Econ 140B Project

Aram Dovlatyan

**Component 1**

1) Preparing the data for analysis meant removing any strange values. Since income is a continuous variable, it is likely a value used for a numeric code would be extremely high or low. It can be inferred that the income values of 9999999 are a numeric code for missing data, whereas a 0 for income means the person has no source of income / possibly unemployed. So to remove the values of 9999999, I subset the data.frame excluding all the observations of 9999999 in the incwage column so as to avoid skewed data and outliers.

2) For education, to prepare the data for analysis, since we are only given categorical values and no numeric value of years of education, we will have to break up education into dummy variables,

   using "**LHS**" as my reference category, "LHS" pertaining to less than high school. This will fall into the intercept. All the education values of grades 1 to 12 will be grouped into this variable.

   "**HS**" which takes a value of 1 meaning at least high school diploma, this also includes the data from people with AA degrees, and some college but no degree.

   "**Bach**" which takes a value of 1 for bachelors degree and 0 if else.

   "**Grad**" which takes a value of 1 for any higher education above bachelors which includes the educational values of master's degree, professional degree and doctorate degree.

   For gender I am going to create one dummy variable, the variable will be called **Male** and if it takes a value of 1 it represents a male and 0 if else.
   The average wage for females will fall into the reference category aka the intercept with LHS.

   For race I will split up the categories into white and non-white. Non-whites will fall into the reference category, so that will include all other races. Therefore I will create 1 dummy variable, named "**white**". For values of 1 it is the respective race, and 0 if else. The motivation behind only 1 race is that because when I

want to check interaction terms in the regression model I will have to create many new variables, which I believe could lead to overfitting. In addition there are many observations for white so the distribution would be more normal and the sample is larger.

For marital status, I will create a dummy variable called "**Married**" taking a value of 1 if the person is married and 0 if else. I will group Married, spouse present and spouse absent into married. I will group separated, divorced, widowed, never married/single and widowed or divorced into the value of 0 which in turn falls under the intercept. Now the intercept would be the average income for a non-white non-married female with less than high school education.

## Code for component 1

```
setwd("C:/Users/aramd/Desktop/Fall 2018 classes stuff/Econ 140B"
    df <- read.csv("cps_earnings.csv")
    View(df)
    summary(df)
    str(df)

##### Component 1 - preparing the data for analysis
library(tidyr)
library(dplyr)
library(AER)

#1) Removing/subsetting income reported as 99999999
df1 <- subset(df,
              subset = incwage != 9999999)
hist(df1$incwage)

#2) Examine data on edu,gender,race and marital status
names(df1$married)
summary(df1$educ)
str(df1)
View(df1)

#2) Regrouping categorical variables into dummy variables (Component 1
part 2)
```

```
df1$hs      <- ifelse(df1$educ == "high school diploma or equivalent" | df1$educ
== "some college but no degree" |
                    df1$educ == "associate's degree, occupational/vocational
program"  |
                    df1$educ == "associate's degree, academic program", 1, 0)
df1$married <- ifelse(df1$marst == "married, spouse present" | df1$marst ==
"married, spouse absent ", 1, 0)
df1$male    <- ifelse(df1$sex  == "male", 1, 0)
df1$white   <- ifelse(df1$race == "white", 1, 0)
df1$bach    <- ifelse(df1$educ == "bachelor's degree", 1, 0)
df1$grad    <- ifelse(df1$educ == "master's degree" | df1$educ == "professional
school degree" |
                    df1$educ == "doctorate degree", 1, 0)
```

**Component 2**

1) I estimated a model that explains income as a function only of education by creating a multivariable linear regression consisting of dummy variables for each education category and an intercept. I decided to run a regression to be able to create predictions from the model.

2) When including gender and race in the model, I decided to create interaction terms between the explanatory variables for all possible combinations because I wanted to see if there is a difference in income between males and females holding education level constant, as well as a difference in income between white males & females and non white males & females holding education level constant.

3) When including marital status, I included 2 more interaction terms to see if there are differences between married males and non married males. In addition to see if there are differences between married white males and non married non white males.

**Code for component 2**

```
##### Component 2 - Constructing and estimating a model

# Part 1 - Regress income as a function only of education
# Part 1 -Intercept/Reference category is Less than high school observations
reg1 <- lm(incwage ~ hs + bach + grad, data = df1)
```

```
summary(reg1)

Residuals:
   Min      1Q  Median      3Q     Max
-70013  -23477   -7214   11523 1219986

Coefficients:
             Estimate     Std. Error   t value      Pr(>|t|)
(Intercept)   7213.8       333.8        21.61       <2e-16 ***
hs           16263.1       385.1        42.23        <2e-16 ***
bach         41038.0       473.1        86.74       <2e-16 ***
grad         62798.7       551.8       113.80        <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53530 on 143622 degrees of freedom
Multiple R-squared:  0.1062,         Adjusted R-squared:  0.1062
F-statistic:  5688 on 3 and 143622 DF,  p-value: < 2.2e-16
```

**#Part 2 - Include gender and race and interaction terms**
**#Also include interactions to see if there is difference in income between**
**# males and females holding education level constant, as well as**
**# difference in income between white males & females and non white males**
**# & females holding education level constant**

```
reg2 <- lm(incwage ~ hs + bach + grad + male + white + male*hs + male*bach +
male*grad + white*hs + white*bach + white*grad + male*white*hs + male*white*bach +
male*white*grad, data = df1)
summary(reg2)

Residuals:
   Min      1Q  Median      3Q     Max
-90428  -17076   -6474   12410 1237409
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 4028.72 | 931.91 | 4.323 | 1.54e-05 | *** |
| hs | 12781.55 | 1073.20 | 11.910 | < 2e-16 | *** |
| bach | 30780.51 | 1329.54 | 23.151 | < 2e-16 | *** |
| grad | 50460.00 | 1547.68 | 32.604 | < 2e-16 | *** |
| male | 2444.99 | 1345.46 | 1.817 | 0.06919 | . |
| white | 339.36 | 1075.68 | 0.315 | 0.75240 | |
| hs:male | 5058.30 | 1559.83 | 3.243 | 0.00118 | ** |
| bach:male | 16361.62 | 1975.19 | 8.284 | < 2e-16 | *** |
| grad:male | 27500.86 | 2269.17 | 12.119 | < 2e-16 | *** |
| hs:white | -73.89 | 1236.83 | -0.060 | 0.95236 | |
| bach:white | 314.02 | 1521.79 | 0.206 | 0.83652 | |
| grad:white | -2237.92 | 1772.44 | -1.263 | 0.20673 | |
| male:white | 4350.06 | 1541.05 | 2.823 | 0.00476 | ** |
| hs:male:white | 3467.25 | 1784.15 | 1.943 | 0.05197 | . |
| bach:male:white | 7152.50 | 2240.26 | 3.193 | 0.00141 | ** |
| grad:male:white | 3542.38 | 2584.20 | 1.371 | 0.17044 | |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52580 on 143610 degrees of freedom
Multiple R-squared: 0.1374, Adjusted R-squared: 0.1373
F-statistic: 1525 on 15 and 143610 DF, p-value: < 2.2e-16

**#Part 3 - Include marital status in model**
**# Interaction term for married & males and married white males to see**
**# if there is a difference bettwen married males and non married males.**
**# Also to see if there is a difference between married white males**
**# and non married white males. This is also applied to females too**
**# we can check all those differences for females too.**

```
reg3 <- lm(incwage ~ hs + bach + grad + male + white + male*hs + male*bach +
male*grad + white*hs + white*bach + white*grad + male*white*hs + male*white*bach +
male*white*grad + married + married*male + married*male*white, data = df1)
summary(reg3)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -94490 | -18576 | -5389 | 11424 | 1239535 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3317.7 | 938.7 | 3.534 | 0.000409 *** |
| hs | 12188.2 | 1074.1 | 11.347 | < 2e-16 *** |
| bach | 29492.1 | 1350.3 | 21.841 | < 2e-16 *** |
| grad | 48857.0 | 1576.7 | 30.988 | < 2e-16 *** |
| male | -348.1 | 1360.2 | -0.256 | 0.798032 |
| white | 163.2 | 1087.5 | 0.150 | 0.880728 |
| married | 3876.7 | 841.0 | 4.609 | 4.04e-06 *** |
| hs:male | 2669.6 | 1564.3 | 1.707 | 0.087912 . |
| bach:male | 12044.1 | 2007.5 | 5.999 | 1.98e-09 *** |
| grad:male | 21375.0 | 2322.8 | 9.202 | < 2e-16 *** |
| hs:white | -213.1 | 1240.2 | -0.172 0 | .863583 |
| bach:white | 479.6 | 1546.3 | 0.310 | 0.756442 |
| grad:white | -1874.0 | 1804.1 | -1.039 | 0.298905 |
| male:white | 2255.8 | 1564.2 | 1.442 | 0.149266 |
| male:married | 12098.6 | 1233.4 | 9.809 | < 2e-16 *** |
| white:married | -756.4 | 954.3 | -0.793 | 0.427990 |
| hs:male:white | 2275.7 | 1792.1 | 1.270 | 0.204145 |
| bach:male:white | 5846.7 | 2278.3 | 2.566 | 0.010282 * |
| grad:male:white | 2637.8 | 2642.6 | 0.998 | 0.318193 |
| male:white:married | 2887.1 | 1395.6 | 2.069 | 0.038582 * |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52250 on 143606 degrees of freedom
Multiple R-squared: 0.1484, Adjusted R-squared: 0.1483
F-statistic: 1317 on 19 and 143606 DF, p-value: < 2.2e-16

**Component 3**

1) One might be concerned that education is endogenous because the coefficient on education might be biased in its effect on income because the coefficient on education contains a composite effect of ability and other omitted variables. One might be concerned that education is correlated with the error term or in other words our omitted variables. The implication of education being endogenous is that our estimate of the coefficient on education is biased.

2) I believe the only potential instruments given our dataset are region, state and health. I believe region and state have a similar degree of validity as they are just geographical locations. It can be argued that where you grow up is like a random draw of a card and thus it would be uncorrelated with error term, and it could be argued that the type of region people grew up in across educational levels is not the same. By type of region I mean a city's demographic and infrastructure. There could be a possible positive correlation between living near a university and pursuing higher levels of education. It could very well influence whether someone pursues higher education, to see the intuition behind this I believe it's helpful to think of an extreme case, comparing two people where one grew up in the rural midwest and another in an urban city like Los Angeles. There are less colleges in rural areas so it is less likely for someone to attend colleges in those areas versus someone growing up in an urban city where there are many colleges.

Since the two assumptions for an instrument are met in region/state, it is a potential instrument, however the categories for the region and state data are too generalized and too big, they cover too big of a region of the U.S. so I don't believe they will spit out a significant estimated coefficient in the 1st stage of a two stage least squares estimate. If the data was more specific or in detail to city it might be possible to be a good instrument. State would most likely be a better instrument than region since it covers less area and is more specific. There would be more correlation between higher education and state as midwestern states have significantly less colleges than coastal states.

Another problem with using these as an instrument is that we have multiple binary endogenous variables, we took education as categorical then split it into multiple dummy variables. So to run a 2SLS on education, the endogenous variable, would mean running separate 2SLS estimates for every education dummy variable we made (HS, Bach and Grad). There would be multiple 1st stage regressions, for each endogenous education variable. In addition all the potential instruments are categorical variables, meaning they would have to be split into dummy variables or converted to

numeric variables, so we would break up health into multiple binary instrumental variables. There would also have to be at least as many instruments as endogenous variables. In addition all the explanatory variables and instruments would be binary, to my knowledge we have not seen an example of this in this class and I believe attempting to construct and estimate a 2SLS model given this dataset is outside the scope of this class, nonetheless I gave it a shot.

Another possible instrument in our dataset could be health because I think health and education levels are correlated. I believe that people with better health are more likely to have pursued higher education levels than people with poorer health. It seems logical to me that people with poor health are not pursuants of higher education as they can't afford the education investment or are not in good health to even go to school. In addition I think a person in better health, although health is a vague term to me, would be more likely to pursue higher education, so there will be more people with higher education in better health than lower education levels. Again another problem with this instrument is that health would have to be split up into multiple dummy variables, to regress all the educational dummy variables on the health dummy variables.

I attempted to construct and estimate a model with health as an instrument. I did not run separate 1st stage regressions for each binary endogenous education variable, I instead combined the higher education categories of bachelor degrees and anything greater than masters degrees, as I believe this is where a correlation between health and education would be strong. So I regressed this combined education variable on multiple instruments, each instrument being a binary variable split up from the categorical health variable. I then tried to run a a tsls model with only one instrument, the one that had the highest coefficient and correlation with education. This is where things get hazy, I had trouble trying to understand my code and what to do next. Using the package AER, I tried to run a tslsmodel but had trouble reading the instruction on how to properly use the package to run a 2SLS model. In addition I didn't know how to do it with default R.

## Attempted code for component 3

```
##### Component 3
#Part 3 testing if health level is valid as instrument for getting bachelors degree
# and higher, if health level is correlated with higher education
df1$higheredu         <- ifelse(df1$bach == 1 | df1$grad == 1, 1, 0)
df1$excelhealth       <- ifelse(df1$health == "excellent", 1, 0)
df1$verygoodhealth    <- ifelse(df1$health == "very good", 1, 0)
```

```
df1$averagehealth     <- ifelse(df1$health == "good" | df1$health == "fair", 1, 0)
df1$poorhealth        <- ifelse(df1$health == "poor", 1, 0)
```

**#1st stage of 2SLS estimate, see coefficients of instruments to see which instrument is best**
**#instruments are split up binary variables from health category**

```
reg4 <- lm(higheredu ~ (excelhealth) + (verygoodhealth) + (poorhealth) + male + white + married + married*male + married*male*white , data = df1)
summary(reg4)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.4872 -0.2952 -0.2027  0.5593  0.9728

Coefficients:
                    Estimate      Std. Error      t value         Pr(>|t|)
(Intercept)         0.135963      0.004319      31.478          < 2e-16 ***
excelhealth         0.145474    0.002893      50.280          < 2e-16 ***
verygoodhealth      0.103913  0.002790     37.250          < 2e-16 ***
poorhealth         -0.069575   0.006200      -11.222          < 2e-16 ***
male               -0.037184      0.006232       -5.966          2.43e-09 ***
white              0.002150       0.004866         0.442             0.6586
married            0.205748      0.006763      30.422          < 2e-16 ***
male:married        0.009045   0.009881    0.915           0.3600
white:married      -0.048638   0.007679   -6.334          2.40e-10 ***
male:white         -0.004187   0.007328      -0.571          0.5677
male:white:married  0.024003   0.011187   2.146          0.0319 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.435 on 143615 degrees of freedom
Multiple R-squared:  0.06273,       Adjusted R-squared:  0.06266
F-statistic: 961.2 on 10 and 143615 DF,  p-value: < 2.2e-16
```

**#2SLS estimate with AER package and highest T value instrument which was excelhealth**

#I believe excelhealth captured most variation in higheredu so I only include that

```
tslsmodel <- ivreg(incwage ~ higheredu +  male +  white + white*male |
                (excelhealth) + male + white + white*male, data = df1)
summary(tslsmodel)

Residuals:
   Min      1Q  Median      3Q     Max
-111462  -12834    6657   22166 1198028

Coefficients:
             Estimate      Std. Error      t value          Pr(>|t|)
(Intercept)  -5568.3     1115.7        -4.991            6.02e-07 ***
higheredu    98627.8    3724.7       26.480           < 2e-16 ***
male         11661.5        670.3         17.398          < 2e-16 ***
white        -1088.6        520.3         -2.092          0.0364 *
male:white    7829.4       759.2       10.313           < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60170 on 143621 degrees of freedom
Multiple R-Squared: -0.1293,        Adjusted R-squared: -0.1294
Wald test: 897.4 on 4 and 143621 DF,  p-value: < 2.2e-16
```

**#Compare the new value of estimated coefficient of higheredu accounting for instrument excelhealth since it was best instrument**
**#Now put new value of higheredu back into original regression to see new coefficients and see if bias was reduced or endogeneity was resolved**

```
reg5 <- lm(incwage ~ higheredu + male + white + white*male, data = df1)
summary(reg5)

Residuals:
   Min      1Q  Median      3Q     Max
 -66983  -21954  -11301    13017 1241624
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 11300.576 | 409.320 | 27.608 | <2e-16 | *** |
| higheredu | 37069.969 | 313.885 | 118.101 | <2e-16 | *** |
| male | 10653.297 | 592.840 | 17.970 | <2e-16 | *** |
| white | 4.815 | 458.337 | 0.011 | 0.992 | |
| male:white | 7953.916 | 674.210 | 11.797 | <2e-16 | *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53440 on 143621 degrees of freedom
Multiple R-squared: 0.1092, Adjusted R-squared: 0.1092
F-statistic: 4402 on 4 and 143621 DF, p-value: < 2.2e-16