CHAPTER 4. METHODOLOGY AND VARIABLES

This paper uses panel data analysis to examine the determinants of regional foreign direct investment (FDI) distribution across provinces in China and over time (from 1997 to 2006). Moreover, this paper adopts the ordinary least squares model to avoid multicollinearity.[55] Therefore, in this chapter, this paper will interpret the ordinary least squares model. Subsequently, this paper expresses the empirical results of this model in order to investigate how region's political power in the central government affects regional FDI in China after 1997.

4.1 Empirical Model

Panel data analysis is adopted because this paper examines the determinants of FDI distribution across provinces and over time. A major advantage of using the panel data method, as pointed out by Hsiao (1986), is to resolve or reduce the magnitude of a key econometric problem that often arises in empirical studies, namely, the omitted (mis-measured, not observed) variables that are correlated with explanatory variables.

Generally, data sets are composed by two dimensions. One dimension is time, and the other is the cross-section dimension. There are assumed to be m cross-sectional units and T time periods, for a total of $n = mT$ observations. The traditional equation is:

$$Y_{i,t} = X_{i,t} + \varepsilon_{i,t} \qquad i = 1...N \qquad t = 1...T \qquad (4\text{-}1)$$

where $X_{i,t}$ is a $1 \times k$ vector of observations on explanatory variable. If sum of each

[55] The methodology of this main study refer to Greene (2003).

expectation $\varepsilon_{i,t}$ equal zero, ordinary least squares could be used to estimate. Panel data not only owns dynamic characteristic of time-series data but also has cross-section data in order to express different properties between samples. Compared with cross-section data, panel data analysis can observe the dynamic variation which cross-section data can not. Moreover, panel data also use fixed effects model to analyze fixed effects of observations and to understand individual characteristics. By using panel data analysis, this study can decrease estimated bias of population parameters. Consequently, if we want to analyze long-term determinants of some observations, the adoption of panel data can obtain more correct and conscientious results. It is exactly what purpose this paper needs.

In addition, this paper adopts panel data of 30 provinces in China from 1997 to 2006.[56] Since policies between these provinces have regional variations and the Chinese government allocates resources to 30 provinces differently, regional economic growth in 30 provinces may have different impacts. If this paper neglects regional-specific effects in regression models, it might be biased in estimation results.[57] Thus, when opting econometric method, it is proper to use fixed effects model that could control regional-specific diversities to estimate. However, limited observations and excess provincial parameters make regression model multicollinearity and inappropriate results so this paper shifts the fixed effects model to the ordinary least squares model.[58]

The population model of the study is linear in its parameters,

[56] The data of Tibet is so incomplete that we could not use it thoroughly. We fulfill this paper by neglecting Tibet. Thus, there are only 30 provinces of China in this paper.

[57] The difference between these areas also includes geographic and cultural diversities.

[58] Hyclak (1996) used panel data of 200 cities in U.S.A to study the influence of structural changes of labor demand on the rate of unemployment. Liu and Huang (2002) discussed about the determinants of different rate of unemployment between Taiwan cities and countries. Kao, Huang and Pan (2005) referred to the relationship between efficiency of regional government in China and FDI attraction. These studies have the same problem with this paper in choosing estimation methods.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \quad (4\text{-}2)$$

where $y, x_1, x_2, x_3, ...x_k$ are observable random scalars (that is, we can observe them in a random sample of the population), ε is the unobservable random disturbance or error, and $b_0, b_1, b_2, ...b_k$ are the parameters (constants) this paper would like to estimate.

The error form of the model in equation (4-2) is useful for presenting a unified treatment of the statistical properties of various econometric procedures. Nevertheless, the steps one uses for getting to equation (4-2) are just as important. A structural equation can be obtained from an economic model, or it can be obtained through informal reasoning.[59] Sometimes the structural model is directly estimable. Other times we must combine auxiliary assumptions about other variables with algebraic manipulations to arrive at an estimable model. In addition, we will often have reasons to estimate nonstructural equations, sometimes as a precursor to estimating a structural equation.

The error term ε can consist of a variety of things, including omitted variables and measurement error. The parameters b_j hopefully correspond to the parameters of interest, that is, the parameters in an underlying structural model. Whether this is the case depends on the application and the assumptions made. The key condition needed for regression model to consistently estimate the b_j (assuming this paper has available a random sample from the population) is that the error (in the population) has mean zero and is uncorrelated with each of the regressors:

$$E(\varepsilon) = 0, \qquad Cov(x_j, \varepsilon) = 0, \qquad j = 1,2,...k \quad (4\text{-}3)$$

[59] Goldberger (1972) defined a structural model as one representing a causal relationship, as opposed to a relationship that simply captures statistical associations.

The zero-mean assumption is for free when an intercept is included, and we will restrict attention to that case in what follows. It is the zero covariance of ε with each x_j that is important. Complete regression model includes not only equation (4-2) but also 5 basic assumptions or classical assumptions. Regression model relies on certain assumptions:

1. Normality ε is normally distributed, that is $\varepsilon \overset{i.i.d}{\sim} N(0,\sigma^2)$.

2. Zero-mean: an expected value of ε is zero, that is $E(\varepsilon)=0$.

3. Homocedasticity: ε has the same variance at all values of x_j, that is $Var(\varepsilon)=\sigma^2$.

4.Unautocorrelation: values of ε are independent of one another, that is $Cov(\varepsilon_t,\varepsilon_s)=0, t\neq s$.

5. ε is uncorrelated with x_j, that is $Cov(x_j,\varepsilon)=0$.

Complete linear regression model includes equation (4-2) and 5 basic assumptions, which is called classical normal linear regression model. The former 3 assumptions mean that the error term ε is a random variable. The mean of ε is zero and the covariance of ε is unknown. The fourth assumption emphasizes different error terms are independent and uncorrelated, that is the error terms are mutual independent random variables. In addition, if x_j is correlated with the error terms, it means x_j is correlated with unmeasured variables that are influencing y. Since we cannot purge their influence from the effect of x_j on y, we will consistently over-estimate $\beta_j\,(j=1,2,...k)$.

In order to explore the role of region's political power in the central government

in regional FDI in China, this paper primarily collects all related variables from various years of the *China Statistical Yearbooks*. In addition, the *Chongqing Statistical Yearbooks* are also used to complement related variables if necessary.[60] Due to the data limitations, the research period covered by this paper is limited from 1997 to 2006. The data set used in this paper is a panel data set of China's 30 provinces/regions form 1997 to 2006.[61] The advantage of panel data is that, compared with time series or cross-sectional data, such data contains more information and observations. Due to the larger sample size, the use of panel data can increase the number of degrees of freedom and make the estimation more accurate.

According to the literature, regional FDI in China is a function of the degree of a region's political power (CCPB), gross provincial product (GPP), gross provincial product per capita (PER), provincial average wage (WAGE), length of paved highway(km) divided by area (1000km^2) (HIWAY), the degree of provincial openness(OP), and the accumulated foreign direct investment (AFDI). Thus, this paper establishes a function of regional FDI in China and other independent variables presented as follows:

$$FDI = f(GPP, PER, WAGE, HIWAY, OP, AFDI, CCPB, TimeTrend) \tag{4-4}$$

Function (4-4) can be further expressed by regression model presented as (4-5):

$$\begin{aligned} \log FDI_{i,t} = {} & \beta_{0\,i,t} + \log \beta_1 GPP_{i,t-1} + \beta_2 \log PER_{i,t-1} + \beta_3 \log WAGE_{i,t-1} \\ & + \beta_4 HIWAY_{i,t-1} + \beta_5 OP_{i,t-1} + \beta_6 AFDI_{i,t-1} + \beta_7 CCPB_{i,t-1} \\ & + \beta_8 COASTAL + \beta_9 T + \varepsilon_{i,t} \end{aligned} \tag{4-5}$$

In equation (4-5), i represents the i^{th} province and t means the t^{th} year,

[60] Chongqing City is a municipality that has been detached from the Sichuan Province since 1997. Therefore, its information prior to 1997 must be collected from the *Sichuan Statistical Yearbook* and *Chongqing Statistical Yearbook* in order to obtain a complete and true data set.

[61] The data of Tibet is incomplete that we could not use it thoroughly. We fulfill this paper by neglecting Tibet. Thus, there are only 30 provinces of China in this paper.

where i=1, 2, …, 30 and t=1997, 1998, …, 2006.[62] T expresses time trend to investigate whether time affects the determinants of regional FDI or not. $FDI_{i,t}$ denotes the amount of foreign direct investment of the i^{th} region in year t. In addition, $\varepsilon_{i,t}$ represents the error term with zero mean and variance σ^2.[63] It is worth noting that all dependent variables are lagged one year in order to avoid any potential endogeneity problems between any dependent and independent variables. In order to eliminate any fluctuations in prices, all value variables in this paper are adjusted by the CPI deflator (base year=1997). The definitions, descriptive statistics, and expected signs of all of the variables are listed and described in Table 10 and analyzed in the paragraphs below.

4.2 Data and Variables

$GPP_{i,t}$ is gross provincial product of the i^{th} region in year t, which denotes the size of a province's economy.[64] According to Coughlin and Segev (2000), numerous studies of FDI location have used a measure of economic size,[65] suggesting that larger economies attract more investment because there is more potential market demand. On the other hand, Sun et al. (2002) mentioned the larger the market size of a particular province is, other things being constant, the more FDI the province should attract. Therefore, this paper expects $GPP_{i,t}$ to be positive. In addition, $PER_{i,t}$, which is gross provincial product per capita of the i^{th} region in year t and represents purchasing power and living level of a region is a positive impact on FDI.[66] Also,

[62] The data of Tibet is so incomplete that we could not use it thoroughly. We fulfill this paper by neglecting Tibet. Thus, there are only 30 provinces of China ($i = 1,2,...30$) in this paper.

[63] It is normally distributed and also satisfies the general assumption of an independent and identical distribution (iid).

[64] Sun et al. (2002) adopted gross domestic product to measure market demand and market size.

[65] Broadman and Sun (1997) also found GPP to be a positive, statistically significant determinant of FDI in China through 1992.

[66] Sun et al. (2002) used GDP, GDP per capita, retail sales, and retail sales per capita to capture

$PER_{i,t}$ is expected to a positive variable in this paper.

Labor cost, as measured by $WAGE_{i,t}$, which expresses provincial average wage of the i^{th} region in year t, is expected to be a negative factor to FDI. All else equal, higher wages should deter foreign investment.[67] However, since higher wages might be due to higher productivity, ideally employee productivity should be controlled for in the regression analysis.[68] Past studies of FDI have found somewhat conflicting results for the effect of wages, but this is likely due to some extent to the omission of a productivity variable.[69]

$HIWAY_{i,t}$, denoting a region's infrastructure of the i^{th} region in year t, is the length of paved highway divided by area. In this paper, $HIWAY_{i,t}$ is expected to be a positive influence on FDI. Both Broadman and Sun (1997) and Coughlin and Segev (2000) found transportation infrastructure to have positive, statistically significant relationships with FDI inflows.

$OP_{i,t}$ is the degree of provincial openness of the i^{th} region in year t which is equal to the amount of import divided by gross provincial product (GPP). As pointed out by Sun et al. (2002), a more open economy attracts FDI because it welcomes foreign capital and foreign investors are more familiar with the host economy. Thus,

---

demand and size effect.

[67] However, as pointed out by Sun et al. (2002), higher wage may well reflect higher labor quality. Hence, it is conceivable that wages in those provinces that can attract relatively more FDI can be higher, too.

[68] Coughlin and Segev (2000),"Foreign Direct Investment in China: A Spatial Econometric Study." *World Economy*, 23(1),1-23.

[69] For example, using state level data, Luger and Shetty (1985), Coughlin et al. (1990 and 1991), and Friedman et al. (1992) found wages to be a negative determinant of FDI in the United States, as expected. Ondrich and Wasylenko (1993), however, did not find a statistically significant relationship. But among these studies only Friedman et al. (1992) explicitly controlled for productivity, which was a positive determinant of foreign plant location.

this paper expects $OP_{i,t}$ to be positive. Furthermore, $AFDI_{i,t}$ represents agglomeration effect. The level of agglomeration of a particular province expected to be positively related to the FDI. Following Wheeler and Mody (1992), this paper uses cumulative foreign investment to capture the agglomeration benefits. The cumulative FDI amount captures the possible "herding effect" among foreign investors. Other than the absolute measure, this paper constructs a relative measure, the ratio of the cumulative FDI relative to the cumulative domestic investment, CFDI/CINV,[70] standing for the level of FDI.

Most of all, $CCPB_{i,t}$ represents the index of provincial political power of the i^{th} region in year t in the Central Committee of the Chinese Communist Party or the Central Political and Legislative Affairs Committee. According to Hsu (2007), [71] provincial political power affects the resource of political allocation in the central government positively. Thus, this paper expects $CCPB_{i,t}$ to be a positive variable. In addition, this paper investigates the relationship between provincial political power and time (year) which uses $CCPB \times T$ to examine whether the degree of $CCPB$ increases or decreases by years and affects FDI or not.

Finally, this paper adopts a dummy variable to differentiate among provinces that lie on the coastal and inland areas.[72] The role of this variable is used to control the influence of determinants, which may differ systematically between the coastal and non-coastal provinces. These may include superior access to sea-routes,

[70] Sun et al. (2002) used CFDI/CINV as an explanatory variable to estimate the level of agglomeration of a particular province.
[71] Hsu (2007), "The Horizontal and Vertical Competition in China's Central-Provincial Relations, 1993-2004: Issue Linkage between Elite Politics and Fixed Asset Investment." *Mainland China Studies*, 50(2), 1-33.
[72] Coughlin and Segev (2000) also used a dummy variable to differentiate among provinces that lie on the coast and those that do not.

geographical proximity to foreign countries, and the increased experience of coastal provinces in absorbing FDI, especially as many of these provinces have enjoyed preferential treatment during China's early experimentation with FDI. Broadman and Sun (1997) found a statistically significant preference for investing in the coastal provinces. Dummy variable, *COAST*, takes the value of one for the 12 coastal provinces, and zero otherwise.[73] This paper expects *COAST* to be a positive variable. In addition, this paper also adopts 8 dummy variables of time trend, they are *Y99*, *Y00*, *Y01*, *Y02*, *Y03*, *Y04*, *Y05* and *Y06*, separately representing the year of 1999, 2000, 2001, 2002, 2003, 2004, 2005 and 2006.

[73] The coastal provinces are Beijing, Jiangsu, Hainan, Guangxi, Guangdong, Tianjin, Fujian, Zhejiang, Shandong,Shanghai, Hebei and Liaoning.

Table 10：Descriptive Statistics and Definitions of Variables

Variables	Definitions	Mean (S.D.)	Expected Sign
FDI	The amount of foreign direct investment (100 million RMB)	3.918 (1.733)	
GPP	Gross provincial product (100 million RMB)	7.912 (0.938)	+
PER	Gross provincial product per capita (RMB)	9.035 (0.601)	+
WAGE	Provincial average wage (RMB)	9.247 (0.393)	−
HIWAY	The length of paved highway divided by area (1/km)	0.354 (0.231)	+
OP	The degree of provincial openness, import/GPP (%)	13.937 (18.286)	+
AFDI	The level of foreign direct investment, CFDI/CINV (%)	9.105 (9.374)	+
CCPB	The degree of a region's political power	26.657 (28.920)	+
$CCPB \times T$	The relationship between provincial political power and time	-0.150 (0.358)	?
COASTAL	Dummy variable (Beijing, Jiangsu, Hainan, Guangxi, Guangdong, Tianjin, Fujian, Zhejiang, Shandong, Shanghai, Hebei, Liaoning = 1;others=0)	0.400 (0.491)	+
*Y*99	Dummy variable (1999=1,others=0)	0.111 (0.315)	?
*Y*00	Dummy variable (2000=1,others=0)	0.111 (0.315)	?
*Y*01	Dummy variable (2001=1,others=0)	0.111 (0.315)	?
*Y*02	Dummy variable (2002=1,others=0)	0.111 (0.315)	?
*Y*03	Dummy variable (2003=1,others=0)	0.111 (0.315)	?
*Y*04	Dummy variable (2004=1,others=0)	0.111 (0.315)	?
*Y*05	Dummy variable (2005=1,others=0)	0.111 (0.315)	?
*Y*06	Dummy variable (2006=1,others=0)	0.111 (0.315)	?

Source: *China Statistical Yearbooks* and *Chongqing Statistical Yearbooks*.
Note: The numbers in the parentheses are standard deviations.