ECON 121 HW1

Wanjun Gu

1/31/2020

Import all the libraries and data

```
knitr::opts_chunk$set(echo = TRUE)
library(foreign)
library(readxl)
library(readstata13)
library(knitr)
library(ggplot2)
ps1_cps = read_xlsx(path = "ps1_cps.xlsx")
nlsy = read.dta13(file = "nlsy79.dta")
```

Question 1

In the Mincerian Wage Equation: $ln(w_i) = \beta_0 + \beta_1 ed_i + \beta_3 exper_i + \beta_4 exper^2 + \epsilon_i$, β_1 means the education wage return given the same years of experience. Or, with same years of experience, how much more (in log scale) does the individuals with one more year of education earn. The reason why experience has a square term $exper^2$ is because people expect the return of experience to be non-linear. That is, people would normally expect the more experience one tends to have, the more valuable the experience is.

Question 2

```
# Data Processing
cps = as.data.frame(read_excel(path = "ps1_cps.xlsx"))
educ_yr = cbind(c("Associate's degree, occupational/vocational program",
                  "Master's degree"                                    ,
                  "Grades 7 or 8"                                      ,
                  "High school diploma or equivalent"                  ,
                  "Bachelor's degree"                                  ,
                  "Some college but no degree"                         ,
                  "Doctorate degree"                                   ,
                  "12th grade, no diploma"                             ,
                  "Associate's degree, academic program"               ,
                  "Grade 10"                                           ,
                  "Grade 11"                                           ,
                  "None or preschool"                                  ,
                  "Professional school degree"                         ,
                  "Grade 9"                                            ,
                  "Grades 1, 2, 3, or 4"                               ,
                  "Grades 5 or 6"  ),
                c(16, 18, 7.5, 12, 16, 15,
                  22, 12, 14, 10, 11, 0,
                  14, 9, 2.5, 5.5))

cps$uhrsworkt = as.numeric(cps$uhrsworkt) # Assign NA if work hour varies
```

```
## Warning: NAs introduced by coercion
```

```
assign_educ = function(x){
  return(educ_yr[which(x == educ_yr[,1]),2])
}
cps$educ = as.numeric(sapply(cps$educ, assign_educ))
cps$age = as.numeric(cps$age)
cps$sex = as.factor(cps$sex)
cps$exper = cps$age - cps$educ - 5
cps$exper2 = cps$exper^2
cps$white = as.numeric(cps$race == "White")
cps$black = as.numeric(cps$race == "Black/Negro")
cps$other = as.numeric(cps$race != "White" & cps$race != "Black/Negro")
cps$race = as.factor(cps$race)
cps$hwage = log(cps$incwage / (cps$uhrsworkt * cps$wkswork1))
cps$hwage[cps$hwage == -Inf] = 0
cps = na.omit(cps) # Get rid of all the NA values
cps = cps[cps$uhrsworkt >= 35 & cps$wkswork1 >= 50,] # Get rid of part-time
kable(head(cps)) # Display data
```

	age	sex	race	uhrsworkt	educ	wkswork1	incwage	exper	exper2	white	black	other	hwage
2	55	Female	White	40	18	52	56000	32	1024	1	0	0	3.292984
6	59	Male	White	50	12	52	50000	42	1764	1	0	0	2.956512
7	29	Male	White	45	12	52	40000	12	144	1	0	0	2.838729
8	30	Female	White	40	16	52	30000	9	81	1	0	0	2.668830
9	40	Male	White	50	15	52	0	20	400	1	0	0	0.000000
10	46	Female	White	40	16	52	45000	25	625	1	0	0	3.074295

```
summary(cps) # Summarize data
```

```
##       age            sex                                     race      
##  Min.   :25.00   Female:22495   White                           :39825  
##  1st Qu.:35.00   Male  :28784   Black/Negro                     : 5923  
##  Median :43.00                  Asian only                      : 3717  
##  Mean   :43.41                  American Indian/Aleut/Eskimo  :  634  
##  3rd Qu.:52.00                  White-American Indian         :  342  
##  Max.   :64.00                  Hawaiian/Pacific Islander only:  268  
##                                 (Other)                       :  570  
##    uhrsworkt          educ          wkswork1        incwage       
##  Min.   : 35.0   Min.   : 0.00   Min.   :50.00   Min.   :      0  
##  1st Qu.: 40.0   1st Qu.:12.00   1st Qu.:52.00   1st Qu.:  30000  
##  Median : 40.0   Median :15.00   Median :52.00   Median :  48500  
##  Mean   : 43.5   Mean   :14.57   Mean   :51.95   Mean   :  64346  
##  3rd Qu.: 45.0   3rd Qu.:16.00   3rd Qu.:52.00   3rd Qu.:  75002  
##  Max.   :170.0   Max.   :22.00   Max.   :52.00   Max.   :1609999  
##                                                                    
##      exper           exper2           white            black       
##  Min.   :-2.00   Min.   :   0.0   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:15.00   1st Qu.: 225.0   1st Qu.:1.0000   1st Qu.:0.0000  
##  Median :23.00   Median : 529.0   Median :1.0000   Median :0.0000  
##  Mean   :23.84   Mean   : 691.6   Mean   :0.7766   Mean   :0.1155  
##  3rd Qu.:33.00   3rd Qu.:1089.0   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :58.00   Max.   :3364.0   Max.   :1.0000   Max.   :1.0000  
##                                                                    
##      other            hwage       
##  Min.   :0.0000   Min.   :-6.254  
##  1st Qu.:0.0000   1st Qu.: 2.668  
##  Median :0.0000   Median : 3.074  
##  Mean   :0.1079   Mean   : 2.996  
##  3rd Qu.:0.0000   3rd Qu.: 3.520  
##  Max.   :1.0000   Max.   : 6.428  
## 
```

Question 3

```
kable(summary(lm(data = cps, hwage ~ educ + exper + exper2))$coefficient)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2566103	0.0292396	42.976331	0
educ	0.1028490	0.0014890	69.072179	0
exper	0.0181678	0.0016023	11.338398	0
exper2	-0.0002787	0.0000319	-8.741979	0

Based on the regression results, the return of education is 0.1% increase of wage for one year of education.

Question 4

```
kable(summary(lm(data = cps, hwage ~ white + black + sex + educ + exper + exper2))$coefficient)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.1530248	0.0320641	35.959958	0.0000000
white	-0.0214552	0.0129493	-1.656865	0.0975529
black	-0.1340282	0.0168742	-7.942800	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
sexMale	0.1875892	0.0080700	23.245342	0.0000000
educ	0.1057297	0.0014885	71.032201	0.0000000
exper	0.0169517	0.0015935	10.637776	0.0000000
exper2	-0.0002535	0.0000317	-7.995401	0.0000000

The regression result shows that after controling for sex and race, the coefficient on education becomes more significant. Also, the sex and age variable themselves are significantly correlated with wage. This indicates that sex and age differences also explains variations in wage return of education and they also explains wage differences.

Question 5

```
l = summary(lm(data = cps, hwage ~ white + black + sex + educ + exper + exper2))$coefficient
white_coef = l[2,1]
white_se = l[2,2]
male_coef = l[4,1]
male_sd = l[4,2]
se = (white_se^2+male_sd^2)^0.5
t_score = (white_coef - male_coef)/se
print(t_score)
```

```
## [1] -13.70059
```

As the result shows, t statistics is way greater than 1.96, therefore the differnece is significant.

Question 6

```
cps_male = cps[cps$sex == "Male",]
cps_female = cps[cps$sex == "Female",]
ml = summary(lm(data = cps_male, hwage ~ white + black + educ + exper + exper2))$coefficient
fl = summary(lm(data = cps_female, hwage ~ white + black + educ + exper + exper2))$coefficient
print("Regression results for males")
```

```
## [1] "Regression results for males"
```

```
kable(ml)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3543112	0.0446224	30.350502	0.0000000
white	-0.0269620	0.0188256	-1.432198	0.1520979
black	-0.1629416	0.0255723	-6.371796	0.0000000
educ	0.1026864	0.0020453	50.206012	0.0000000
exper	0.0208734	0.0023426	8.910550	0.0000000
exper2	-0.0003353	0.0000461	-7.279945	0.0000000

```
print("Regression results for females")
```

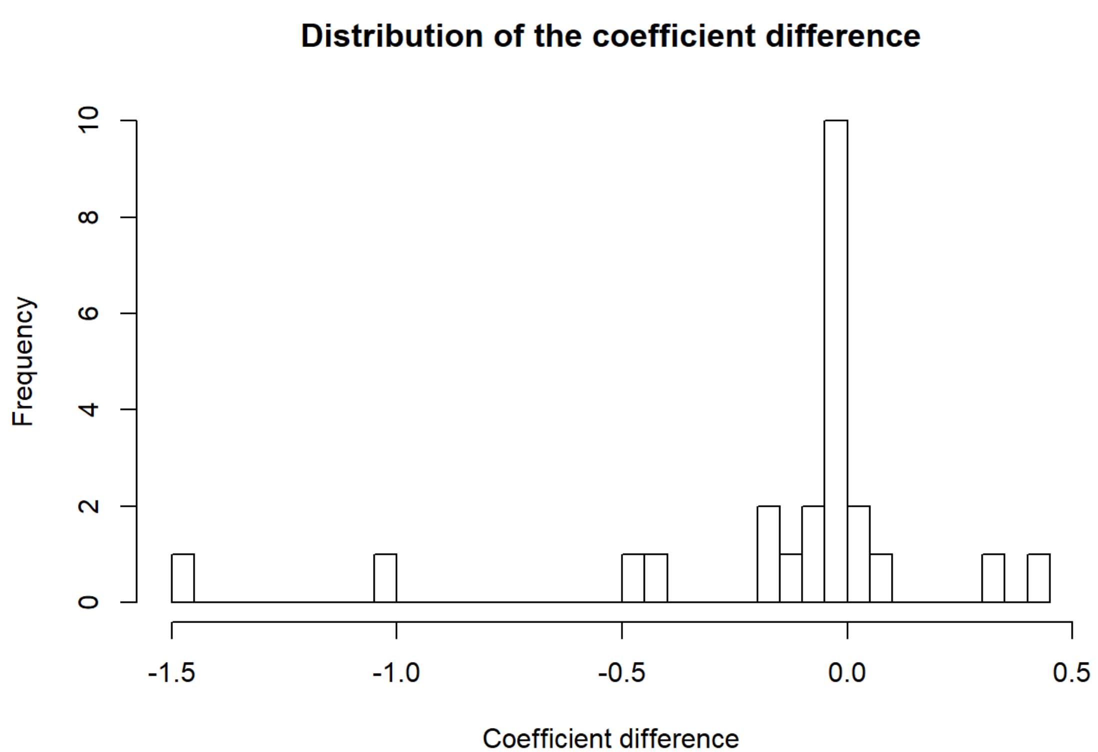
```
## [1] "Regression results for females"
```

```
kable(fl)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.1138004	0.0436557	25.5132882	0.0000000
white	-0.0167805	0.0171791	-0.9767937	0.3286818
black	-0.1068967	0.0214618	-4.9807839	0.0000006
educ	0.1106325	0.0021375	51.7574453	0.0000000
exper	0.0122206	0.0020913	5.8434101	0.0000000
exper2	-0.0001500	0.0000422	-3.5521645	0.0003829

Although there is no way to synthesize the standard error of the two coefficients of male and female since the samples are innately different, we can use bootstrap to determine the standard deviation of the difference between the two samples.

```
# Construct a bootstrap engine
dl_list = vector()
for(i in 1:1000){
  cps_male_sample = cps_male[sample(1:dim(cps_male)[1], 1000, replace = TRUE),]
  cps_female_sample = cps_male[sample(1:dim(cps_female)[1], 1000, replace = TRUE),]
  ml = summary(lm(data = cps_male_sample, hwage ~ white + black + educ + exper + exper2))$coefficient
  fl = summary(lm(data = cps_female_sample, hwage ~ white + black + educ + exper + exper2))$coefficient
  dl = ml - fl
  dl_list = c(dl_list, dl)
}
hist(dl, breaks = 50, main = "Distribution of the coefficient difference",
     xlab = "Coefficient difference")
```



```
print(paste0("Mean Difference: ", mean(dl)))
```

```
## [1] "Mean Difference: -0.133405598431262"
```

```
print(paste0("Difference SE: ", sd(dl)))
```

```
## [1] "Difference SE: 0.395711981006444"
```

As shown in the bootstap result, the difference is not sigificant. Therefore, the education returns in male and female seperately are not statistically significant.

Question 7

```
cps$male = as.numeric(as.factor(cps$sex)) - 1
interact_l = summary(lm(data = cps, hwage ~ white + black + educ + exper + exper2 + I(male*educ)))$coefficient
kable(interact_l)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2723547	0.0315085	40.381267	0.0000000
white	-0.0202005	0.0129536	-1.559453	0.1188954
black	-0.1343747	0.0168809	-7.960163	0.0000000
educ	0.0980912	0.0014953	65.598967	0.0000000
exper	0.0169183	0.0015942	10.612222	0.0000000
exper2	-0.0002541	0.0000317	-8.011397	0.0000000
I(male * educ)	0.0121233	0.0005409	22.414143	0.0000000

The result got from Q7 is the same the one from Q6. This suggests that the results got from bootstap and interaction variable are the same.

Question 8

```
# Display data
kable(head(nlsy[,1:8]))
```

age79	foreign	urban14	mag14	news14	lib14	educ_mom	educ_dad
20	1	1	0	1	1	8	8
20	1	1	1	1	1	5	8
17	0	1	1	1	1	10	12
16	0	1	1	1	0	11	12
19	0	1	1	1	1	12	12
18	0	1	0	1	1	12	12

```
# Unweighted summary
black = na.omit(nlsy$black)
hisp = na.omit(nlsy$hisp)

black_mean = sum(black, na.rm = TRUE)/length(black)
hisp_mean = sum(hisp, na.rm = TRUE)/length(hisp)

black_sd = ((black_mean * (1-black_mean))/length(black))^0.5
hisp_sd = ((hisp_mean * (1-hisp_mean))/length(hisp))^0.5


#Weighted summary
black_mean_weight = sum(black*nlsy$perweight, na.rm = TRUE)/sum(nlsy$perweight)
hisp_mean_weight = sum(hisp*nlsy$perweight, na.rm = TRUE)/sum(nlsy$perweight)

black_sd_weight = ((black_mean_weight * (1-black_mean_weight))/length(black))^0.5
hisp_sd_weight = ((hisp_mean_weight * (1-hisp_mean_weight))/length(hisp))^0.5

Q8 = data.frame(
  black_mean = sum(black, na.rm = TRUE)/length(black),
  hisp_mean = sum(hisp, na.rm = TRUE)/length(hisp),

  black_sd = ((black_mean * (1-black_mean))/length(black))^0.5,
  hisp_sd = ((hisp_mean * (1-hisp_mean))/length(hisp))^0.5,


  #Weighted summary
  black_mean_weight = sum(black*nlsy$perweight, na.rm = TRUE)/sum(nlsy$perweight),
  hisp_mean_weight = sum(hisp*nlsy$perweight, na.rm = TRUE)/sum(nlsy$perweight),

  black_sd_weight = ((black_mean_weight * (1-black_mean_weight))/length(black))^0.5,
  hisp_sd_weight = ((hisp_mean_weight * (1-hisp_mean_weight))/length(hisp))^0.5)

kable(Q8)
```

black_mean	hisp_mean	black_sd	hisp_sd	black_mean_weight	hisp_mean_weight	black_sd_weight	hisp_sd_weight
0.2501971	0.1578118	0.0038455	0.0032368	0.1418509	0.0654364	0.0030977	0.0021956

The weighted summary better discribes the population ratio becuase in the unweighted sampings, due to the fact that black and hispanics are Ethnic minorities, they are upsampled and therefore over-represented Considering weight can eliminate the differences.

Question 9

```
hour_wage = log(nlsy$laborinc07/nlsy$hours07)
hour_wage[hour_wage == -Inf] = 0

experience = nlsy$age79 + 28 - nlsy$educ - 5

nlsy_plus = cbind(nlsy, experience, hour_wage)
nlsy_plus = subset(nlsy_plus, hours07 > 1750)
nlsy_plus = na.omit(nlsy_plus)

rm(experience, hour_wage)

# For black
print("Unweighted summary for the black")
```

```
## [1] "Unweighted summary for the black"
```

```
kable(summary(lm(data = nlsy_plus, hour_wage ~ educ +
                 I(experience^2) +
                 experience + black + male))$coefficient)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.7016775	0.6577952	4.107171	0.0000409
educ	0.1255954	0.0084000	14.951855	0.0000000
I(experience^2)	0.0026465	0.0008244	3.210304	0.0013369
experience	-0.1342427	0.0448467	-2.993366	0.0027769
black	-0.2188915	0.0317476	-6.894745	0.0000000
male	0.2645347	0.0278221	9.508090	0.0000000

```
print("Weighted summary for the black")
```

```
## [1] "Weighted summary for the black"
```

```
kable(summary(lm(data = nlsy_plus, hour_wage ~ educ +
                 I(experience^2) +
                 experience + black + male, weights = perweight))$coefficient)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.3082053	0.6918313	3.336370	0.0008569
educ	0.1304702	0.0082055	15.900332	0.0000000
I(experience^2)	0.0023259	0.0008866	2.623437	0.0087394
experience	-0.1136084	0.0478590	-2.373817	0.0176545
black	-0.2308859	0.0447666	-5.157550	0.0000003
male	0.3060472	0.0279673	10.943041	0.0000000

```
# For Hispanic
print("Unweighted summary for the hispanic")
```

```
## [1] "Unweighted summary for the hispanic"
```

```
kable(summary(lm(data = nlsy_plus, hour_wage ~ educ +
                 I(experience^2) +
                 experience + hisp + male))$coefficient)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.8177221	0.6630552	4.2496040	0.0000219
educ	0.1273055	0.0084832	15.0067059	0.0000000
I(experience^2)	0.0029466	0.0008290	3.5543455	0.0003835
experience	-0.1501018	0.0451024	-3.3280229	0.0008829
hisp	0.0238112	0.0375895	0.6334539	0.5264752

	Estimate	Std. Error	t value	Pr(>|t|)
male	0.2765659	0.0279506	9.8948241	0.0000000

```
print("Weighted summary for the hispanic")
```

```
## [1] "Weighted summary for the hispanic"
```

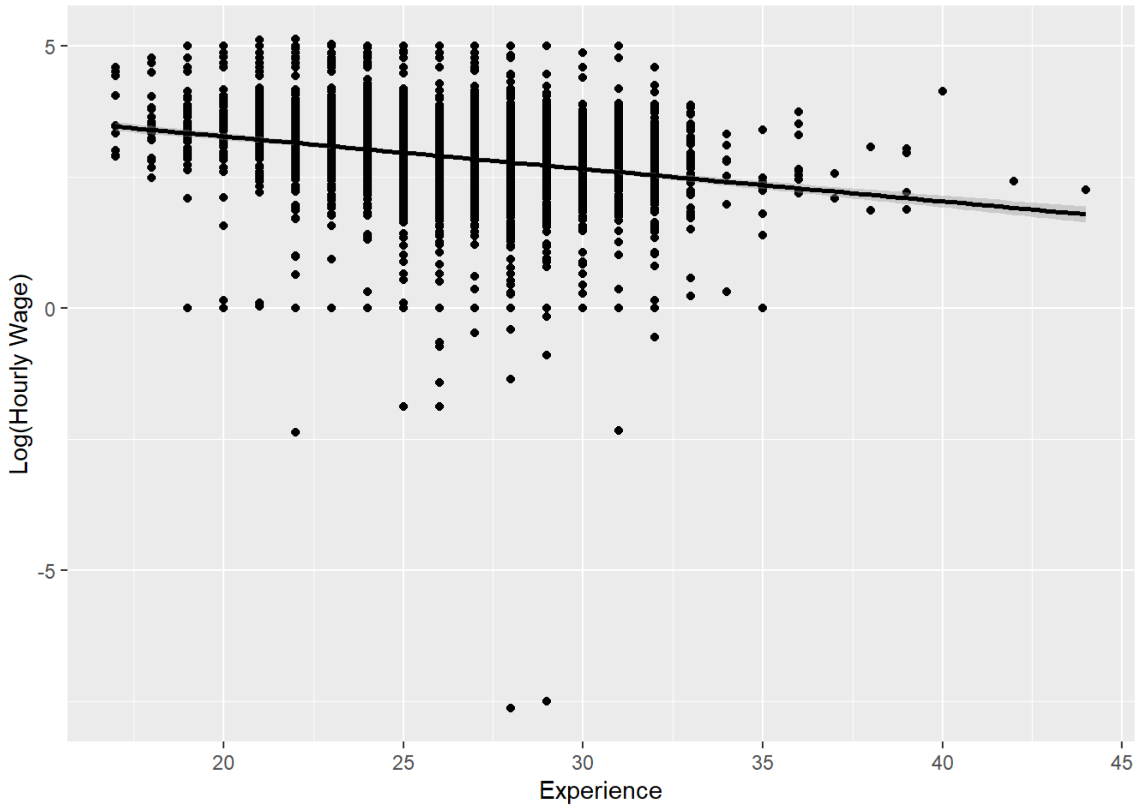
```
kable(summary(lm(data = nlsy_plus, hour_wage ~ educ +
                   I(experience^2) +
                   experience + hisp + male, weights = perweight))$coefficient)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.3432575	0.6946602	3.3732429	0.0007503
educ	0.1317033	0.0082426	15.9783250	0.0000000
I(experience^2)	0.0024453	0.0008896	2.7487750	0.0060100
experience	-0.1198860	0.0480238	-2.4963905	0.0125883
hisp	0.0041861	0.0625679	0.0669041	0.9466615
male	0.3130156	0.0280345	11.1653726	0.0000000

The usage of sampling weight can slightly change the coefficient and increase the standard error of the statistics. However, Weighted regression is preferred because it takes consideration of the Over sampling of minority populations such as hispanic or black. Therefore, I prefer the weighted regression.

Question 10

```
ggplot(data = nlsy_plus, aes(x = experience, y = hour_wage)) +
  geom_point() +
  geom_smooth(method = "lm", color = "black") +
  xlab("Experience") +
  ylab("Log(Hourly Wage)")
```



The coefficient of education obtained from NLSY is similar to the coefficient of education obtained from CPS. Both are similar to around 0.13. However, the coefficients of experience and experience^2 are different acrosstwo datasets. Particularly, the coefficients from NLSY is negative, which is realistically unlikely. This suggests that either: - The two samples are innately different in terms of populaiton component - The way how part-time workers and non-working force are excluded influences the result. In NLSY, workers who work for less than 1750 hours are dropped however in CPS, workers who work less than 35 hours per week or less than 50 weeks per year are dropped.

Question 11

I think β_1 does not represent the causal effect of education. This is because although the correlation between education and hour wage is found to be significant. There is no evidence indicating causality. In fact, there may be many other variables correlated with both education and wage that are ommited in this regression For instance, family income is traditionally believed to be correlated with education because richer households tend to afford more education. Wealth status is also related to wage given that difference in parents' income may suggest difference in access to resources. Therefore, we cannot prove that the regression is not subjective to ommited variable bias and we cannot determin causal correlation.

Question 12

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.7527504	0.6590189	4.177043	0.0000302
educ	0.1245644	0.0084399	14.758978	0.0000000
I(experience^2)	0.0026805	0.0008248	3.249907	0.0011644
experience	-0.1361035	0.0448682	-3.033405	0.0024344
black	-0.2298632	0.0329400	-6.978241	0.0000000
hisp	-0.0483787	0.0387629	-1.248066	0.2120835
male	0.2630655	0.0278449	9.447518	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.7232917	0.6551837	5.682821	0.0000000
educ	0.0662963	0.0098559	6.726588	0.0000000
I(experience^2)	0.0030033	0.0008127	3.695461	0.0002226
experience	-0.1672974	0.0442697	-3.779045	0.0001598
black	-0.0424899	0.0370793	-1.145918	0.2519008
hisp	0.0709117	0.0403266	1.758437	0.0787533
male	0.2288108	0.0275976	8.290959	0.0000000
urban14	0.0660547	0.0331545	1.992329	0.0464061
afqt81	0.0073370	0.0006767	10.843085	0.0000000

I think living at urban places at the age of 14 and AFQT should be included in the regression. As the statistics shows, the coefficients of both Urban14 and AFQT81 are significant. As an explaination, the AFQT score is a good indicator of one's cognitive ability and should be correlated with earning, since we expect smart people to earn more. Also living in urban environment at a early age can be significant to access to resource and education quality. Therefore, I expect these two factors to be added to the regression.

Question 13

In a natural experiment or survey setting, it's hard for OLS to indicate causal relationship because non of the conditons are randomly assigned. Therefore, it is hard to include/control all the variables that are potentially correlated with the error term. The coefficients are thus only good for passvie prediction but not causation.