Panel Data of Individual Wages

——*Xiaoxuan Ma*

Agenda

- Overview
- Descriptive Facts
- Distribution of Log Wage
- Examining and Cleaning Outlier
- Panel Data Models
- Brief Answers about Other Questions
- Appendix

Overview

	exp	wks	bluecol	ind	south	smsa	married	sex	union	ed	black	lwage	id	time
1-1	3	32	no	0	yes	no	yes	male	no	9	no	5.56068	1	1
1-2	4	43	no	0	yes	no	yes	male	no	9	no	5.72031	1	2
1-3	5	40	no	0	yes	no	yes	male	no	9	no	5.99645	1	3
1-4	6	39	no	0	yes	no	yes	male	no	9	no	5.99645	1	4
1-5	7	42	no	1	yes	no	yes	male	no	9	no	6.06146	1	5
1-6	8	35	no	1	yes	no	yes	male	no	9	no	6.17379	1	6
1-7	9	32	no	1	yes	no	yes	male	no	9	no	6.24417	1	7

Panel data characteristics

- A panel of 595 individuals from 1976 to 1982, they have both **cross-sectional** and **time-series** dimensions.
- The data includes **logarithm of wages**, other 3 **continuous variables** and 8 **categorical variables**.
- Panel data are **balanced**, which means all individuals are observed in all time periods.

- Due to the limited time provided, I will **focus on the following questions**:

- 1) Highlight three **descriptive facts** from the data with supporting analysis and graphs.
- 2) Pick a continuous variable of interest, what is the **distribution** of this variable?
- 3) Continuing from above, how would **you examine and clean outlier**?
- 5) **Build a model** from data, share with us the insights and results.

- Also, I'll briefly talk about my thoughts about some other questions.

Descriptive Facts: Means and Variations of Variables

		Standard Deviation		
	Means	Overall	Between	Within
exp	19.85	10.97	**10.79**	**1.98**
wks	46.81	5.13	**3.28**	**3.94**
ed	12.85	2.79	**2.79**	**0**
lwage	6.68	0.46	**0.39**	**0.24**

For panel data, there are 3 kinds of variations:

- **Overall variation**: variation over time and individuals.
- **Between variation**: variation between individuals.
- **Within variation**: variation within individuals over time.

- We need to characterize the variables from two dimensions: **cross-sectional and time-series.**

- The **log wage** is the dependent variable for the further research, it has more between variation than within variation, which means the variation between different persons is larger than the variation of a person over time.
- For the **experience**, the within variation is almost deterministic, because next year people will get one more year of experience, and the between variation is much larger.
- For the **weeks worked**, the standard deviation is almost equally split.
- The **education** variable has zero within variation, so it's time-invariant. That is to say, the education for a person doesn't change over time.

Descriptive Facts: Wage Trends of Males and Females

- From the left graph, we can find the average log wages of both males and females **keep increasing** during the seven years.
- The trends are close to two straight lines with similar slopes, which means the **growth rates** are **stationary** over time, and are **close** for males and females.
- Males' average log wages are **always higher** than females' during the seven years. And the difference doesn't change much.

Notes: The similar analysis can also apply to other attributes. For example, wage trends of the blue-collar vs nonblue-collar.

Descriptive Facts: Samples' Distribution Characteristics

Now, I use **categorical variables** to discover samples' distribution characteristics.

Attributes	bluecol	ind	south	smsa	married	sex	union	black
sample:595	no:290	no:369	no:421	no:213	no:116	male:528	no:377	no:552
	yes:305	yes:226	yes:174	yes:382	yes:479	female:67	yes:218	yes:43

- For 1982, **the most of samples** are not from the manufacturing industry, not from the south, from the standard metropolitan statistical area, married, male, not from the union, not black.

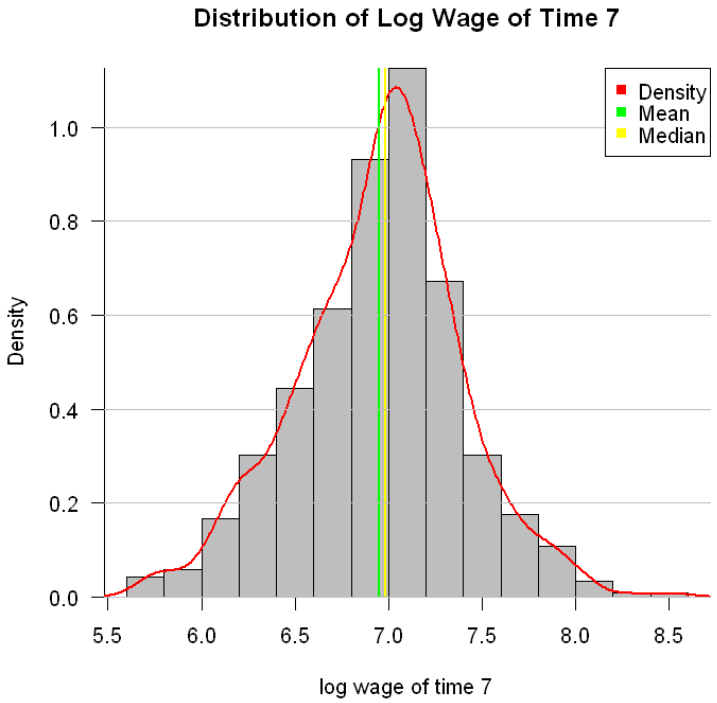
Number of all samples : 595							
not blue collar				blue collar			
49%				51%			
male		female		male		female	
87%		13%		90%		10%	
not black	black	not black	black	not black	black	not black	black
95%	5%	95%	5%	95%	5%	55%	45%

- Having a closer look. The proportions of blue-collar and nonblue-collar are close.
- The proportion of females in nonblue-collar is higher than in blue-collar.
- The proportion of black people in female blue-collar is much higher than in female nonblue-collar.

Notes: The similar analysis can also apply to other attributes. For example, manufacturing industry -> union -> married.

Distribution of Log Wage

Taking the **log wages** from 1982 as an example, let's try to find the distribution.



- From the left graph, we can find the distribution **looks like normal distribution**.
- And the mean < the median < the mode, which presents **slight negative skew**.

- The **Shapiro-Wilk normality test** shows:
  W = 0.992, p-value = 0.003.
- Supposing alpha level is 0.05, then the null hypothesis that the data are normally distributed is **rejected**.

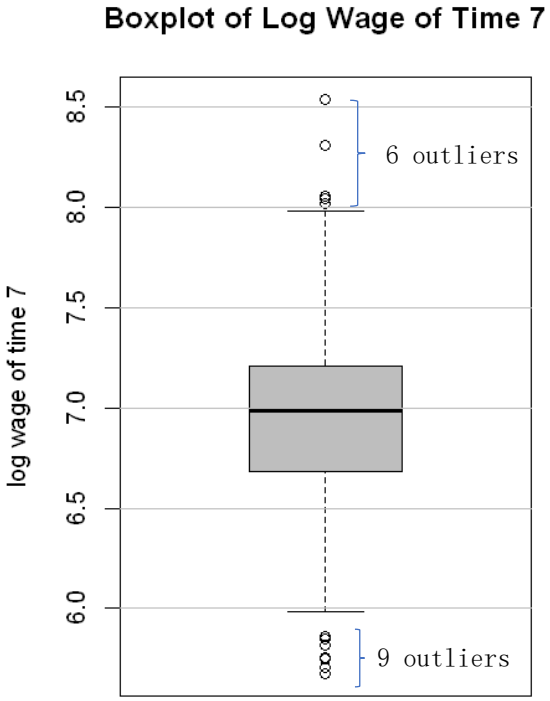
- Skewness = -0.11, Kurtosis = 3.39

- Overall, we can conclude that the log wage is **slightly leptokurtic & negative skew distributed**.

Notes: In this example, the mode means the value with the highest density. And in the Jupyter Notebook, there is an extra QQ-plot.

Examining and Cleaning Outlier

The leptokurtic distribution shows heavy tails, indicating the exist of outliers.



Box Plot

- Boxplots are based on a **five-number** summary: “minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”.
- Outliers are data which are out of **[Q1 -1.5*IQR, Q3 + 1.5*IQR]**.

- Still taking the log wages from 1982 as an example. From the left graph, we can find **6 large outliers** and **9 small outliers**, this is consistent with our observations of the distribution before.

Different methods to deal with outlier:

1. **Drop the outlier records.** Mainly used when doing general analysis about cross-sectional data.
2. **Keep the outlier records.** Mainly used when trying to analyze different behaviors of different groups and analyze special/extreme situations.
3. **Assign a new value**, like mean, mode or last record. Mainly used for time series analysis.
4. **Try a transformation.** For example, try creating a percentile version of the original field.

Notes: Interquartile range (IQR): 25th to the 75th percentile. “minimum” : Q1 -1.5*IQR. “maximum” : Q3 + 1.5*IQR.

Panel Data Models: Introduction

I considered three basic panel data regression models, used package(**plm**) in R.

Pooled OLS Model

- It uses both the between and within variation to estimate the parameters.
- It is obtained by stacking the data over i and t into one long regression with NT observations and estimating it by OLS:

$$y_{it} = \alpha + \mathbf{x}_{it}'\beta + (\alpha_i - \alpha + e_{it})$$

Between Model

- It only uses the between variation (across individuals).
- This is an OLS estimation of the time-averaged dependent variable on the time-averaged regressors for each individual:

$$\overline{y_i} = \alpha + \bar{\mathbf{x}}_i'\beta + (\alpha_i - \alpha + \bar{e}_i)$$

Random Effects Model

- The random effects estimates are a weighted average of the between and within estimates.
- This is an OLS estimation of the transformed model:

$$y_{it} - \hat{\lambda}\overline{y_i} = (1 - \hat{\lambda})\mu + (\mathbf{x}_{it} - \hat{\lambda}\bar{\mathbf{x}}_i)'\beta + \nu_{it}$$

$$\nu_{it} = (1 - \hat{\lambda})\alpha_i + (e_{it} - \hat{\lambda}\bar{e}_i)$$

$$\lambda = 1 - \sigma_e / \sqrt{\sigma_e^2 + \sigma_\alpha^2}$$

Panel Data Models: Results and Insights

Results of 3 descriptive regression models are showed in the table.

	Pooling Model	Between Model	Random Effect Model
Intercept	5.304	5.169	4.298
exp	0.011	0.009	0.052
wks	0.004	0.008	0.002
ed	0.074	0.071	0.107
sex	-0.350	-0.302	-0.330
married	0.062	0.124	-0.073
black	-0.152	-0.139	-0.227
Adj. R-Squared	0.345	0.424	0.383

Model:
lwage ~ exp + wks + ed + sex + married + black

Total number of observations: 595*7
All coefficients are significant(P-value < 0.05).
Hausman Test: one model is inconsistent

- From the table, we can find the **between model** has the highest adjusted R-Squared value, indicating the best model performance. And Hausman Test suggests the random model is inconsistent.
- For the **coefficients**, the three models showed the **similar results**. Experience, weeks worked, education and married have **positive effect** on the log wage. For example, in the between model, the coefficient of ed is 0.071, that means if the average education of a person increases by one year, then the average log wage will increase by 7.1%.
- Sex and black have **negative effect**. In the pooling model, with the same other attributes, if a person is female / black, her log wage will decrease by 35.0% / 15.2%.

Notes: Due to the limited time, I did not perform data cleaning and standardization. For predictive models, we need to split data into fitting and testing sets or perform cross validation.

Brief Answers about Other Questions

6) Find a way to classify this dataset into 3 homogenous groups.

- This is a clustering problem. Popular methods includes K-Means, Hierarchical, There are two important points.
- Turning panel data into cross-sectional data, possible ways: taking average of continuous data or flatting time series.
- How to characterize and interpret the groups?

7) What size of sample would you recommend?

- The Cochran formula allows us to calculate an ideal sample size given a desired level of precision, desired confidence level, and the estimated proportion of the attribute present in the population.
- e is the desired level of precision (i.e. the margin of error),
- p is the (estimated) proportion of the population which has the attribute in question,
- q is 1 – p.

$$n_0 = \frac{Z^2 pq}{e^2}$$

8) How would you select & allocate these samples?

- Stratified sampling
- Cluster sampling

9) What are the pros and cons of sampling?

- Advantages: Saving time, money. Lower requirement of computer.
- Disadvantages: Biasness. Affecting the level of accuracy. Not easy to determine size and methods.

Appendix

Codes in R:

- https://github.com/XiaoxuanMa/Panel-Data-of-Individual-Wages-Analysis/blob/master/Panel%20Data%20Analysis%20Codes%20in%20R.ipynb

Main Reference:

- Data set: Wages Data (https://vincentarelbundock.github.io/Rdatasets/doc/plm/Wages.html)
- Panel Data Econometrics in R: The plm Package
- Sample Size in Statistics (How to Find it): Excel, Cochran's Formula, General Tips

Other Project Sample: Location Recommendation for a New Nike Store (Python)

- Used Google Map API and Foursquare API to collect the data of nearby venues (including the number of restaurants, stadiums, shops, bus stops and so on) of 80 existing Nike stores and 70 unsuitable locations for stores.
- Built a decision tree model based on the data to predict which location in Santa Fe might be appropriate for a future Nike store.
- Made a radar plot to characterize the existing Nike stores and the unsuitable locations, visualized the two predicted locations on a Folium map.

- Link: https://github.com/XiaoxuanMa/Next-Nike-Store-in-New-Mexico/blob/master/Next%20Nike%20Store%20In%20New%20Mexico.pdf

Thank you for reading!

Xiaoxuan Ma
Masters in Mathematical Finance
Boston University Questrom School of Business
mxx@bu.edu | www.linkedin.com/in/xxmbu