

Homework 1: Stata Practice

陳力行
r13627024

November 1, 2024

1 Read Data

Question 1.1

The dataset which I use is the information about the Survey of Family Income and Expenditure, it contains a lot of variables, such as year, household income, composition, type, expenditures, etc.

Question 1.2

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

```
1 cd "/Users/coco/Desktop/台大計量/stata"  
2  
3 use "inc108.dta"  
4 append using "inc109.dta"  
5 append using "inc110.dta"  
6 append using "inc111.dta"  
7 append using "inc112.dta"
```

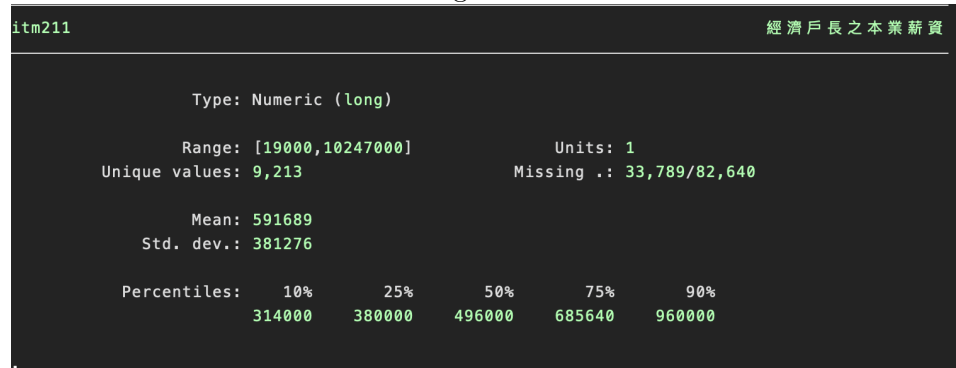
2 Examine Data

Question 2.1

I use the following code to find missing value from data in item211. I found 33,789 missing data entries.

```
1 codebook itm211
```

Figure 1:

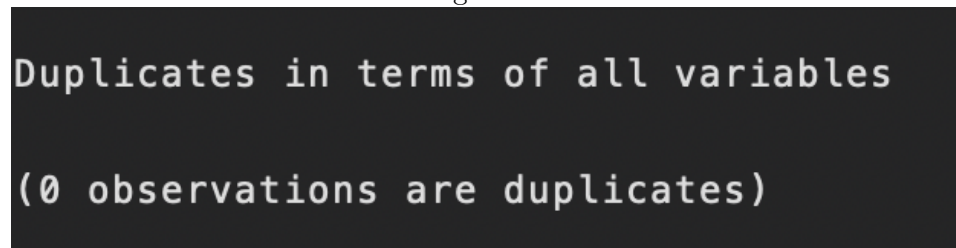


Question 2.2

I use the following code to check for duplicates and remove them.

```
1 duplicates drop
```

Figure 2:

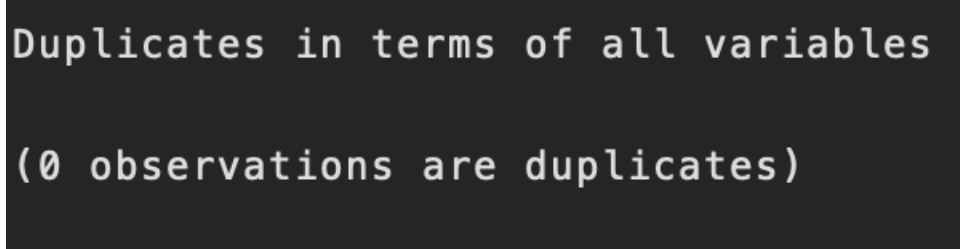


Question 2.3

I used the following command to verify whether all duplicate values in the data have been removed.

```
1 duplicates list
```

Figure 3:



Duplicates in terms of all variables
(0 observations are duplicates)

3 Create Sample for Analysis

Question 3.1

The original dataset contains many columns, but I am only focusing on the relationship between the household head's salary income and age. I selected occupation and gender as control variables and excluded unnecessary columns to facilitate further analysis.

Question 3.2

I used the following code to clean the data, retain the desired columns, and generate dummy variables.

```
1 keep id a5 a6 a7 a11 itm211 year
2 rename a5 job
3 tabulate job
4 tabulate job, gen(job_dummy)
5
6 rename a6 age
7 tabulate age
8
9 rename a7 gender
10 tabulate gender, gen (gender_dummy)
11
12 rename a11 edu
13 tabulate edu
14 recode edu (1/6 = 1) (7/8 = 2) (9/10 = 3)
15 label define edu_labels 1 "高中(含高中及五專前三年)以下" 2 "大學" 3 "研
    究所以上"
16 label values edu edu_labels
```

```

17 tabulate edu, gen (edu_dummy)
18
19 drop if missing(itm211)
20 generate ln_income = log(itm211)
21 describe ln_income

```

4 Visualize Data

Question 4.1

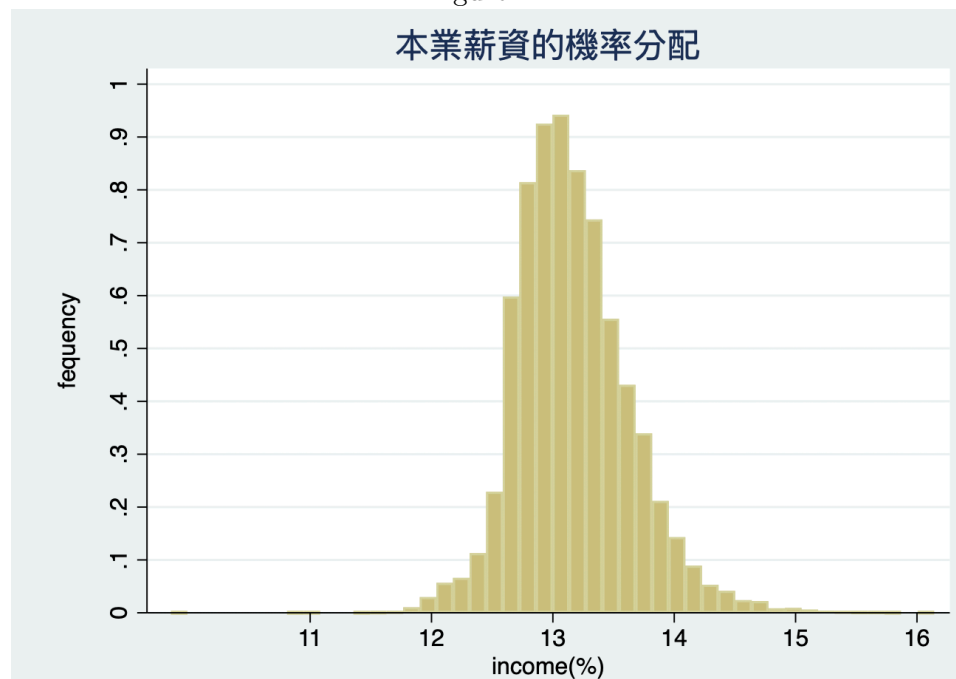
The following figure describes the probability distribution of salary data, where I took the natural logarithm of the salary.

```

1 histogram ln_income, ///
2 title("本業薪資的機率分配") ///
3 xtitle("income(%)") ///
4 ytitle("fequency", margin(medium)) ///
5 xlabel(11(1)16) ///
6 ylabel(0(0.1)1)

```

Figure 4:

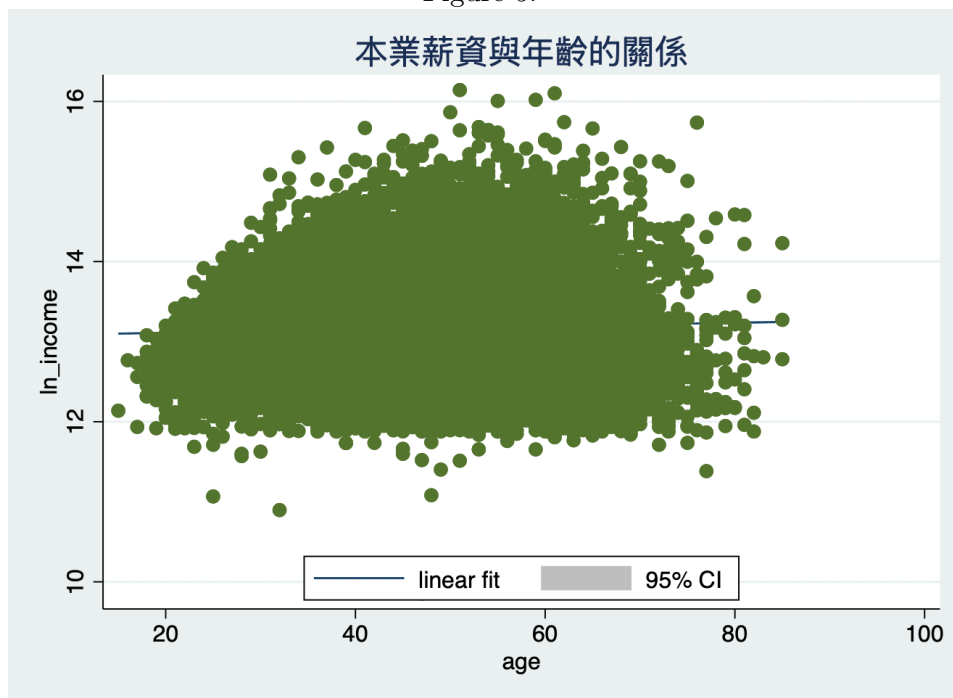


Question 4.2

The following figure describes the relationship between salary and age in the data. It shows that as age increases, salary also increases, indicating a positive correlation between the two.

```
1 graph twoway (lfitci ln_income age) ///  
2 (scatter ln_income age) ///  
3 , title("本業薪資與年齡的關係") ///  
4 ytitle("ln_income") ///  
5 xtitle("age") ///  
6 legend(ring(0) order(2 "linear fit" 1 "95% CI"))
```

Figure 5:



5 Preliminary Analysis

Question 5.1

I want to explore the causal relationship between salary income and age, so I set salary as the outcome and age as the treatment.

$$\ln(\text{income})_i = \alpha + \beta \cdot \text{age}_i + \varepsilon_i$$

```
1 regress ln_income age
```

Figure 6:

Source	SS	df	MS	Number of obs	=	48,851
Model	25.4110285	1	25.4110285	F(1, 48849)	=	115.41
Residual	10755.7617	48,849	.220183867	Prob > F	=	0.0000
				R-squared	=	0.0024
				Adj R-squared	=	0.0023
Total	10781.1728	48,850	.220699545	Root MSE	=	.46924

ln_income	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
age	.0020868	.0001943	10.74	0.000	.0017061	.0024676
_cons	13.06885	.0092859	1407.39	0.000	13.05065	13.08705

First, we can observe that the p-value is highly significant, indicating that the model has explanatory power. Additionally, the treatment (age) is also highly significant, with a positive coefficient, suggesting that as age increases, the model predicts higher income.

Question 5.2

Although the results of the simple regression model are significant, I believe there may be omitted variables, such as occupation, education level and gender. Including these control variables could potentially alter the coefficient for age, weakening its explanatory power.

Question 5.3

The omitted-variable bias formula can be written as:

$$\hat{\alpha} \rightarrow \alpha + \beta \frac{Cov(X, age)}{Var(age)}$$

Where X is an omitted variable (occupation, education level and gender), and age is the treatment. If age and salary are positively correlated, and if factors such as gender, education level, and certain occupations also have a positive correlation with salary, using a simple regression model could lead to an overestimation of the coefficient. In reality, the impact of age on salary may be much smaller than initially perceived.

Question 5.4

This model represents my multiple regression analysis after adding the control variables.

```
1 regress ln_income age gender_dummy1 edu_dummy2-edu_dummy3 job_dummy1-  
    job_dummy12
```

Figure 7:

Source	SS	df	MS	Number of obs = 48,851		
				F(16, 48834) = 2720.03		
Model	5080.44021	16	317.527513	Prob > F = 0.0000		
Residual	5700.73255	48,834	.116736957	R-squared = 0.4712		
				Adj R-squared = 0.4711		
Total	10781.1728	48,850	.220699545	Root MSE = .34167		
ln_income	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
age	.0052366	.0001515	34.56	0.000	.0049396	.0055336
gender_dummy1	.1668003	.0035905	46.46	0.000	.1597629	.1738377
edu_dummy2	.1347836	.0040711	33.11	0.000	.1268041	.1427631
edu_dummy3	.3522409	.0064393	54.70	0.000	.3396199	.3648619
job_dummy1	-1.34711	.2461431	-5.47	0.000	-1.829554	-.8646665
job_dummy2	.6604529	.0474577	13.92	0.000	.5674352	.7534705
job_dummy3	.3424694	.0472636	7.25	0.000	.2498321	.4351067
job_dummy4	.1991342	.0471349	4.22	0.000	.1067492	.2915192
job_dummy5	-.0495797	.0473077	-1.05	0.295	-.1423035	.043144
job_dummy6	-.0800751	.0471729	-1.70	0.090	-.1725346	.0123844
job_dummy7	-.0216927	.0471231	-0.46	0.645	-.1140545	.0706691
job_dummy8	-.0649965	.0471343	-1.38	0.168	-.1573804	.0273874
job_dummy9	-.3150491	.0473374	-6.66	0.000	-.4078311	-.2222672
job_dummy10	.3679886	.0501574	7.34	0.000	.2696793	.4662978
job_dummy11	-.3105969	.0544677	-5.70	0.000	-.4173543	-.2038395
job_dummy12	-.0956752	.1295982	-0.74	0.460	-.3496893	.1583389
_cons	12.61112	.0476672	264.57	0.000	12.51769	12.70455

This result differs from our original expectations (we anticipated that the coefficient for age would decrease), but it is evident that the control variables are statistically significant. The coefficient for males is positive, which aligns with our initial expectations, and both coefficients for education level are positive, indicating that higher education is associated with higher salary expectations. Notably, the coefficient for those with a master's degree is significantly larger than that for university graduates.

Regarding the dummy variables for occupation, most show statistical significance, while the less significant occupations include administrative support personnel, skilled trades workers, machine operators and assemblers, and forestry production workers.

Question 5.5

Gender:

Gender often plays a significant role in determining income due to various factors such as discrimination, occupational segregation, and differences in work experience. Controlling for gender helps isolate the effect of age on income from potential gender-related income disparities.

Education:

Education is a key determinant of income. It influences the types of jobs individuals can attain and their earning potential. By controlling for educational attainment, you account for differences in human capital that could otherwise confound the relationship between age and income.

Job Types:

Different job categories have varying salary ranges. By including dummy variables for specific job types, we control for the impact of occupation on income. This helps to ensure that any observed relationship between age and income is not simply reflecting the differences in earnings across job categories.

Theoretical Justification:

The inclusion of these covariates is supported by economic and sociological theories. For instance, human capital theory suggests that education and job experience directly impact earning potential, while labor market theories highlight the role of gender in wage differentials. This theoretical grounding reinforces the appropriateness of these controls.

Facilitating Interpretation:

By controlling for these variables, we can better interpret the coefficient on age. It allows us to claim that the effect of age on income is independent of the variations introduced by gender, education, and job type. This provides a clearer picture of how age impacts income over time.

In summary, these control variables help to account for confounding factors, thereby making the estimation of the impact of age on income more accurate (within-sample predictions).

Question 5.6

I used the following code to perform a fixed effects (year) multiple regression.

```
1 reghdfe ln_income age gender_dummy1 edu_dummy2-edu_dummy3 job_dummy1-  
job_dummy12, absorb(idd year)
```

Model Overview

Sample Size: A total of 47,318 observations, reduced from 48,851, due to the dropping of 1,533 singleton observations.

F Statistic: $F(16, 32648) = 1535.76$, indicating the overall significance of the model is extremely high ($\text{Prob} > F = 0.0000$).

R-squared: $R^2 = 0.6554$, adjusted $R^2 = 0.5006$, indicating the model explains about 65.54% of the variance in the dependent variable, significantly higher than the previous 47.12%.

Coefficient Estimates

Age: For each additional year of age, the expected income in natural logarithm increases by approximately 0.51185%, slightly lower than before.

Gender: Males are expected to have an income approximately 17.05% higher than females, similar to previous results but with a slightly increased effect.

Education: The impact of education on income remains significant in the fixed effects model, indicating that higher education levels lead to higher income.

Role of Fixed Effects

Controlling for Confounding Variables: By including individual fixed effects (id) and time fixed effects (year), the model can control for those individual characteristics that do not change over time (e.g., birthplace, family background) and time impacts (such as policy changes and economic fluctuations), which helps reduce confounding bias.

Improving Model Accuracy: The fixed effects model provides more reliable estimates of explanatory variables as it better reflects causal relationships between variables by controlling for potential confounding factors.

Comparison of Result Impacts

Increased Explanatory Power: The R^2 value of the fixed effects model has significantly increased, indicating that the model better explains the variation in income after controlling for potential confounding factors.

Changes in Coefficients: The coefficients of certain variables show noticeable changes, indicating that the relationships among variables have become stronger or weaker after considering fixed effects, highlighting the importance of controlling for confounding variables.

Figure 8:

HDFE Linear regression Absorbing 2 HDFE groups						Number of obs = 47,318 F(16, 32648) = 1535.76 Prob > F = 0.0000 R-squared = 0.6554 Adj R-squared = 0.5006 Within R-sq. = 0.4294 Root MSE = 0.3318
ln_income	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
age	.0051185	.0001806	28.35	0.000	.0047646	.0054724
gender_dummy1	.1704982	.0042833	39.81	0.000	.1621027	.1788936
edu_dummy2	.1232887	.0048961	25.18	0.000	.1136921	.1328853
edu_dummy3	.3249877	.0077068	42.17	0.000	.3098821	.3400933
job_dummy1	-1.296531	.2773526	-4.67	0.000	-1.840152	-.7529094
job_dummy2	.6252994	.0594622	10.52	0.000	.5087513	.7418476
job_dummy3	.32171	.059236	5.43	0.000	.2056053	.4378147
job_dummy4	.1876024	.0590792	3.18	0.001	.0718049	.3033999
job_dummy5	-.0502259	.0592547	-0.85	0.397	-.1663672	.0659154
job_dummy6	-.0815002	.0591279	-1.38	0.168	-.1973931	.0343927
job_dummy7	-.0093566	.0590379	-0.16	0.874	-.125073	.1063597
job_dummy8	-.0399688	.0590696	-0.68	0.499	-.1557475	.0758098
job_dummy9	-.286991	.0592662	-4.84	0.000	-.4031548	-.1708271
job_dummy10	.3932334	.0624761	6.29	0.000	.2707779	.5156888
job_dummy11	-.2265532	.068145	-3.32	0.001	-.3601199	-.0929865
job_dummy12	-.1156991	.1563467	-0.74	0.459	-.4221445	.1907462
_cons	12.62548	.0596915	211.51	0.000	12.50848	12.74248
Absorbed degrees of freedom:						
Absorbed FE	Categories	- Redundant	= Num. Coefs			
idd	14650	0	14650			
year	5	1	4			