ECON 121 FA23 Problem Set 4

Solution

Question 2

Summary statistics appear below. 21 percent of the sample participated in HS. 32 percent of the sample is black, and 20 percent is Hispanic. Average mother's education is 12 years. 3 in 10 repeat a grade, another 3 in 10 go to college, and 7 in 10 graduate high school. Also worthy of note is the number of NA values, which is very high for ppvt_3. This high level of "missingness" will be important later.

```
summary(nlsy_kids)
```

```
##    head_start        sibdiff           mom_id         hispanic
##  Min.   :0.0000   Min.   :0.0000   Min.   :    3   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 3448   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median : 6400   Median :0.0000
##  Mean   :0.2066   Mean   :0.2321   Mean   : 6227   Mean   :0.2005
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.: 8870   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :12667   Max.   :1.0000
##
##      black             male          firstborn        lninc_0to3
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 3.909
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 9.586
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :10.118
##  Mean   :0.3203   Mean   :0.5097   Mean   :0.4045   Mean   :10.070
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:10.584
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :13.423
##                                                     NA's   :218
##      momed       dadhome_0to3       ppvt_3            lnbw
##  Min.   : 1.0   Min.   :0.000   Min.   :  0.00   Min.   :1.792
##  1st Qu.:10.0   1st Qu.:0.250   1st Qu.: 12.00   1st Qu.:4.635
##  Median :12.0   Median :1.000   Median : 19.00   Median :4.745
##  Mean   :11.7   Mean   :0.678   Mean   : 21.88   Mean   :4.718
##  3rd Qu.:13.0   3rd Qu.:1.000   3rd Qu.: 30.00   3rd Qu.:4.852
##  Max.   :20.0   Max.   :1.000   Max.   :101.00   Max.   :5.434
##  NA's   :6      NA's   :1603    NA's   :3591     NA's   :145
##  comp_score_5to6 comp_score_7to10 comp_score_11to14    repgrade
##  Min.   : 0.00   Min.   : 0.00    Min.   : 0.6667   Min.   :0.0000
##  1st Qu.:29.50   1st Qu.:26.00    1st Qu.:23.5000   1st Qu.:0.0000
##  Median :44.50   Median :45.00    Median :42.6667   Median :0.0000
##  Mean   :45.42   Mean   :45.19    Mean   :43.7758   Mean   :0.3158
##  3rd Qu.:62.38   3rd Qu.:63.92    3rd Qu.:62.0000   3rd Qu.:1.0000
##  Max.   :98.50   Max.   :99.00    Max.   :99.0000   Max.   :1.0000
##  NA's   :1845    NA's   :1019     NA's   :1384      NA's   :1026
##     learndis           hsgrad          somecoll           idle
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :1.0000   Median :0.0000   Median :0.0000
```

```
##  Mean   :0.04102   Mean   :0.7152   Mean   :0.3152   Mean   :0.1591
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  NA's   :121       NA's   :1077     NA's   :1077     NA's   :1078
##     fphealth
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.0988
##  3rd Qu.:0.0000
##  Max.   :1.0000
##  NA's   :1077
```

The question asks about the backgrounds of kids who participated in HS. HS participants are more likely to be black, have lower family income, and have less educated mothers, on average. They are also more likely to repeat a grade and less likely to go to college. However, these differences in long-term outcomes may reflect selection bias rather than the effects of HS. In other words, HS participants may have worse outcomes because they come from disadvantaged backgrounds.

```
nlsy_kids |>
    group_by(head_start) |>
    summarize(black = mean(black, na.rm = TRUE),
              lninc_0to3 = mean(lninc_0to3, na.rm = TRUE),
              momed = mean(momed, na.rm = TRUE),
              repgrade = mean(repgrade, na.rm = TRUE),
              somecoll = mean(somecoll, na.rm = TRUE))
```

```
## # A tibble: 2 x 6
##   head_start black lninc_0to3 momed repgrade somecoll
##        <dbl> <dbl>      <dbl> <dbl>    <dbl>    <dbl>
## 1          0 0.269      10.1   11.8    0.289    0.329
## 2          1 0.518       9.78  11.5    0.407    0.269
```

Question 3

Average scores are 5.8 points lower for participants than for non-participants. The association is highly statistically significant and represents roughly one-quarter of a standard deviation in test scores. If we assumed participation is exogenous, then we would conclude that HS reduces test scores by one- quarter of a standard deviation on average. However, we already know that participation is associated with several background characteristics that are likely to have independent effects on test scores, which implies that the residual is correlated with HS participation. As a result, participation is not exogenous, and we should not interpret the association as a causal effect. The bias is probably negative, since disadvantaged families select into HS, and kids from disadvantaged families may tend to have worse long-term outcomes.

```
# Run an OLS regression of the age 5-6 test score on the HS indicator,
# clustering standard errors by mom_id.
feols(comp_score_5to6 ~ head_start,
      data = nlsy_kids,
      vcov = ~mom_id)
```

```
## NOTE: 1,845 observations removed because of NA values (LHS: 1,845).
```

```
## OLS estimation, Dep. Var.: comp_score_5to6
## Observations: 2,420
## Standard-errors: Clustered (mom_id)
##             Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 46.65384   0.616964 75.61845  < 2.2e-16 ***
```

```
## head_start  -5.84207   1.209494 -4.83018 1.5113e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 22.2   Adj. R2: 0.010934
```

```
# For reference, compute the standard deviation of the test score.
sd(nlsy_kids$comp_score_5to6, na.rm=TRUE)
```

```
## [1] 22.37593
```

Question 4

The estimated coefficient on HS participation is now even more negative than the one from question 3. That is consistent with family-level omitted variables: kids from disadvantaged families enroll in HS, and they have have lower average test scores due to their disadvantage.

```
# First create a data frame of families instead of kids. We can do so
# using group_by(), as follows:
nlsy_families <-
  nlsy_kids |>
  drop_na(comp_score_5to6, head_start) |>
  group_by(mom_id) |>
  summarise(mean_test = mean(comp_score_5to6),
            mean_head_start = mean(head_start))

# Now estimate OLS using the family averages
feols(mean_test ~ mean_head_start,
      data = nlsy_families,
      vcov = 'hetero')
```

```
## OLS estimation, Dep. Var.: mean_test
## Observations: 1,426
## Standard-errors: Heteroskedasticity-robust
##                 Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)     47.26384   0.622140 75.96982  < 2.2e-16 ***
## mean_head_start -7.58640   1.366079 -5.55341 3.3379e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 20.0   Adj. R2: 0.018928
```

Question 5

The fixed effect model suggests that HS participation raises test scores, in contrast to the negative effects suggested by OLS and the between effect model. The likely reason is that between-family variation in HS participation is correlated with family disadvantage, which biases us toward finding a negative association in the pooled and between effect models. The full-sample fixed effect model without controls indicates that HS raises test scores by 7.6 points, or one-third of a SD, on average.

```
# Estimate the individual-level model with mother fixed effects.
feols(comp_score_5to6 ~ head_start | mom_id, data = nlsy_kids)
```

```
## NOTE: 1,845 observations removed because of NA values (LHS: 1,845).
```

```
## OLS estimation, Dep. Var.: comp_score_5to6
## Observations: 2,420
## Fixed-effects: mom_id: 1,426
## Standard-errors: Clustered (mom_id)
```

```
##            Estimate Std. Error t value   Pr(>|t|)
## head_start  7.63285    2.01362  3.7906 0.00015655 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 10.7     Adj. R2: 0.442754
##              Within R2: 0.016246
```

Question 6

In the fixed effect regression, we can include child-level covariates only. We cannot control for any family-level variables that do not vary between siblings. I choose male, firstborn, lninc_0to3, dadhome_0to3, and lnbw as covariates. I do not use ppvt_3 because it is available for few observations. When I include it, the sample shrinks and changes composition a lot. This was a judgment call, and you could have done it differently. As researchers, we often face tradeoffs between having more information (by controlling for PPVT) and maintaining the composition of the sample (by not controlling for PPVT).

```
feols(comp_score_5to6 ~ head_start + male + firstborn + lninc_0to3 +
                        dadhome_0to3 + lnbw | mom_id,
      data = nlsy_kids)
```

```
## NOTE: 2,370 observations removed because of NA values (LHS: 1,845, RHS: 1,732).
```

```
## OLS estimation, Dep. Var.: comp_score_5to6
## Observations: 1,895
## Fixed-effects: mom_id: 1,251
## Standard-errors: Clustered (mom_id)
##              Estimate Std. Error   t value Pr(>|t|)
## head_start    5.64711    2.35257  2.400400 0.016523 *
## male         -2.81106    1.27581 -2.203352 0.027752 *
## firstborn     1.66089    1.17064  1.418783 0.156212
## lninc_0to3    2.27392    1.73535  1.310356 0.190316
## dadhome_0to3 -3.26060    3.27771 -0.994781 0.320035
## lnbw          6.91016    3.42362  2.018376 0.043765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 9.48045     Adj. R2: 0.492933
##                 Within R2: 0.029561
```

The estimate is still positive and statistically significant, but it is slightly smaller, in magnitude: HS participation raises test scores by 5.6 points on average. It is useful to check whether this is due to omitted variable bias or the different composition of the subsample with non-missing covariates. I re-estimate the model with no pre-HS covariates, but this time using the sub-sample with non-missing covariates. This was not necessary for full credit, but it is good practice.

```
nlsy_kids_subsample <-
  nlsy_kids %>%
  drop_na(male, firstborn, lninc_0to3, dadhome_0to3, lnbw)

feols(comp_score_5to6 ~ head_start | mom_id,
      data = nlsy_kids_subsample)
```

```
## NOTE: 638 observations removed because of NA values (LHS: 638).
```

```
## OLS estimation, Dep. Var.: comp_score_5to6
## Observations: 1,895
## Fixed-effects: mom_id: 1,251
## Standard-errors: Clustered (mom_id)
```

```
##            Estimate Std. Error t value Pr(>|t|)
## head_start    5.971     2.3642 2.52559 0.011673 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 9.5773     Adj. R2: 0.486544
##                Within R2: 0.009632
```

The coefficient on HS is much closer to the regression with pre-HS covariates. This suggest that within-family OVB is *NOT* the issue, but rather that individuals with missing data on covariates have larger effects. The estimates are robust to controlling for pre-HS covariates.

Question 7

Standardize the outcome variables by subtracting mean and dividing by SD. The scale() function in R does this in one step:

```
nlsy_kids <-
  nlsy_kids %>%
  mutate(std_5to6 = scale(comp_score_5to6),
         std_7to10 = scale(comp_score_7to10),
         std_11to14 = scale(comp_score_11to14))
```

You were not expected to know this function. You could have also used `comp_score_5to6 - mean(comp_score_5to6, na.rm = TRUE))/sd(comp_score_5to6, na.rm = TRUE)`, etc., which would generate exactly the same variables.

Now we run a FE regression of each standardized score on HS participation,
finding that the estimated effects shrink as children get older. HS raises scores by 0.34 standard deviations on average at ages 5-6, by 0.16 standard deviations at ages 7-10, and by 0.15 standard deviations at ages 11 to 14.

```
feols(std_5to6 ~ head_start | mom_id,
      data = nlsy_kids)
```

```
## NOTE: 1,845 observations removed because of NA values (LHS: 1,845).
```

```
## OLS estimation, Dep. Var.: std_5to6
## Observations: 2,420
## Fixed-effects: mom_id: 1,426
## Standard-errors: Clustered (mom_id)
##            Estimate Std. Error t value   Pr(>|t|)
## head_start 0.341119   0.089991  3.7906 0.00015655 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.478179     Adj. R2: 0.442754
##                  Within R2: 0.016246
```

```
feols(std_7to10 ~ head_start | mom_id,
      data = nlsy_kids)
```

```
## NOTE: 1,019 observations removed because of NA values (LHS: 1,019).
```

```
## OLS estimation, Dep. Var.: std_7to10
## Observations: 3,246
## Fixed-effects: mom_id: 1,546
## Standard-errors: Clustered (mom_id)
##            Estimate Std. Error t value Pr(>|t|)
## head_start 0.159245    0.06204 2.56682 0.010357 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.526513     Adj. R2: 0.470368
##                  Within R2: 0.004229
```

```
feols(std_11to14 ~ head_start | mom_id,
      data = nlsy_kids)
```

```
## NOTE: 1,384 observations removed because of NA values (LHS: 1,384).
```

```
## OLS estimation, Dep. Var.: std_11to14
## Observations: 2,881
## Fixed-effects: mom_id: 1,346
## Standard-errors: Clustered (mom_id)
##            Estimate Std. Error t value Pr(>|t|)
## head_start 0.153001    0.06088 2.51317 0.012081 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.511791     Adj. R2: 0.508071
##                  Within R2: 0.004263
```

You may notice that the sample changes across regressions due to missingness. You could have also held the sample constant, as we did above for adding covariates. The effect on the test score at age 5-6 is still largest, although the relative effect sizes at 7-10 and 11-14 are flipped.

```
nlsy_kids_subsample <-
  nlsy_kids |>
  drop_na(std_5to6, std_7to10, std_11to14)

feols(std_5to6 ~ head_start | mom_id,
      data = nlsy_kids_subsample)
```

```
## OLS estimation, Dep. Var.: std_5to6
## Observations: 1,728
## Fixed-effects: mom_id: 1,021
## Standard-errors: Clustered (mom_id)
##            Estimate Std. Error t value  Pr(>|t|)
## head_start 0.321301   0.103263 3.11147 0.0019133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.472877     Adj. R2: 0.449221
##                  Within R2: 0.014944
```

```
feols(std_7to10 ~ head_start | mom_id,
      data = nlsy_kids_subsample)
```

```
## OLS estimation, Dep. Var.: std_7to10
## Observations: 1,728
## Fixed-effects: mom_id: 1,021
## Standard-errors: Clustered (mom_id)
##            Estimate Std. Error  t value Pr(>|t|)
## head_start 0.091516   0.095592 0.957356  0.33861
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.437615     Adj. R2: 0.536713
##                  Within R2: 0.001435
```

```
feols(std_11to14 ~ head_start | mom_id,
      data = nlsy_kids_subsample)
```

```
## OLS estimation, Dep. Var.: std_11to14
## Observations: 1,728
## Fixed-effects: mom_id: 1,021
## Standard-errors: Clustered (mom_id)
##            Estimate Std. Error t value Pr(>|t|)
## head_start 0.182914   0.101884 1.79531   0.0729 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.443297     Adj. R2: 0.524396
##                  Within R2: 0.005564
```

Question 8

We run FE regressions for longer-term outcomes. We find that HS participation reduces grade repetition by 5 percentage points, reduces learning disability diagnosis by 4 percentage points, raises high school graduation by 13 percentage points, raises college attendance by 7 percentage points, reduces idleness (not working or studying) by 7 percentage points, and reduces fair/poor health by 7 percentage points. All of these results but one (for grade repetition) are significant at the 5 percent level. The grade repetition result is significant at the 9 percent level.

```
feols(repgrade ~ head_start | mom_id,
      data = nlsy_kids)
```

```
## NOTE: 1,026 observations removed because of NA values (LHS: 1,026).
```

```
## OLS estimation, Dep. Var.: repgrade
## Observations: 3,239
## Fixed-effects: mom_id: 1,450
## Standard-errors: Clustered (mom_id)
##             Estimate Std. Error  t value Pr(>|t|)
## head_start -0.054403   0.031696 -1.71642 0.086299 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.293357     Adj. R2: 0.278761
##                  Within R2: 0.001695
```

```
feols(learndis ~ head_start | mom_id,
      data = nlsy_kids)
```

```
## NOTE: 121 observations removed because of NA values (LHS: 121).
```

```
## OLS estimation, Dep. Var.: learndis
## Observations: 4,144
## Fixed-effects: mom_id: 1,714
## Standard-errors: Clustered (mom_id)
##             Estimate Std. Error  t value  Pr(>|t|)
## head_start -0.037349   0.013224 -2.82444 0.0047912 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.144667     Adj. R2: 0.092616
##                  Within R2: 0.003505
```

```
feols(hsgrad ~ head_start | mom_id,
       data = nlsy_kids)
```

```
## NOTE: 1,077 observations removed because of NA values (LHS: 1,077).

## OLS estimation, Dep. Var.: hsgrad
## Observations: 3,188
## Fixed-effects: mom_id: 1,367
## Standard-errors: Clustered (mom_id)
##            Estimate Std. Error t value   Pr(>|t|)
## head_start 0.131179   0.030895 4.24594 2.3239e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.31008     Adj. R2: 0.17344
##                 Within R2: 0.009208
```

```
feols(somecoll ~ head_start | mom_id,
      data = nlsy_kids)
```

```
## NOTE: 1,077 observations removed because of NA values (LHS: 1,077).

## OLS estimation, Dep. Var.: somecoll
## Observations: 3,188
## Fixed-effects: mom_id: 1,367
## Standard-errors: Clustered (mom_id)
##            Estimate Std. Error t value Pr(>|t|)
## head_start 0.073996   0.030749 2.40648 0.016239 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.310531     Adj. R2: 0.217764
##                  Within R2: 0.00294
```

```
feols(idle ~ head_start | mom_id,
      data = nlsy_kids)
```

```
## NOTE: 1,078 observations removed because of NA values (LHS: 1,078).

## OLS estimation, Dep. Var.: idle
## Observations: 3,187
## Fixed-effects: mom_id: 1,367
## Standard-errors: Clustered (mom_id)
##             Estimate Std. Error  t value Pr(>|t|)
## head_start -0.072788   0.031397 -2.31828 0.020581 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.263083     Adj. R2: 0.093811
##                  Within R2: 0.003961
```

```
feols(fphealth ~ head_start | mom_id,
      data = nlsy_kids)
```

```
## NOTE: 1,077 observations removed because of NA values (LHS: 1,077).

## OLS estimation, Dep. Var.: fphealth
## Observations: 3,188
## Fixed-effects: mom_id: 1,367
## Standard-errors: Clustered (mom_id)
##             Estimate Std. Error  t value  Pr(>|t|)
## head_start -0.065942   0.023907 -2.75822 0.0058891 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## RMSE: 0.224664     Adj. R2: 0.007413
##                  Within R2: 0.004454
```

Question 9

The easiest way to test for heterogeneous effects by race, ethnicity, and sex is include interactions of the HS dummy with race, ethnicity, and sex dummies. We also need to control for the main effect of sex, but not for the main effects or race and ethnicity because they are collinear with the mother fixed effects. I do this below for the high school graduation outcome. The results do not show strong evidence of heterogeneity in effects by race, ethnicity, or sex. The coefficients on the interaction terms are large, but none are significant at the 5% level.

```
# Here I use R's nice approach to interaction terms, but you could have also
# directly generated new variables for the interaction terms.
feols(hsgrad ~ head_start*(hispanic + black) | mom_id,
      data = nlsy_kids)
```

```
## NOTE: 1,077 observations removed because of NA values (LHS: 1,077).
```

```
## The variables 'hispanic' and 'black' have been removed because of collinearity (see $collin.var).
```

```
## OLS estimation, Dep. Var.: hsgrad
## Observations: 3,188
## Fixed-effects: mom_id: 1,367
## Standard-errors: Clustered (mom_id)
##                      Estimate Std. Error  t value Pr(>|t|)
## head_start           0.059740   0.077431 0.771528  0.44053
## head_start:hispanic 0.066048   0.096913 0.681520  0.49566
## head_start:black    0.100056   0.087448 1.144178  0.25275
## ... 2 variables were removed because of collinearity (hispanic and black)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.30996     Adj. R2: 0.173172
##                 Within R2: 0.009976
```

Question 10

The evidence suggests that HS participation has lasting effects on children's outcomes, which provides some justification for the program's existence. Whether the government should expand or cut funding for this and similar programs depends on its cost-effectiveness compared with other potential uses of funds. In general, it is difficult to extrapolate the effects of program expansion from our estimated average effects of treatment on the treated because the effects may be different in the new subpopulations that would gain access if the program expanded. At the same time, the lack of significant treatment effect heterogeneity in Problem 8 suggests that perhaps we can extrapolate. Many answers could receive full credit for this question.