Problem Set 3

Wanjun Gu

2/20/2020

```
knitr::opts_chunk$set(echo = TRUE)
suppressPackageStartupMessages({
  library(readstata13)
  library(ggplot2)
  library(gridExtra)
  library(margins)
  library(knitr)
  library(MASS)
  library(lme4)
  library(nlme)
  library(plm)
})
graphics.off()
rm(list = ls())

suppressWarnings({
  nd = read.dta13(file = "nlsy_deming.dta", nonint.factors = TRUE)
  source('data_description.R')
})
```

Q1

```
get_sd = function(x){
  x = na.omit(x)
  if(length(unique(x)) == 2){
    return(((mean(x)*(1-mean(x)))/length(x))^0.5)
  }else{
    return(sd(x))
  }
}
sum_nd = rbind.data.frame(apply(nd, 2, mean, na.rm = TRUE),
                          apply(nd, 2, get_sd))
colnames(sum_nd) = names(nd)
rownames(sum_nd) = c("Mean", "Std")
print(sum_nd)
```

```
##       head_start     sibdiff   mom_id  hispanic       black        male
## Mean 0.206565064 0.232121923 6227.424 0.2004689 0.320281360 0.509730363
## Std  0.006199042 0.006464645 3320.364 0.0061303 0.007144485 0.007654701
##        firstborn lninc_0to3     momed dadhome_0to3   ppvt_3      lnbw
## Mean 0.404454865 10.0699249 11.701338    0.6780303 21.88131 4.7184022
## Std  0.007515067  0.7627533  2.407231    0.4317648 13.12782 0.2290714
##      comp_score_5to6 comp_score_7to10 comp_score_11to14      repeat    learndis
## Mean        45.42266         45.19414          43.77577 0.315838222 0.041023166
## Std         22.37593         24.12119          24.80608 0.008167822 0.003081121
##           hsgrad    somecoll        idle    fphealth   HS2_FE90
## Mean 0.715181932 0.315244668 0.159083778 0.098808030 0.25985091
## Std  0.007993425 0.008228723 0.006478855 0.005285008 0.01011985
```

```
nd_head_start = nd[nd$head_start == 1,]
sum_nd_head_start = rbind.data.frame(apply(nd_head_start, 2, mean, na.rm = TRUE),
                                     apply(nd_head_start, 2, get_sd))
colnames(sum_nd_head_start) = names(nd)
rownames(sum_nd_head_start) = c("Mean", "Std")
print(sum_nd_head_start)
```

```
##      head_start    sibdiff   mom_id   hispanic     black      male  firstborn
## Mean          1 0.52667423 6646.919 0.18842225 0.5175936 0.5346198 0.42338252
## Std           0 0.01682144 3111.197 0.01317477 0.0168350 0.0168050 0.01664648
##      lninc_0to3     momed dadhome_0to3   ppvt_3      lnbw comp_score_5to6
## Mean  9.7835740 11.502838    0.5042258 18.12687 4.7108991        40.81176
## Std   0.6565652  2.273424    0.4575741 11.42641 0.2159896        20.62690
##      comp_score_7to10 comp_score_11to14    repeat    learndis     hsgrad
## Mean         38.18787          36.19352 0.4072581 0.042045455 0.71840659
## Std          22.42834          22.47849 0.0180128 0.006765356 0.01666981
##        somecoll       idle   fphealth HS2_FE90
## Mean 0.26923077 0.19230769 0.09615385        1
## Std  0.01643943 0.01460682 0.01092609        0
```

```
diff_df = sum_nd_head_start - sum_nd
diff_df[2,] = (sum_nd[2,]^2 + sum_nd_head_start[2,]^2)^0.5
print(diff_df)
```

```
##       head_start    sibdiff    mom_id    hispanic      black       male
## Mean 0.793434936 0.29455231  419.4955 -0.01204669 0.19731228 0.02488939
## Std  0.006199042 0.01802089 4550.2051  0.01453118 0.01828827 0.01846625
##       firstborn lninc_0to3      momed dadhome_0to3   ppvt_3        lnbw
## Mean 0.01892765 -0.2863509 -0.1985007   -0.1738045 -3.75444 -0.00750308
## Std  0.01826421  1.0064147  3.3110755    0.6291223 17.40409  0.31484156
##      comp_score_5to6 comp_score_7to10 comp_score_11to14     repeat    learndis
## Mean       -4.610894         -7.00627         -7.582255 0.09141984 0.001022289
## Std         30.432731         32.93725         33.475724 0.01977813 0.007433932
##           hsgrad    somecoll       idle     fphealth   HS2_FE90
## Mean 0.003224661 -0.04601390 0.03322391 -0.002654184 0.74014909
## Std  0.018487225  0.01838387 0.01597920  0.012137167 0.01011985
```

As we can see from the comparison result:
- Children who participate in the head start program tend to attend different high school from their siblings do.
- There tend to be fewer Hispanics, but the difference is not significant.
- There tend to be more attendee with African American ancestry.
- More attendees are males.
- More attendees are first born.
- They tend to come from families with lower income.
- They tend to have mothers with lower education level.
- They tend to not live with their fathers during early ages.
- Etc

In general, they tend to come from more handicap families with lower income, less parental care and low parent education and social status

Q2

```
fm2 = comp_score_5to6 ~ head_start
lm2 = rlm(formula = fm2, data = nd)
kable(summary(lm2)$coefficient)
```

	Value	Std. Error	t value
(Intercept)	46.659643	0.5533979	84.314817
head_start	-6.401757	1.2054803	-5.310544

If we assume the effect of test scores during the age of five to six is exogenous, then we will be confidently concluding that joining the head start program has a negative impact on test scores ($p << 0.001$). However, this conclusion does not make much logical sense because as much as the head start program can be unimpactful, it should do no harm to participants grades. This means that there are some other omitted variables that further dictate participants' grades. Thus, head start participation is not exogenous.

Q3

```
fm3 = comp_score_5to6 ~ head_start + (1|mom_id)
lm3 = lmer(fm3, data = nd, REML=TRUE)
kable(summary(lm3)$coefficient)
```

	Estimate	Std. Error	t value
(Intercept)	46.17539	0.5761476	80.145059
head_start	-2.55297	1.1400671	-2.239316

Considering the random effect of different mother makes the coefficient on head start participation less significant. To be more specific, considering mom education level decreases the coefficient of the head start program participation. This result means that to some degree, mom education level has effect on the test scores and head start program participation. Hence, we are even less confident about announcing the negative causal effect of head start on test scores.

Q4

```
fm4.1 = comp_score_5to6 ~ head_start + factor(mom_id)
fm4.2 = comp_score_5to6 ~ head_start + factor(mom_id) +
  hispanic + black + firstborn + lninc_0to3 +
  momed + dadhome_0to3

lm4.1 = lm(fm4.1, data = nd)
lm4.2 = lm(fm4.2, data = nd)
kable(summary(lm4.1)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.666668	16.703389	2.135295	0.0329811
head_start	7.632852	1.884853	4.049574	0.0000553

```
kable(summary(lm4.2)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.49688	24.445993	0.6748294	0.5000132
head_start	6.30668	2.331502	2.7049856	0.0070015

In order to take account of fixed effects, we can only include pre-head start variables. In addition, the variables should logically have an impact on the test score. In other words, the variables should be confounding variables instead of mediating variables. In this case, variables taking account of race, parental care and education and sibling status are the ones that we should including. Changes that happens later in the children's lives should be excluded.

As we use a simple fixed effect model to consider the fixed effect of different mothers, the coefficient on head start dramatically increases, from significantly negative to significantly positive. This result indicates that after controlling for fixed effect of different mothers, participating the head start program is proved to have a positive influence on children's grades.

Furthermore, as we control pre-Head Start variables, the coefficient on head start drops to a smaller positive value. This indicates that race, parental care and education and sibling status can also explain some variations within different mother groups.

The fixed effect estimator is drastically different from the random effect estimator. With that said, both models, compared to OLS, increases the coefficient of head start. This means that the effect of head-start on test score predicted by OLS is far from being causal. The discrepancies between RE and FE are likely caused by heteroskedasticity in mother groups. As a result, some family groups with higher odd ratios are severely under-weighted.

Q5

```
to_z = function(x){
  return((x - mean(x, na.rm = TRUE))/sd(x, na.rm = TRUE))
}
nd$zscore_5to6 = to_z(nd$comp_score_5to6)
nd$zscore_7to10 = to_z(nd$comp_score_7to10)
nd$zscore_11to14 = to_z(nd$comp_score_11to14)
fm5.1 = zscore_5to6 ~ head_start + factor(mom_id)
fm5.2 = zscore_7to10 ~ head_start + factor(mom_id)
fm5.3 = zscore_11to14 ~ head_start + factor(mom_id)
lm5.1 = lm(fm5.1, data = nd)
lm5.2 = lm(fm5.2, data = nd)
lm5.3 = lm(fm5.3, data = nd)
kable(summary(lm5.1)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.4360038	0.7464892	-0.5840725	0.5593042
head_start	0.3411189	0.0842358	4.0495735	0.0000553

```
kable(summary(lm5.2)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.5469937	0.5146026	-1.062944	0.2879585
head_start	0.1592453	0.0592830	2.686187	0.0072977

```
kable(summary(lm5.3)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.5351822	0.4049399	-1.321634	0.1864872
head_start	0.1530010	0.0597015	2.562765	0.0104789

As the result shows, the positive effect of head start program on test scores when controlling for fixed effect is fading as time progresses. The coefficient on head start is decreasing over time and the significance of the coefficients is decreasing as well. However, what should be noted is that as time progresses, there may be more and more factors that can potentially impact test scores which are not taken into the model. To further determine if the effect of head start program is fading or not, more data needs to be gathered on the subjects about their later development.

Q6

```
nd$repeat_grade = nd$`repeat`
fm6.1 = repeat_grade ~ head_start + factor(mom_id)
fm6.2 = learndis ~ head_start + factor(mom_id)
fm6.3 = hsgrad ~ head_start + factor(mom_id)
fm6.4 = somecoll ~ head_start + factor(mom_id)
fm6.5 = idle ~ head_start + factor(mom_id)
fm6.6 = fphealth ~ head_start + factor(mom_id)
fm6.7 = nd ~ head_start + factor(mom_id)
lm6.1 = lm(fm6.1, data = nd)
lm6.2 = lm(fm6.2, data = nd)
lm6.3 = lm(fm6.3, data = nd)
lm6.4 = lm(fm6.4, data = nd)
lm6.5 = lm(fm6.5, data = nd)
lm6.6 = lm(fm6.6, data = nd)
kable(summary(lm6.1)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0000000	0.2791923	3.581760	0.0003504
head_start	-0.0544033	0.0312228	-1.742424	0.0816062

```
kable(summary(lm6.2)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000000	0.1090953	0.000000	1.0000000
head_start	-0.0373492	0.0127786	-2.922781	0.0035013

```
kable(summary(lm6.3)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6666667	0.2369391	2.813663	0.0049507
head_start	0.1311789	0.0318957	4.112740	0.0000408

```
kable(summary(lm6.4)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000000	0.2372837	0.000000	1.0000000
head_start	0.0739958	0.0319421	2.316557	0.0206384

```
kable(summary(lm6.5)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3333333	0.2010510	1.657954	0.0974991
head_start	-0.0727877	0.0270646	-2.689402	0.0072234

```
kable(summary(lm6.6)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000000	0.1716709	0.000000	1.0000000
head_start	-0.0659418	0.0231096	-2.853436	0.0043738

More regressions suggest that joining the head start program is beneficial to children's growth even in terms of long-term outcomes. The benefits include less chances of repeating the same grade, less chances of suffering from learning disabilities, higher chances of graduating high school successfully, higher chances of attending and graduating from colleges and ultimately better self-reported health conditions.

Q7

```
fm6.1 = repeat_grade ~ head_start + hispanic + black + male + factor(mom_id)
fm6.2 = learndis ~ head_start + hispanic + black + male + factor(mom_id)
fm6.3 = hsgrad ~ head_start + hispanic + black + male + factor(mom_id)
fm6.4 = somecoll ~ head_start + hispanic + black + male + factor(mom_id)
fm6.5 = idle ~ head_start + hispanic + black + male + factor(mom_id)
fm6.6 = fphealth ~ head_start + hispanic + black + male + factor(mom_id)
fm6.7 = nd ~ head_start + hispanic + black + male + factor(mom_id)
lm6.1 = lm(fm6.1, data = nd)
lm6.2 = lm(fm6.2, data = nd)
lm6.3 = lm(fm6.3, data = nd)
lm6.4 = lm(fm6.4, data = nd)
lm6.5 = lm(fm6.5, data = nd)
lm6.6 = lm(fm6.6, data = nd)
kable(summary(lm6.1)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0000000	0.2776277	3.601946	0.0003245
head_start	-0.0557439	0.0310492	-1.795343	0.0727680

```
kable(summary(lm6.2)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000000	0.1086103	0.000000	1.0000000
head_start	-0.0384509	0.0127239	-3.021935	0.0025379

```
kable(summary(lm6.3)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6666667	0.2354573	2.831370	0.0046859
head_start	0.1324541	0.0316973	4.178712	0.0000307

```
kable(summary(lm6.4)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000000	0.2341822	0.000000	1.000000
head_start	0.0758184	0.0315257	2.404974	0.016273

```
kable(summary(lm6.5)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3333333	0.2007925	1.660088	0.0970691
head_start	-0.0722580	0.0270308	-2.673177	0.0075809

```
kable(summary(lm6.6)$coefficient[1:2,])
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.000000	0.1707181	0.000000	1.0000000
head_start	-0.065069	0.0229821	-2.831286	0.0046871

Regression results show that gender and race have little effect on the coefficients of head start in models predicting long-term outcomes. This suggests that the benefits that head start programs have may be universal, regardless of gender and race.

Q8

The result of this study advocates an expansion of early-childhood education programs. This is supported by the positive effects of the head-start program on variables indicating test-scores and long-term success (negative effects of head-start program on variables indicating long-term drawbacks). However, the data used in this study is likely insufficient to predict the effect of an expansion of early-childhood education programs. This is because the model explains very little of the variations of either test scores or long-term success markers. To better assess the effect of the programs, we need to collect more comprehensive data.