The Differences of Micro Data and Macro Data Used in Statistical Analysis: C-D Production Function

Wangyue Li, Ting Dai
Renmin University of China, Beijing 100872, China
Corresponding author: Wangyue Li, e-mail:liwangyue2010@sina.com

Abstract

In empirical studies, we discussed the differences of micro data and macro data used in statistical analysis. Based on the 2008 Economic Census, we not only analyzed the differences of micro data and macro data in one variable descriptive statistics and correlation of two variables, but also we discussed their differences in multiple regression analysis. In descriptive analysis, we discovered that macro data are much closer to normal distribution than micro data, but not the case after logarithm of the data. In the correlation analysis of two variables, the correlation calculated on macro data is higher than correlation calculated on micro data. In the regression model analysis, we used OLS method to estimate C-D production function, and found that when heteroscedasticity and multicollinearity didn't be eliminated, the estimation based on macro data is quite different from the result based on micro data in the economy of scale, the marginal contribution of production factors, and the explanatory power of factors to output. After eliminating heteroscedasticity and multicollinearity, the difference of the estimation in the explanatory power of factors to output still exists. And when we tested whether the model satisfied the conditions of OLS, we discovered that micro data are more prone to producing heteroskedasticity while macro data are prone to producing multicollinearity.

Key Words: C-D production function, descriptive statistics analysis, macro data, micro data, regression analysis

1. Introduction

In modern economics, empirical analysis is highly respected, while the analysis is in the premise of corresponding data conditions. That's to say, data collection and collation is the first crucial step when research purposes are determined. In reality, data can be obtained with different degree of difficulty, however. For researchers, the data on the yearbooks are available. But most of these data are aggregated by region, or by industries which are summarized according to two-digit codes, the underlying data about enterprises are generally difficult to gain. It's no problem analyzing the differences between regions or sectors with these data. Zhang(2011) analyzed the impact of infrastructure on the economic gap among these regions using the panel data of 28 provinces on the provincial-level, ranging from 1989 to 2008. Wang & Zhan(2011) studied how the processing trade to affect the income gap among China's 30 manufacturing industries based on these industrial-level data from 2005 to 2008. Sun & He(2011) used the panel data of 34 industrial sectors in China from 1998 to 2006, to analyze the relationship of R&D among these industries empirically.

However, some other scholars used industrial data to make model, explaining the behavior of enterprises.Wang(2006) collected panel data of China's FDI on industry level, and tried to make regression model to examine the impact of FDI on the capability of independent innovation of China's national enterprises. However, it still deserves more study for the reasonability and effectiveness when explaining enterprise behavior using data on industry level. Input-output and production efficiency are some fundamental problems in the study of economics. Generally, discussions of these issues can not be separated from the production function. No matter macro or micro data are all used to estimate the production function in existing studies.

For example, Cao(2007) estimated the production function using macroeconomic data from 1979 to 2005. And Zhang & Xu(2009)estimated the function with time-series data about our country from 1979 to 2005 and inter-provincial panel data from 1978 to 2005. Jia & Gan(2010) also estimated it in order to analyze the differences of the functions in various regions with macroeconomic data from 1990 to 2005. Besides, Wang & Shi(2008) used the data about 114,838 private enterprises among these industrial enterprises above designated size, which was obtained from the first national economic census to estimate the production function. Base on the function, they analyzed the productivity and investment efficiency of Chinese private enterprises. After that, Mou(2012) used the data covering 1,816,661 enterprises of 38 industries from the second economic census in 2008 to estimate the transcendental logarithmic production function.

However, we should notice that if there are any differences to analyze the same question about production functions when using micro data or macro data. If there are some differences, how significant are they?

Taking these issues into consideration, this paper will aggregate the data from the economic census in 2008, according to the industry code of national economy (The codes are four-digit codes, three-digit codes and two-digit codes respectively). Estimate Cobb-Douglas production function (C-D production function) separately using both the aggregated data on industrial-level and the original data of micro-enterprises. The results are compared to illustrate the statistical differences between the micro data and the macro data.

2. Data and Methods

This article uses the data of industrial enterprises above designated size from the 2008 national economic census for empirical analysis. The total number of enterprises is 421832. But we should make sure that the instructors including the main business income, net fixed assets, as well as the number of employees at the end of the enterprises' accounting time are all greater than 0, as a result, we finally used the 417295 enterprises among these ones as our samples. This paper aggregates the data of these 417,295 enterprises into different levels of aggregation according to the industry code of national economy. The codes are four-digit codes, three-digit codes and two-digit codes, respectively. So that we can analyze the differences among data on enterprise-level and three types of aggregated data on industrial-level.

This paper chose C-D production function to make a model comparing the differences. That is,

$$Y = AK^{\alpha}L^{\beta}$$

In this function, *Y* indicates the main business income, *K* indicates the net fixed assets, *L* is the number of employees at the end of the enterprises' accounting time.

3. Differences in Simple Statistic Between Original Micro-data and Macro-data

First, we calculated the value of simple statistics of the main business income, the results shown in Table 1.

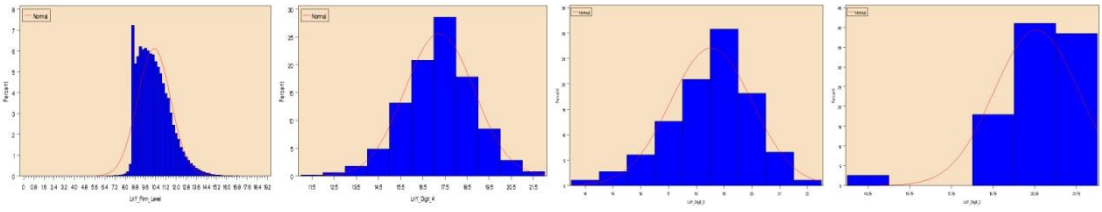
Table 1 Comparison of micro data and macro data in descriptive statistics

Main business income				
Statistics	Enterprise level	Industry level		
		Four-Digit	Three-Digit	Two-Digit
Mean	117526.98	99278180.8	269469348	1257523623
Sd	1224770.57	245474216	421352883	1212413811
CV	1042.12	247.26	156.36	96.41
Skewness	73.90	7.40	3.89	1.14
Kurtosis	8764.29	78.22	23.00	3.47
N	417295	494	182	39

4. Differences in Simple Statistic Between Logarithm of Micro data and Macro data

Logarithm of data is a commonly used method in econometric analysis. Firstly, we made main business income, net fixed assets, and the number of employees logarithm in micro data and macro data, separately. We denoted them by LnY, LnK, LnL, and the histograms of these indicators shown in Figure 1.

In Figure 1, (a1),(a2),(a3) and (a4) represent the histograms of Ln(main business income) in enterprise data, four-digit industry data, three-digit industry data and two-digit industry data, separately.



(a1) (a2) (a3) (a4)

Figure 1 Histogram comparison of micro data and macro data - Based on logarithm of the data

Based on the logarithm of data, the monotonic of some statistics has been damaged, so we calculated the simple statistics of Ln(main business income), Ln(net fixed assests), Ln(the number of employees) and the correlation of each other. After that, we draw Figure 2.

From figure 2, we can see that the skewness and coefficient of variation of Ln(main business income), Ln(net fixed assets) and Ln(number of employees) are monotonous, which means skewness and CV are smaller with increasing levels of data aggregation. However, the kurtosis and correlation aren't monotonous.

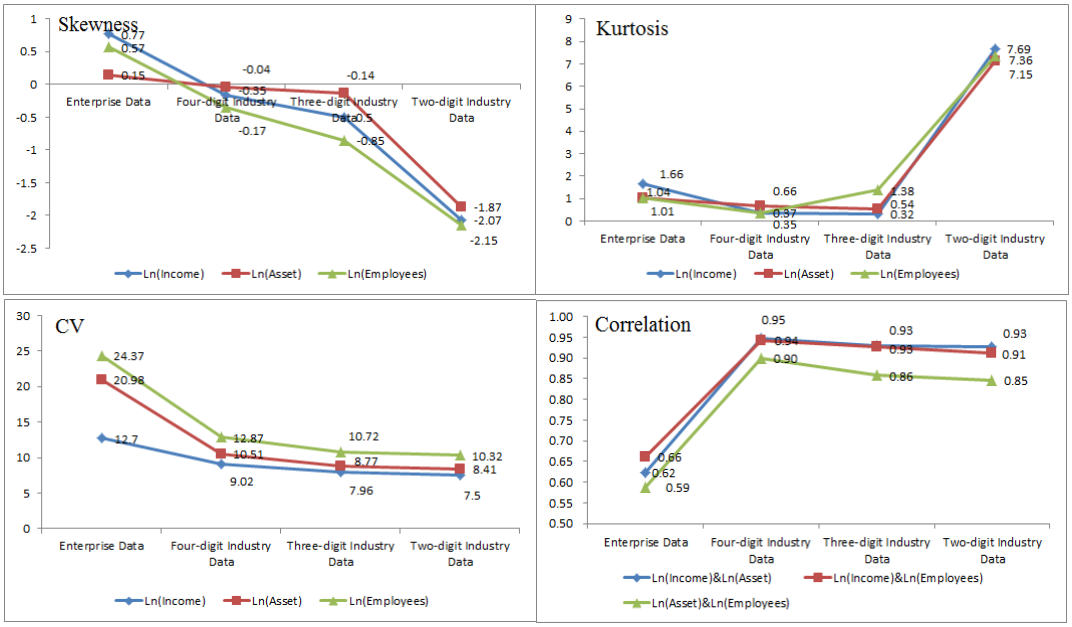


Figure 2 Difference of enterprise data and industry data in descriptive statistics

5. Regression Analysis from Micro Data and Macro Data Differences

In this paper, ordinary linear square (OLS) method is used to estimate the C-D production based on micro data and macro data, separately, and the result is showed in Table 2.

Table 2 Comparison of preliminary regression results based on micro data and macro data[a]

LnY	Firm-level	Industry-level		
		Four-digit	Three-digit	Two-digit
LnK	0.27***	0.49***	0.50***	0.52***
LnL	0.54***	0.51***	0.55***	0.47***
A	5.63***	3.79***	3.21***	3.80***
Adj-R^2	0.52	0.94	0.92	0.91
F	227849	3894.19	1182.06	193.48
n	417295	494	182	39

*** significant at the 0.001 level.

[a] Regression results after eliminating heteroscedasticity and multicollinearity

It can be seen from Table 2, the regression result of micro data and macro data is different in the economy of scale, the marginal contribution rate and the explanatory power of factors to output, before elimination of heteroscedasticity and multicollinearity.

The residuals of the regression are tested in this section. The method of OLS needs to meet Gauss-Markov condition[12], while Residual autocorrelation doesn't exist in cross-sectional data in generally and the endogenous doesn't exist in C-D production function. Consequently, we test the normality, homoscedasticity, and multicollinearity of residuals, and the test results showed that we should eliminate heteroscedasticity in enterprise data and the multicollinearity in industry data. And then we used OLS methods again to estimate the C-D production function, the result shown in Table 3.

Table 3 Comparison of Final Regression Result based on Micro Data and Macro Data

LnY	Firm-level	Industry-level		
		Four-digit	Three-digit	Two-digit
LnK	0.27***	0.39***	0.41***	0.40***
LnL	0.54***	0.44***	0.45***	0.42***
A	5.63***	6.18***	5.94***	6.83***
Adj-R^2	0.52	0.94	0.92	0.91
F	227849	3894.19	1182.06	193.48
n	417295	494	182	39

*** significant at the 0.001 level;

[a] Regression results after eliminating heteroscedasticity and multicollinearity

6. Conclusions

This article summarized the statistical differences of micro data and macro data based on 2008 Economic Census data, which is showed in Table 4.

Table 4 Difference of Micro Data and Macro Data in Statistical Analysis

Difference		Micro data	Macro data
Descriptive statistical analysis	1. Sample size	Large	Small
	2. Mean	Small	Large
	3. Standard deviation	Small	Large
	4.CV	Large	Small
	5.Skewness (original data)	Large (Skewed to the right)	Small (Skewed to the right)
	5.Skewness (logarithm of data)	Positive (Skewed to the right)	Negative (Skewed to the left)
	6.Kurtosis (original data)	Large	Small
	6.Kurtosis (logarithm of data)	Uncertainty	
	7. Correlation	Small	Large
Regression Analysis	1. Estimated parameters	Not equal	
	2. Significant of parameters	Yes	Yes
	3. R^2	Small	Large
Model testing	0. Normality test	Not Pass	Not Pass
	1. Heteroscedasticity	Large	Small
	2. Residual autocorrelation	——	
	3. Multicollinearity	Small	Large
	4. Endogenous explanatory variables	——	

This article is based on are 2008 Economic Census data on the production function analysis to illustrate the difference of micro data and macro data used in regression. And as for other functions, this conclusion still needs to be proved. In addition, we used the data collected by Zhongguancun Haidian Park and got the similar conclusion on the difference of micro data and macro data, but there is somewhat difference between four-digit, three digit and two digit industry data. However, as for other data, we still need to further our study.

References

Guang-nan Zhang. (2011)"The project of poverty alleviation through transport construction and regional economic disparities in China: evidence from provincial panel data from 1989 to 2008", *Journal of Finance and Economics*, 08, 26-35.

Huai-min Wang, Chun-long Zhan. (2011) "The income gap between the processing trade and industry-empirical studies based on panel data of 30 industries in China", *World Economy Study*, 08, 44-48.

Hui-huang Sun, Shu-feng He. (2011) "Duplicity of R&D, outside technological opportunity and the R&D relationship between industries ——An empirical study based on the panel data of industries in China" , *Scientific Research Management*, V32, 23-28.

Hong-ling Wang, et al. (2006) "An observation and empirical study of R&D behavior of Chinese manufacturing firms: based on a survey of the manufacturing firms in Jiangsu province", *Economic Research Journal*, 02, 44-55.

Ji –yun CAO. (2007) "The aggregate production function and contribution rate of technical change to economic growth in China", *Quantitative & Technical Economics*, 11, 37-46.

Shang-feng Zhang, Bing Xu. (2009) "Production functions with time-varying elasticities and under the catch-up consensus: total factor productivity", *Economics (quarterly)*, 02, 551-568.

Nan Jia, Li Gan. (2010) "Heterogeneous production functions and regional disparity". *Nankai Economic Research*, 01,19-35.

Zheng Wang, Jinchuan Shi.(2008) "Productivity performance and investment efficiency of China's private enterprises", *Economic Research*, 01, 114-159.

Jun-lin Mou. (2012) "The comparison analysis of productive efficiency between state-owned and non-state-owned industry enterprises-based on 2008 China economic census data", *Economic Survey*, 03,55-59.

Zi-nai Li. (2005) *Econometrics*. Higher Education Press, Beijing.

Shapiro, S. S. and Wilk, M. B. (1965) "An Analysis of Variance Test for Normality (Complete Samples)", *Biometrika*, 52, 591–611.

White, H. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity ", *Econometrics*, 48, 817–838.

Breusch, T. S. and Pagan, A. R. (1979) "A Simple Test for Heteroscedasticity and Random Coefficient Variation", *Econometrica*, 47, 1287–1294.

Xiao-qun He, Wen-qing Liu.(2007) *Applied Regression Analysis*. Renmin University of China Press, Beijing.