     wtmean_hisp = weighted.mean(hisp, w=perweight))
```

```
## # A tibble: 1 x 4
##   mean_black wtmean_black mean_hisp wtmean_hisp
##        <dbl>        <dbl>     <dbl>       <dbl>
## 1      0.250        0.139     0.158      0.0630
```

Question 10

The code first keeps full time workers with positive, non-missing earnings, and generates new variables. It then estimates the unweighted and weighted regressions. Estimates of the Mincerian return to education are extremely similar using OLS (0.121) and WLS (0.121). Since theunweighted OLS estimates are more precise (consistent with the Guass Markov theorem), I will continue the analysis with unweighted regressions. (I could have also said that I prefer to have results that are representative of the coefficient I would obtain in the full population, which would have led me to run weighted regressions for the rest of the analysis.

```
nlsy79 <-
  nlsy79 %>%
  filter(hours18 >= 35*50 & laborinc18>0) %>%
  mutate(lnw = log(laborinc18/hours18),
         exper = age79 + 2018-1979 - educ - 5,
         exper2 = exper^2)

feols(lnw ~ educ + exper + exper2 + black + hisp + male, data = nlsy79, vcov='hetero')
```

```
## OLS estimation, Dep. Var.: lnw
## Observations: 3,570
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)  3.436991   0.948333  3.62425 0.00029385 ***
```

```
## educ         0.120598   0.007248 16.63810  < 2.2e-16 ***
## exper       -0.113352   0.047437 -2.38954 0.01692130 *
## exper2       0.001550   0.000619  2.50210 0.01239023 *
## black       -0.261476   0.027426 -9.53392  < 2.2e-16 ***
## hisp        -0.084562   0.029764 -2.84104 0.00452215 **
## male         0.327012   0.022912 14.27277  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.68309   Adj. R2: 0.224944
```

```
feols(lnw ~ educ + exper + exper2 + black + hisp + male, data = nlsy79, vcov='hetero', weights = ~perwe
```

```
## OLS estimation, Dep. Var.: lnw
## Observations: 3,570
## Weights: perweight
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error   t value  Pr(>|t|)
## (Intercept)  2.259575   1.241747  1.819674  0.068892 .
## educ         0.120667   0.008686 13.891922 < 2.2e-16 ***
## exper       -0.051781   0.062216 -0.832277  0.405309
## exper2       0.000735   0.000814  0.903747  0.366191
## black       -0.256713   0.029351 -8.746226 < 2.2e-16 ***
## hisp        -0.074574   0.032056 -2.326364  0.020055 *
## male         0.390222   0.027136 14.380513 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 392.9   Adj. R2: 0.229852
```

Question 11

The extended Mincerian regression yields an estimated return to education of 12 percent, similar to the CPS. However, while the CPS had a significant positive coefficient on exper and a significant negative coefficient on `exper2`, these coefficients have the opposite signin the NLSY. The difference is likely because NLSY respondents are quite old in 2018, with substantial potential experience. The CPS indicated that wages rise with experience at labor market entry but then flatten out (due to the negative squared term). Since the NLSY in 2007 only had mature workers, the dataset is not well-suited for estimating the returns to experience. Workers may already be retiringtoo.

Question 12

It seems unlikely that the coefficient on education represents the causal effect of education. Education and wages are likely to be correlated with a number of omitted variables, such as innate ability and parental socioeconomic status.

Question 13

To address the concerns above, we can control for childhood background characteristics and cognitive test scores. Doing so reduces the estimated return to education substantially, to 6 percent. Note that I save the estimated model as `long_model` and then report it using `summary()`. This is just so all coefficients are reported in the output. `feols()` didn't automatically include all of them

```
long_model <-
  feols(lnw ~ educ + exper + exper2 + black + hisp + male +
            foreign + urban14 + mag14 + news14 + lib14 +
            educ_mom + educ_dad + numsibs + afqt81,
```

```
      data = nlsy79,
      vcov = 'hetero')
```

```
## NOTE: 640 observations removed because of NA values (RHS: 640).
```

```
summary(long_model)
```

```
## OLS estimation, Dep. Var.: lnw
## Observations: 2,930
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)  5.032151   0.967352  5.201985 2.1083e-07 ***
## educ         0.055994   0.008758  6.393788 1.8782e-10 ***
## exper       -0.167154   0.047604 -3.511312 4.5267e-04 ***
## exper2       0.002004   0.000620  3.230152 1.2510e-03 **
## black       -0.020563   0.035342 -0.581841 5.6072e-01
## hisp         0.076085   0.046287  1.643769 1.0033e-01
## male         0.313766   0.024981 12.560129  < 2.2e-16 ***
## foreign      0.065173   0.041475  1.571368 1.1621e-01
## urban14      0.038455   0.031036  1.239047 2.1543e-01
## mag14        0.070789   0.026952  2.626501 8.6719e-03 **
## news14       0.044962   0.032916  1.365979 1.7205e-01
## lib14        0.012623   0.030347  0.415936 6.7749e-01
## educ_mom     0.007007   0.005795  1.209161 2.2670e-01
## educ_dad     0.007907   0.004620  1.711665 8.7065e-02 .
## numsibs      0.001176   0.005568  0.211246 8.3271e-01
## afqt81       0.006667   0.000627 10.625686  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.656832   Adj. R2: 0.274567
```

However, note that the sample size changed because we do not have all control variables for all observations. We should reestimate the "short" model in the smaller sample to make sure the change in coefficients is not due to sample composition. (This was not essential for full credit, but it is good practice.)

```
nlsy79_subsample <-
  nlsy79 |>
  drop_na(c("foreign","urban14","mag14","news14","lib14",
            "educ_mom","educ_dad","numsibs","afqt81"))
feols(lnw ~ educ + exper + exper2 + black + hisp + male,
      data = nlsy79_subsample,
      vcov = 'hetero')
```

```
## OLS estimation, Dep. Var.: lnw
## Observations: 2,930
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)  3.409089   1.011224  3.37125 7.5801e-04 ***
## educ         0.120081   0.007857 15.28424  < 2.2e-16 ***
## exper       -0.109352   0.050531 -2.16405 3.0541e-02 *
## exper2       0.001467   0.000662  2.21572 2.6787e-02 *
## black       -0.244225   0.031175 -7.83389 6.5660e-15 ***
## hisp        -0.060360   0.032091 -1.88093 6.0081e-02 .
## male         0.347698   0.025152 13.82406  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## RMSE: 0.677306   Adj. R2: 0.231013
```

Why did the return fall when we included additional covariates? It appears that urban residence, paternal education and AFQT scores (a measure of innate ability) are all positively related with wages, and it is likely that they also predict higher education. (You can check this.)