ECON 121 FA23 Problem Set 4

Stephanie Nguyen

Question 1

Verbal: list group members.

Stephanie Nguyen (A16215540)

Robert Tso (A13829791)

Tomas Lopez (A16798775)

Akash Juwadi (A16372772)

I discussed the problems with them but did not work on the code explicitly with them.

Question 2

Code: Load packages and dataset, summarize data.

Verbal: Interpret the summary statistics.

```
# The PDF will show the code you write here but not the output.
# Load packages and dataset here.

library(tidyverse)
library(fixest)
library(mfx)
library(car)

load(url("https://github.com/tvogl/econ121/raw/main/data/nlsy_kids.Rdata"))
```

```
# The PDF will show the code AND output here.
# Summarize the data here.

background_headstart <-
  filter(nlsy_kids, head_start == 1)

summary(background_headstart)
```

```
##    head_start    sibdiff           mom_id         hispanic
##  Min.   :1    Min.   :0.0000   Min.   :  204   Min.   :0.0000
##  1st Qu.:1    1st Qu.:0.0000   1st Qu.: 4483   1st Qu.:0.0000
##  Median :1    Median :1.0000   Median : 7126   Median :0.0000
##  Mean   :1    Mean   :0.5267   Mean   : 6647   Mean   :0.1884
##  3rd Qu.:1    3rd Qu.:1.0000   3rd Qu.: 9048   3rd Qu.:0.0000
##  Max.   :1    Max.   :1.0000   Max.   :12667   Max.   :1.0000
##
##      black             male          firstborn        lninc_0to3
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 6.257
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 9.370
##  Median :1.0000   Median :1.0000   Median :0.0000   Median : 9.745
##  Mean   :0.5176   Mean   :0.5346   Mean   :0.4234   Mean   : 9.784
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:10.191
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :13.299
##                                                     NA's   :46
##      momed       dadhome_0to3        ppvt_3           lnbw
##  Min.   : 3.0   Min.   :0.0000   Min.   : 1.00   Min.   :3.178
##  1st Qu.:10.0   1st Qu.:0.0000   1st Qu.:10.00   1st Qu.:4.632
##  Median :12.0   Median :0.5000   Median :16.00   Median :4.736
##  Mean   :11.5   Mean   :0.5042   Mean   :18.13   Mean   :4.711
##  3rd Qu.:13.0   3rd Qu.:1.0000   3rd Qu.:23.00   3rd Qu.:4.836
##  Max.   :20.0   Max.   :1.0000   Max.   :67.00   Max.   :5.434
##                 NA's   :388      NA's   :747     NA's   :21
##  comp_score_5to6 comp_score_7to10 comp_score_11to14     repeat
##  Min.   : 1.50   Min.   : 1.00    Min.   : 1.00     Min.   :0.0000
##  1st Qu.:25.67   1st Qu.:19.00    1st Qu.:18.00     1st Qu.:0.0000
##  Median :39.42   Median :37.00    Median :34.00     Median :0.0000
##  Mean   :40.81   Mean   :38.19    Mean   :36.19     Mean   :0.4073
##  3rd Qu.:53.88   3rd Qu.:54.50    3rd Qu.:51.00     3rd Qu.:1.0000
```

```
##  Max.   :98.00   Max.   :99.00    Max.   :98.00     Max.   :1.0000
##  NA's   :371     NA's   :142      NA's   :192       NA's   :137
##     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   :0.04205   Mean   :0.7184   Mean   :0.2692   Mean   :0.1923
##  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   :1         NA's   :153      NA's   :153      NA's   :153
##     fphealth
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.09615
##  3rd Qu.:0.00000
##  Max.   :1.00000
##  NA's   :153
```

```
background_noheadstart <-
  filter(nlsy_kids, head_start == 0)

summary(background_noheadstart)
```

```
##    head_start    sibdiff           mom_id         hispanic
##  Min.   :0    Min.   :0.0000   Min.   :    3   Min.   :0.0000
##  1st Qu.:0    1st Qu.:0.0000   1st Qu.: 3259   1st Qu.:0.0000
##  Median :0    Median :0.0000   Median : 6282   Median :0.0000
##  Mean   :0    Mean   :0.1554   Mean   : 6118   Mean   :0.2036
##  3rd Qu.:0    3rd Qu.:0.0000   3rd Qu.: 8846   3rd Qu.:0.0000
##  Max.   :0    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.: 9.679
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :10.219
##  Mean   :0.2689   Mean   :0.5033   Mean   :0.3995   Mean   :10.144
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:10.654
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :13.423
##                                                     NA's   :172
##      momed        dadhome_0to3        ppvt_3           lnbw
##  Min.   : 1.00   Min.   :0.0000   Min.   :  0.00   Min.   :1.792
##  1st Qu.:10.00   1st Qu.:0.3333   1st Qu.: 13.00   1st Qu.:4.635
##  Median :12.00   Median :1.0000   Median : 20.00   Median :4.749
##  Mean   :11.75   Mean   :0.7175   Mean   : 22.81   Mean   :4.720
##  3rd Qu.:13.00   3rd Qu.:1.0000   3rd Qu.: 32.00   3rd Qu.:4.852
##  Max.   :20.00   Max.   :1.0000   Max.   :101.00   Max.   :5.252
##  NA's   :6       NA's   :1215     NA's   :2844     NA's   :124
##  comp_score_5to6 comp_score_7to10 comp_score_11to14     repeat
##  Min.   : 0.00   Min.   : 0.00    Min.   : 0.6667   Min.   :0.0000
##  1st Qu.:30.33   1st Qu.:28.17    1st Qu.:25.6250   1st Qu.:0.0000
##  Median :45.67   Median :48.00    Median :45.6667   Median :0.0000
##  Mean   :46.65   Mean   :47.26    Mean   :46.1591   Mean   :0.2886
##  3rd Qu.:63.50   3rd Qu.:66.42    3rd Qu.:65.0833   3rd Qu.:1.0000
```

```
##  Max.   :98.50   Max.   :99.00    Max.   :99.0000   Max.   :1.0000
##  NA's   :1474    NA's   :877      NA's   :1192      NA's   :889
##     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   :0.04075   Mean   :0.7142   Mean   :0.3289   Mean   :0.1492
##  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   :120       NA's   :924      NA's   :924      NA's   :925
##     fphealth
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.0996
##  3rd Qu.:0.0000
##  Max.   :1.0000
##  NA's   :924
```

For children who completed Head Start, 19% of children are Hispanic and 52% of children are Black. In addition, for those who participated in Head Start, the mom's average education is 11.5 years. Only 20% of Hispanic children and 27% of Black children in the data set did not participate in Head Start. For children who did not participate in Head Start, the mom's average education is 11.75 years. 41% of children who participated in Head Start repeated a grade whereas 29% of children who did not participate in Head Start repeated a grade.

Question 3

Code: Regression.

Verbal: Interpret.

```
# All question 3 code here.

nlsy_subset_5to6scores <- nlsy_kids %>% drop_na(comp_score_5to6)
feols(comp_score_5to6 ~ head_start, data = nlsy_subset_5to6scores, vcov = ~mom_id)
```

```
## 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
```

```
-5.84207/sd(nlsy_subset_5to6scores$comp_score_5to6)
```

```
## [1] -0.2610873
```

```
# more conservative estimate when we include clustered standard errors
# rather than independent standard errors

# I decided to cluster based on mom_id because I assumed that the children with
# the same mother are not independent. However, I assumed that children without
# the same mother are independent.

# if y is standardized, then we can talk about the coefficient in terms of
# standard deviations
```

If we assume Head Start is exogenous, children who participate in Head Start on average score 5.8 points less than children who do not participate in Head Start. Children who participate in Head Start score 0.26 standard deviations lower than children who do not participate in Head Start. It is not reasonable to assume that Head Start participation is exogenous because it might be correlated with other omitted variables which also affect test scores. For example, Head Start participation might be correlated with family income level and family income level could also affect test scores. I think there could also be omitted variables at the family level. For example, family income level and mother's education could be omitted variables. I think that it might bias the estimated coefficient downwards because we have not accounted for these background factors in the regression which could negatively impact test scores.

Question 4

Code: Regression.

Verbal: Interpret.

```
# All question 4 code here

nlsy_families <-
  nlsy_kids %>%
  drop_na(comp_score_5to6, head_start) %>%
  group_by(mom_id) %>%
  summarise(mean_headstart = mean(head_start), # average participation rate of
            # Head Start for different families
            mean_test_score = mean(comp_score_5to6))

feols(mean_test_score ~ mean_headstart, data = nlsy_families, vcov = "hetero")
```

```
## OLS estimation, Dep. Var.: mean_test_score
## 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_headstart -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
```

The estimated coefficient is more negative than in question 3. In families where all children participate in Head Start, test scores on average are 7.6 points lower than families where all children do not participate in Head Start. The standard errors are also larger than in question 3.

Question 5

Code: Regression.

Verbal: Interpret.

```
# All question 5 code here

feols(comp_score_5to6 ~ head_start | mom_id, data = nlsy_subset_5to6scores)
```

```
## 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
```

```
7.63285 / sd(nlsy_subset_5to6scores$comp_score_5to6)
```

```
## [1] 0.3411188
```

My results imply that if we look within families, on average, the sibling who participated in Head Start scores 7.6 points higher than siblings who did not participate in Head Start. The sibling who participated in Head Start scores 0.34 standard deviations higher than the sibling who did not participate in Head Start. My fixed effect results are different from questions 3 and 4 because the coefficient on fixed effects represents an association between Head Start and test score within each family. However, the coefficients on 3 and 4 did not look within each family, but rather differentiated only based on Head Start participation. The mother fixed effects controls for everything that varies across moms, therefore I think estimate with fixed effects most likely reflects the effect of Head Start participation on test scores.

Question 6

Code: Regression.

Verbal: Interpret.

```
# All question 6 code here

nlsy_subset <- nlsy_kids %>% drop_na(comp_score_5to6, head_start, firstborn,
                                     male, lnbw)

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

```
## OLS estimation, Dep. Var.: comp_score_5to6
## Observations: 2,350
## Fixed-effects: mom_id: 1,408
## Standard-errors: Clustered (mom_id)
##            Estimate Std. Error  t value   Pr(>|t|)
## head_start  7.22045   2.023100  3.56901 0.00037038 ***
## firstborn   2.23081   0.924451  2.41312 0.01594394 *
## male       -3.26386   1.049235 -3.11071 0.00190389 **
## lnbw        7.32766   2.761860  2.65316 0.00806396 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 10.5     Adj. R2: 0.449946
##              Within R2: 0.038132
```

I chose to include being first-born, male, and birth weight. I think being first-born might be correlated with higher test scores because they might get more attention from their parents and therefore are more likely to go to join Head Start and also have higher test scores. I also chose male and birth weight because birth weight from our class example seemed to have an effect on test scores. The inclusion of the covariates did not change the estimated coefficient on Head Start since it is approximately 7 points. I conclude that the fixed effects estimate of the effect of Head Start is robust since the coefficient did not change drastically after adding in the covariates.

Question 7

Code: Regressions.

Verbal: Interpret.

```
# All question 7 code here

nlsy_subset_scores <- nlsy_kids %>% drop_na(comp_score_5to6, comp_score_7to10,
                                            comp_score_11to14)

nlsy_subset_scores <-
  mutate(nlsy_subset_scores,
         standard_score_5to6 =
           (comp_score_5to6 - mean(comp_score_5to6)) / (sd(comp_score_5to6)),
         standard_score_7to10 =
           (comp_score_7to10 - mean(comp_score_7to10)) / (sd(comp_score_7to10)),
         standard_score_11to14 =
           (comp_score_11to14 - mean(comp_score_11to14)) / (sd(comp_score_11to14)))

feols(standard_score_5to6 ~ head_start | mom_id, data = nlsy_subset_scores)
```

```
## OLS estimation, Dep. Var.: standard_score_5to6
## Observations: 1,728
## Fixed-effects: mom_id: 1,021
## Standard-errors: Clustered (mom_id)
##            Estimate Std. Error t value  Pr(>|t|)
## head_start 0.322317    0.10359 3.11147 0.0019133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.474372     Adj. R2: 0.449221
##                  Within R2: 0.014944
```

```
feols(standard_score_7to10 ~ head_start | mom_id, data = nlsy_subset_scores)
```

```
## OLS estimation, Dep. Var.: standard_score_7to10
## Observations: 1,728
## Fixed-effects: mom_id: 1,021
## Standard-errors: Clustered (mom_id)
##            Estimate Std. Error  t value Pr(>|t|)
## head_start 0.090983   0.095035 0.957356  0.33861
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.435066     Adj. R2: 0.536713
##                  Within R2: 0.001435
```

```
feols(standard_score_11to14 ~ head_start | mom_id, data = nlsy_subset_scores)
```

```
## OLS estimation, Dep. Var.: standard_score_11to14
## Observations: 1,728
## Fixed-effects: mom_id: 1,021
## Standard-errors: Clustered (mom_id)
```

```
##            Estimate Std. Error t value Pr(>|t|)
## head_start 0.181888   0.101313 1.79531   0.0729 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.440812     Adj. R2: 0.524396
##                  Within R2: 0.005564
```

Head Start participation seems to fade out with age. For example, for siblings who go to Head Start, on average, they score 0.32 standard deviations higher than siblings who do not go to Head Start for test scores between ages 5 and 6. However, for siblings who participate in Head Start, on average, they score 0.09 standard deviations higher than siblings who do not (this is not statistically significant). For siblings who go to Head Start, on average, they score 0.18 standard deviations higher than siblings who do not go to Head Start for test scores between 11 and 14 and this coefficient is not statistically significant.

Question 8

Code: Regressions.

Verbal: Interpret.

```
# All question 8 code here

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
```

```r
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
```

Controlling for fixed mother effects, the probability of being a high school graduate increases by 13 percentage points if the sibling participates in Head Start than the siblings who do not, on average. The probability of attending some college increases by 7 percentage points if the sibling participates in Head Start controlling for mother fixed effects than siblings who do not, on average. In addition, the probability of being idle or reporting fair/poor health decreases by 7 percentage points and 7 percentage points, respectively, if the sibling participated in Head Start.

Question 9

Code: Analysis of heterogeneity.

Verbal: Interpret.

```
# All question 9 code here

feols(hsgrad ~ head_start + black*head_start + hispanic*head_start | mom_id,
      data = nlsy_kids)
```

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

## The variables 'black' and 'hispanic' 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:black    0.100056   0.087448 1.144178  0.25275
## head_start:hispanic 0.066048   0.096913 0.681520  0.49566
## ... 2 variables were removed because of collinearity (black and hispanic)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.30996     Adj. R2: 0.173172
##                 Within R2: 0.009976
```

For children who are not Black or Hispanic, siblings who participate in Head Start are 6 percentage points more likely to graduate high school than siblings who do not participate in Head Start. For Black families, siblings who participate in Head Start are 16 percentage points more likely to graduate high school than siblings who do not participate in Head Start. For Hispanic families, siblings in Head Start are 13 percentage points more likely to graduate high school than siblings who do not participate in Head Start. However, for all of these estimates, it is not statistically significant. We cannot conclude that the estimates are different for different races since the t-value is not greater than 1.96 and the p-value is not less than 0.05.

Question 10

Verbal: Policy implications.

I think expanding federal funding for early-childhood education programs is the better position because when we control for mother fixed effects, there is an increase in the probability of graduating high school and attending some college. There is also a decrease in the probability of being idle or reporting fair/poor health. Although the test scores seem to fade out in later childhood, there are long term outcomes which are statistically significant. Although the coefficient on Head Start is a robust estimator, I would be a bit hesitant to use it to predict expansion of the program. Since the children in this sample could be different from the children who the Head Start could be expanded to, I do not think the effects of Head Start might be similar to these new eligible children. For example, the new eligible children might come from wealthier families so the Head Start program might not be as effective on test scores or long term outcomes.