MINERVA®

Homework I

SS154 - Spring 2019

Taha Bouhoun

Data structure and methodology:

The Current Population Survey is a cross-sectional dataset generated from interviewing randomly selected households throughout the US territory. The CPS data used in this assignment offers a snapshot of the US population's characteristics as of March 2017. The dataset contains **185914** subjects and **697** covariates from which we can find binary variables (sex, employment, etc.), categorical variables (race, education, etc.), and continuous variables (earnings, income)

The research question I am interested in is the differences in income between genders and the significance of the gap in total personal income for male and female having the same educational level i.e., same obtaining the same degree. The attribute of being female or male is an intrinsic characteristics, hence, detecting a causal effect is not possible in this case.

The graphs that represent the data:

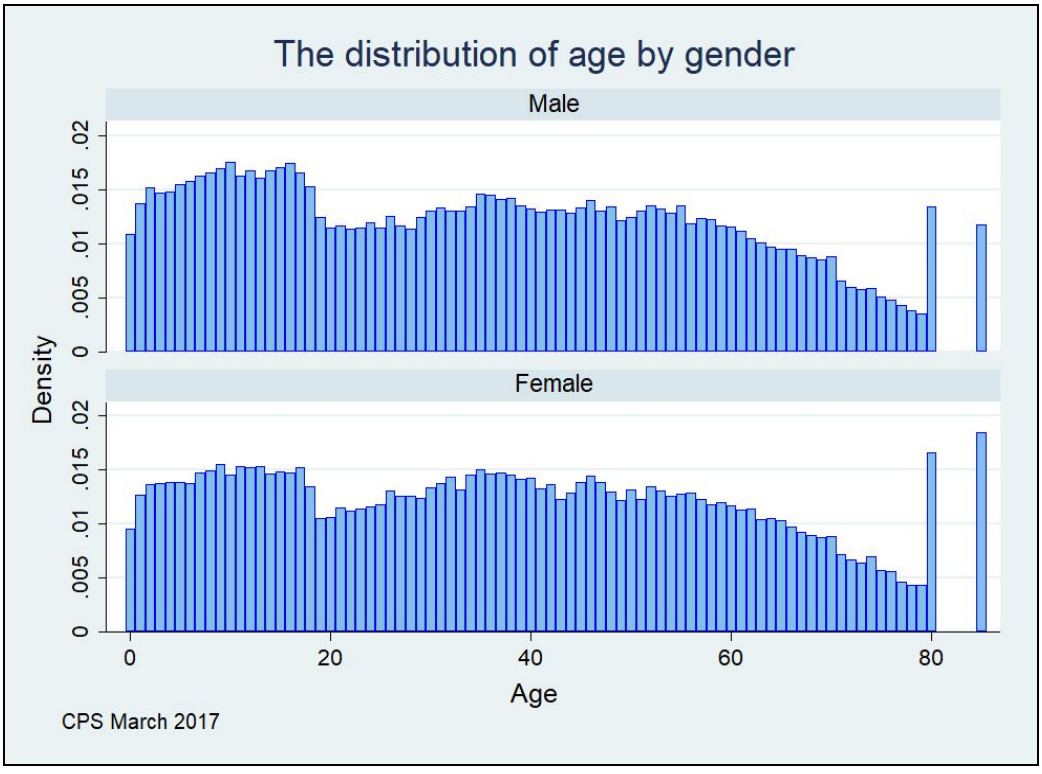

Fig 1. Histograms illustrating the distribution of age by gender from the sample obtained by the Current Population Survey - March 2017

According to Fig 1, the sample has a nearly identical distribution of males and females throughout the age groups which gives us the ability to block for age when comparing between the two genders.

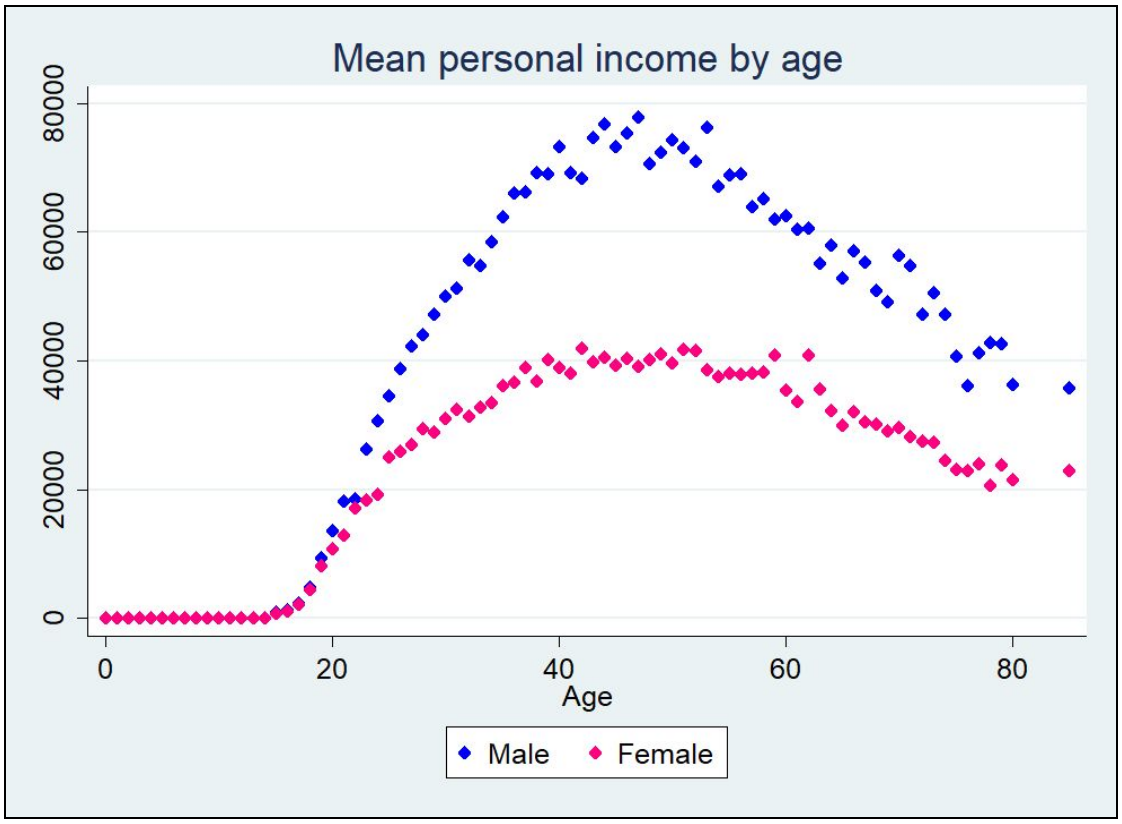

Fig 2. Scatter plot of the mean personal income for Male and Female for all age groups (US dollars)

We start by monitoring the bigger picture of the difference between the mean income of Males and females. Fig 2 offers an overview of the gap that starts getting wider after practically graduating from college (around 25 years old). To further investigate the underlying gap, we can test whether the educational level is behind the disparity in income for males and female. Fig 3 and 4 illustrates the personal income for bachelor's holders.

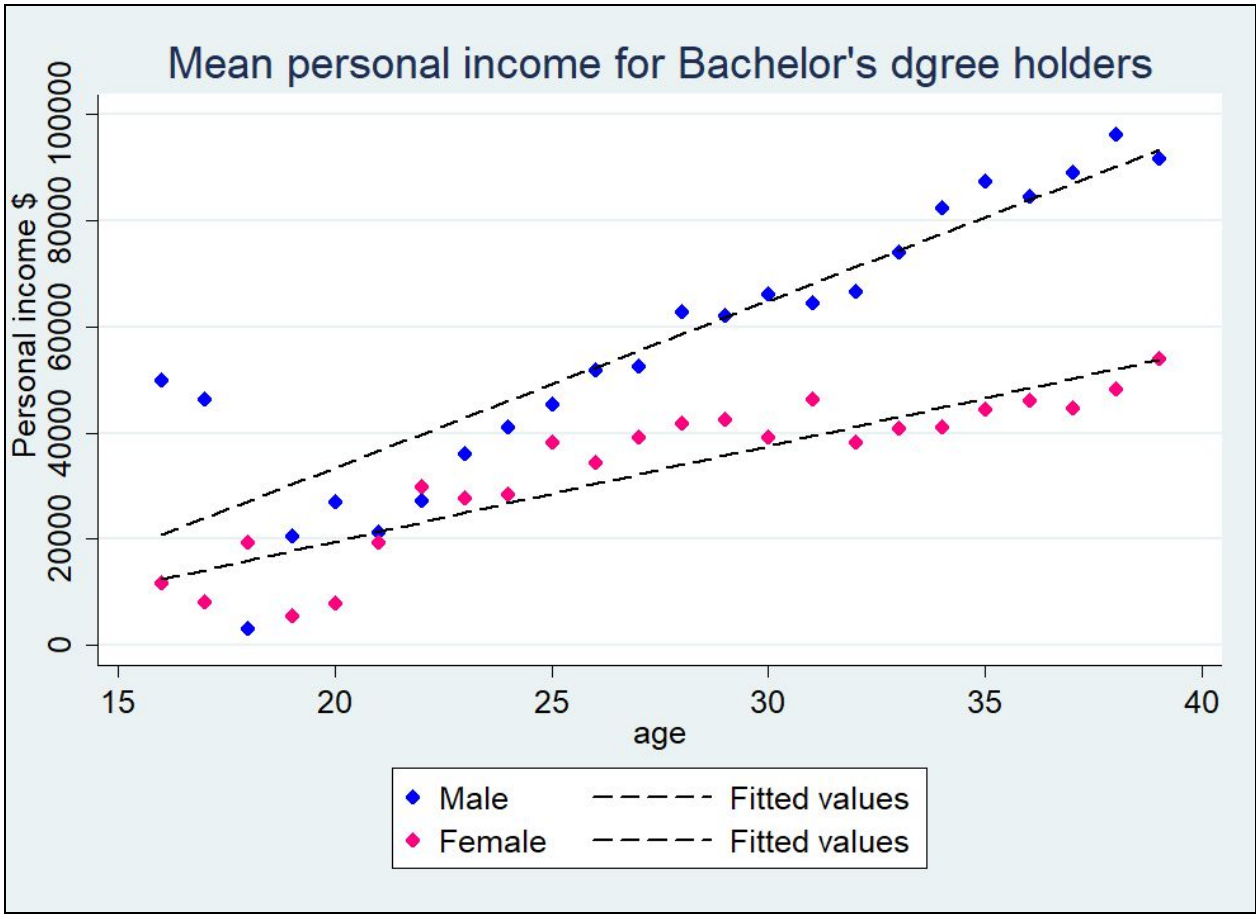

Fig 3. Personal income for Bachelor's degree holders by gender along with the linear regression line. We notice that the slope for males is steeper compared to females.

```
. reg ptotval_m age

      Source |       SS       df       MS              Number of obs =    4568
-------------+------------------------------           F(  1,  4566) =  306.91
       Model |  1.5723e+12     1  1.5723e+12           Prob > F      =  0.0000
    Residual |  2.3392e+13  4566  5.1231e+09           R-squared     =  0.0630
-------------+------------------------------           Adj R-squared =  0.0628
       Total |  2.4964e+13  4567  5.4662e+09           Root MSE      =   71576

------------------------------------------------------------------------------
   ptotval_m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    3683.47   210.2584     17.52   0.000     3271.262    4095.678
       _cons |   -46398.95   6623.926    -7.00   0.000    -59385.05   -33412.85
------------------------------------------------------------------------------
```

Table 1. The summary of the regression line between personal income for male Bachelor's degree holders who are less than 40 years old.

```
. reg ptotval_f age

      Source |       SS       df       MS              Number of obs =    5812
-------------+------------------------------           F(  1,  5810) =   91.71
       Model |  2.2228e+11     1  2.2228e+11           Prob > F      =  0.0000
    Residual |  1.4082e+13  5810  2.4238e+09           R-squared     =  0.0155
-------------+------------------------------           Adj R-squared =  0.0154
       Total |  1.4304e+13  5811  2.4616e+09           Root MSE      =   49232

------------------------------------------------------------------------------
   ptotval_f |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1201.751   125.4901     9.58   0.000     955.7433    1447.758
       _cons |   3408.046   3915.316     0.87   0.384    -4267.432    11083.52
------------------------------------------------------------------------------
```

Table 2. The summary of the regression line between personal income for female Bachelor's degree holders who are less than 40 years old.

Table 1[1]*:*

The linear model is of the form $ptotval - m = \beta_0 + \beta_1 \cdot age + \upsilon$ where "ptotval_m" stands for the personal income of all male with a Bachelor's degree, and the variable "age" is the represents people that are between 15 and 40 years old.

The estimate of the variable "age" is **3693.47 $** which is interpreted to be the amount that male Bachelor's holder earn depending on their age. The standard error is **210.25** and we can notice that 0 doesn't fall within the **95%** confidence interval as the p-value is extremely significant (less than **0.0005**)

Table 2:

The linear model is of the form $ptotval - f = \beta_0 + \beta_1 \cdot age + \upsilon$ where "ptotval_f" stands for the personal income of all female with a Bachelor's degree, and the variable "age" is the represents people that are between 15 and 40 years old.

The estimate of the variable "age" is **1207.75 $** which is interpreted to be the amount that female Bachelor's holder earn depending on their age. The standard error is **125.49** and we can notice that 0 doesn't fall within the **95%** confidence interval as the p-value is extremely significant (less than **0.0005**

[1] #descriptive stats: interpreting the linear regression parameters as well as the significance of the findings and relating the conclusion to the research question proposed in the beginning of the paper.

Hypothesis testing[2]:

1. *Stating the null hypothesis:* the mean difference of personal income of both male and female Bachelor's degree holders is zero $\mu = 0$
2. *Significance level:* we will test the significance level for $\alpha = 0.01$ hence 99% C.I
3. *Test statistic:* we can use a t-test to determine whether we can reject or fail to reject the null hypothesis.
4. *Computation:* we use the **`ttest`** function in Stata for a one sample test.
5. *Interpretation:* reading through the findings and testing its significance level.
6. *Checking the assumptions:* plotting the residuals for both groups to check the hetrostocacity assumption.

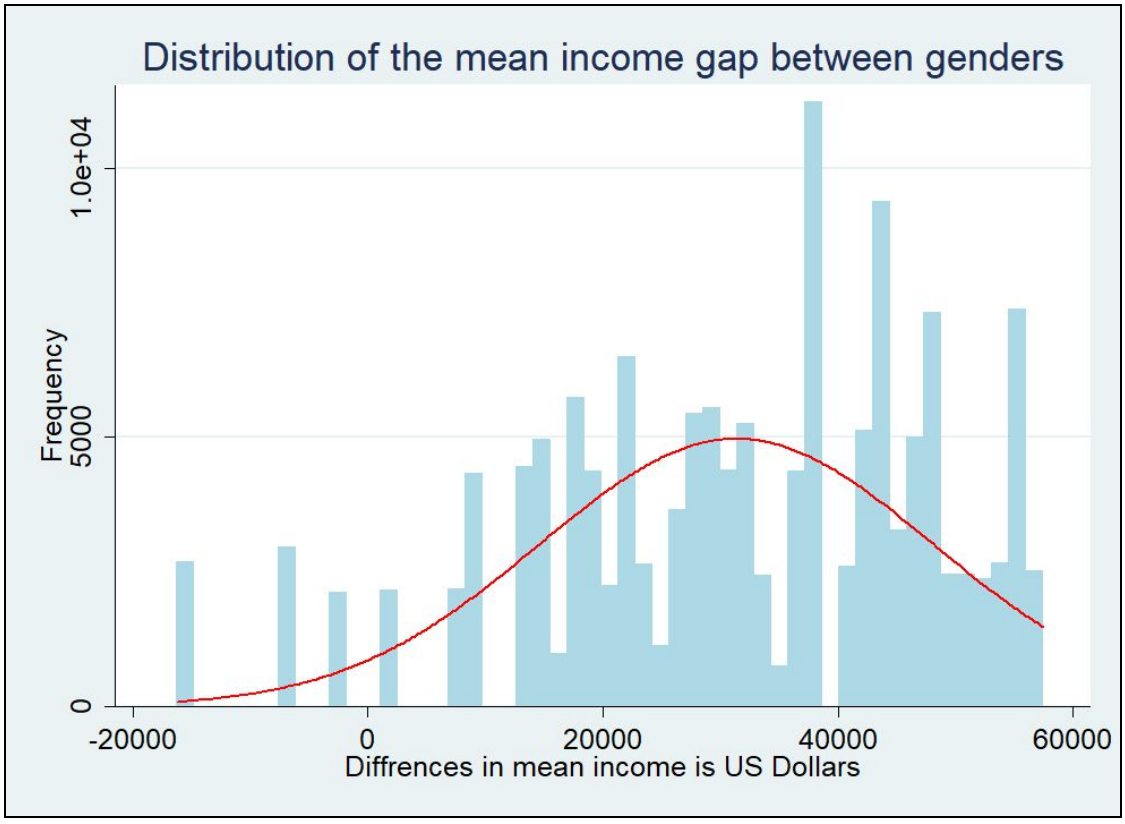

Fig 5. The distribution of the difference of mean income of male and female Bachelor's degree holders.

[2] #hypothesis development: the analysis focuses on the gap of personal income between genders. The assumption is that if having the same educational level, the mean income for both groups have to be equal to zero.

```
. ttest diff==0, level(99)

One-sample t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[99% Conf.	Interval]
diff	185914	34122.68	26.52004	11434.84	34054.36	34190.99

```
    mean = mean(diff)                                     t =  1.3e+03
Ho: mean = 0                             degrees of freedom =   185913

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Table 3. The t test of the hypothesis that the difference in mean income of male and female Bachelor's degree holder is zero.

Table 3:

The results confirm the significance of the gap between male and female personal income although having the same educational level. The distribution of the difference in mean income has a $\mu = 34122.68$ and a standard error $\delta = 26.52$

We can notice that 0 doesn't fall within the 99% confidence interval and the p-value of the test is smaller than 0.0001

The residuals and the assumptions of the least square method:

Linear regression is based on many assumptions. Fig 5 tests for the assumption stated by Guass-Markov of hetrostocasity and whether the variance is steady across the independent variable (in this case age).

We notice that for both groups, the residuals have are approximately centered at 0. However, the variance gets wider as we move along the x-axis which then violates the assumption.

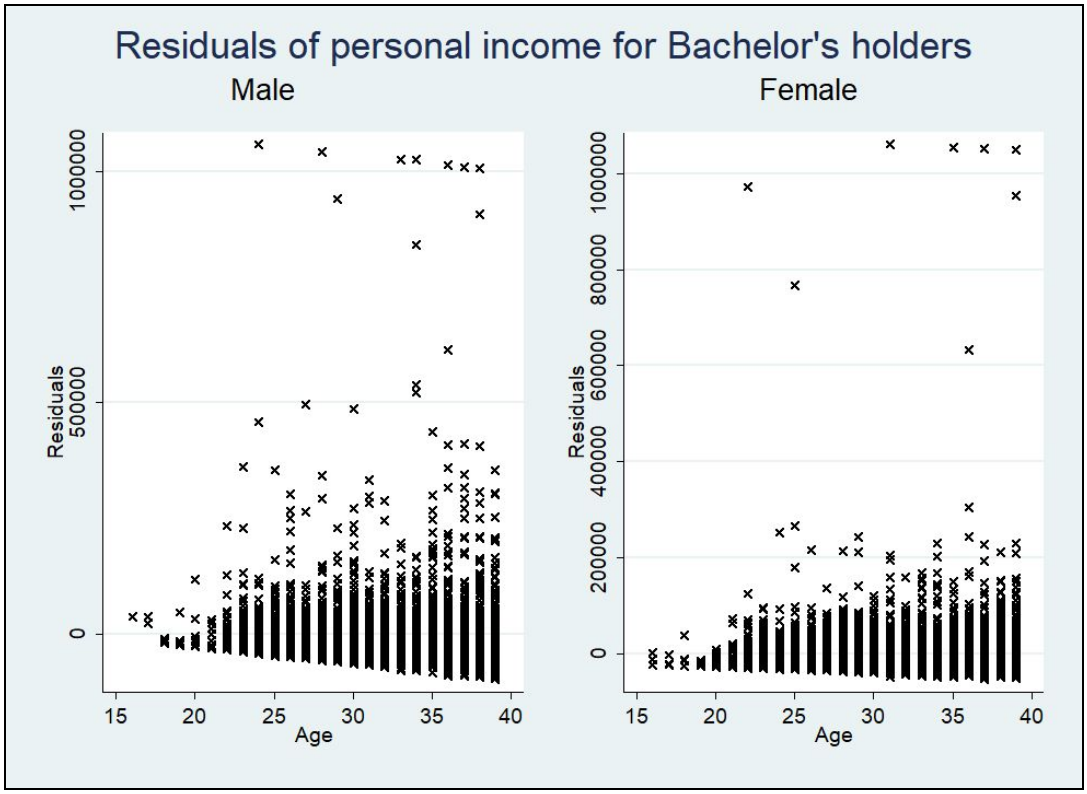

Fig 6. The residual graphs of both mean personal income of male and female Bachelor's degree holders.

Causal effect:

The data structure makes it hard to disentangle any causal inferences as identifying treatment and control groups can lead to confoundedness. The correlation in mentioned in the previous analysis can be attributed to four different reasons:

1. An underlying causal effect: which require further investigation to identify preferably by design eg. using an instrumental variable.
2. Reverse causal effect: where the dependent variable causes the dependent variable
3. Coincidence: when the correlation is strong within one sample and doesn't hold for others
4. Hidden cause: the case where a confounded variable causes the change in both the independent and the dependent variable.

Appendix:

Word count:

1. Stata code link: https://gist.github.com/Tahahaha7/deb754b807feea5d8b35637f92b50ba2
2. CPS data source:

http://www.nber.org/data/cps_progs.html
http://www.nber.org/data/cps.html