Homework 1: Stata Practice

Bo-Yu Chen
r11323006

April 10, 2023

1 Read Data

Question 1.1

The dataset which I use is Manpower Utilization Survey, 2021 by Directorate-General of Budget, Accounting and Statistics, Executive Yuan in Taiwan; it contains a lot of variables, such as area, city, sex, age, number of household members, education level , marital status, , income, remote work or not.

Question 1.2

I use the following code to read my dataset, which is Stata format:

```
cd /Users/bychen/Documents/LaborEcon/term/raw/110manpower
use "mu110.dta", clear
```

2 Examine Data

Question 2.1

I choose ‘b1_a‘ which is monthly income of the individual:

```
sum b1_a, detail
```

and the mean is 41605.43, the median is 35200

Question 2.2

I tab ‘b1_c1‘, which shows that a worker is remote working or not or non-reply, and I use the following code:

```
tab b1_c1
```

the result shows that 84 people work from home and 3864 people didn't work form home. However, there are 23554 people didn't answer this question.

Question 2.3

I use the following code:

```
inspect id a2 a5_3 a21_2 a22 b1_a b1_c1
```

where ‘id‘ is sample id, ‘a2‘ is gender, ‘a5_3‘ is education level, ‘a21_2‘ is industry, ‘a22‘ is occupation, ‘b1_a‘ is monthly income, ‘b1_c1‘ is WFH or not. The result shows that there is no missing value.

Question 2.4

I use the following code:

```
duplicates report
```

The result shows that there is no exactly same observation.

Question 2.5

I use the following code:

```
duplicates report id
```

The result shows that there are a lot of duplicates, but I have no idea why id is not unique.

3 Create Sample for Analysis: Part 1 and Part 2

Question 3.1

I use the following code:

```
gen year = 2021
```

Because I will conduct DID in the future, I need to add year variable.

Question 3.2

I use the following code:

```
egen median_inc = pctile(b1_a), p(50)
```

Where I create the variable called ‘median_inc‘ which is the median of monthly income of all observations.

Question 3.3

In this dataset, all the variable are labeled, so I label the variable crated in Question 3.2. I use the following code:

```
label var median_inc `"The median of monthly income of all
    observations."'
```

Question 3.4

I use the following code:

```
    recode b1_c1 (0 = .)(1 = 1) (2 = 0), generate(wfh)
```

In ‘b1_c1‘ 0 = no-reply, 1 = remote worker, 2 = not remote worker. Here, I recode ‘b1_c1‘ as ‘wfh‘, in which . = no-reply, 1 = remote worker, 0 = not remote worker.

Question 3.5

I use the following code:

```
    clear
    cd /Users/bychen/Documents/LaborEcon/term/work
    use "mu110_edit.dta", replace
    append using "mu109_edit.dta"
```

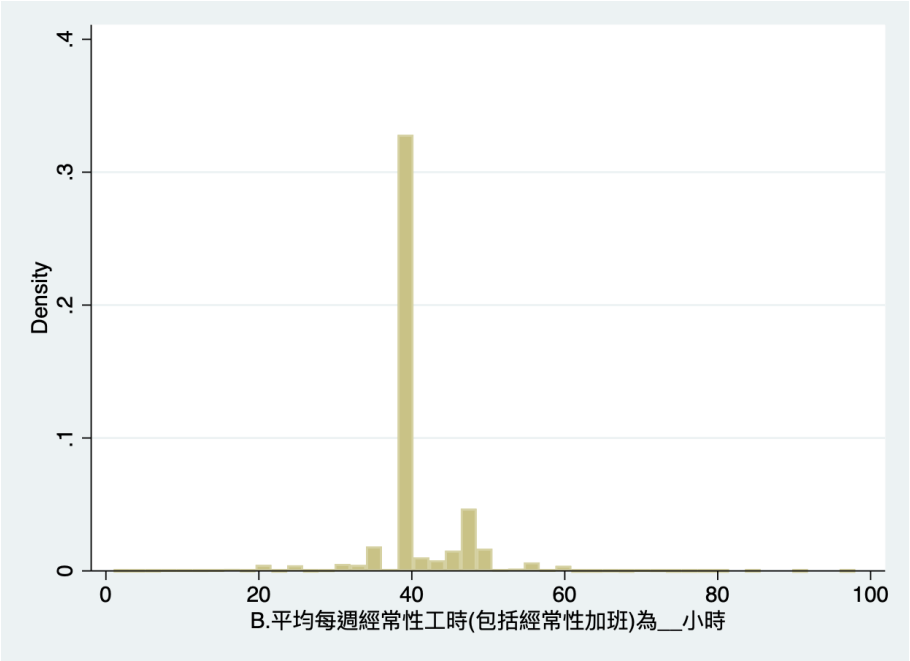
In this dataset, I append ”mu109_edit.dta” on ”mu110_edit.dta”, where ”_edit” means that I add ‘median_inc‘ and ‘year‘ variable in ”mu109.dta” and ”mu110.dta”.

4 Visualize Data

Question 4.1

I use the following code:

```
histogram b2_b
```

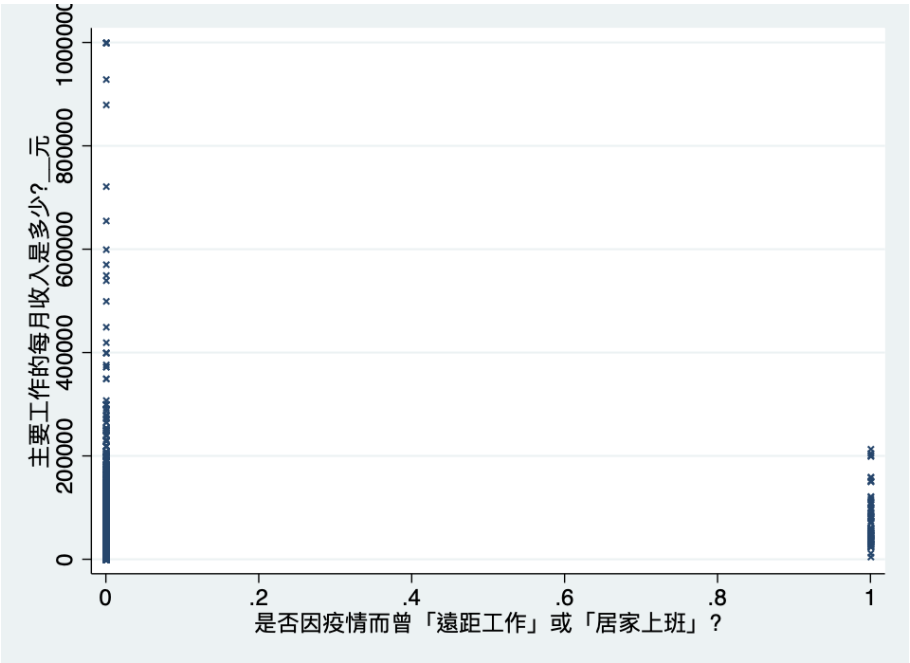


We can notice that working 40 hrs is majority.

Question 4.2

I use the following code:

```
twoway scatter b1_a wfh if wfh !=.
```



We can notice that remote workers seem to earn less if we did not control other variables.

5 Preliminary Analysis

Question 5.1

```
1 regress b1_a wfh if wfh != .
```

The coefficient of ‘wtf‘ is 30306.35, which means that if remote workers earn more than commuter.

Question 5.2

One omitted variable is ”education level”. Consider the following models and the OVB formula.

$$Y_i = \alpha D_i + \beta X_i + e_i$$

$$Y_i = \alpha D_i + u_i$$

$$\hat{\alpha} \to \alpha + \frac{Cov(u_i, D_i)}{Var(D_i)} = \alpha + \beta \frac{Cov(X_i, D_i)}{Var(D_i)}$$

The first equation is true model, The second equation is wrong model we use, where we omit X_i.
And Y_i is monthly salary, X_i is education level, D_i is WFH variable. We know that higher education level earn more and has higher possibility to be allowed remote work, so β and $\frac{Cov(X_i, D_i)}{Var(D_i)}$ are positive. Thus, according to the OVB formula, $\hat{\alpha}$ is overestimated.

Question 5.3

I use the following code:

```
regress b1_a wfh i.a5_3 i.a21_2 i.a22 i.year if wfh!=., r
```

Where 'b1_a' is monthly salary, 'wfh' is wfh or not, 'a5_3' is education level, 'a21_2' is industry, 'a22' is occupation, 'year' is either 2020 or 2021.
I put education level in the regression because higher education level workers are more likely to do a white-collar job and they can easily remote work. And different industries and occupations will affect the probability of being a remote worker, for example, a software engineer can easily work from home compared to farmers, so I put these two variables in the regression.
Those covariates are not bad because one's occupation and industry he works in and education level affect whether he works from home or not and his salary.
The coefficient of 'WFH' is 14346.29, and I think it doesn't make sense. A possible reason leading to this result is that the respondent of the survey doesn't reply to their true job level, for example, a high-level HR manager just reply she is an HR.