universität wien

DISSERTATION

Titel der Dissertation

„Essay in Econometrics“

Verfasserin

Miaomiao Yan

angestrebter akademischer Grad

Doctor of Philosophy (PhD)

Wien, 2012

Studienkennzahl lt. Studienblatt:	A 094 140
Dissertationsgebiet lt. Studienblatt:	Volkswirtschaftslehre (Economics)
Betreuer:	Univ.-Prof. Dipl.-Ing. Dr. Robert M. Kunst

Acknowledgement

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I want to thank the Department of Economics at the university of Vienna for giving me permission and (financial) support to commence this thesis in the first instance, to do the necessary research work.

I am deeply indebted to my supervisor Prof. Robert Kunst from the university of Vienna and Prof. Christine Zulehner whose help, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis. Furthermore, my colleagues from the department supported me in my research work. I want to thank them for all their help, support, interest and valuable hints.

Especially, I would like to give my special thanks to my family whose patient love enabled me to complete this work.

Contents

Chapter 1

Bootstrap Overidentifying Test in Dynamic Panel Data Models

1.1 Introduction

Many economic relationships are in nature dynamic, i.e. the effect of the past on the present can not be neglected. And more and more application of the panel data models also tells us that panel models allow the researcher to better understand the dynamics of adjustment. The dynamic panel data(DPD) models seem to be good consideration. However, with lagged dependent variables as explanatory variables, strict exogeneity of regressors no longer holds, making the estimation of DPD models more complicated than static models. It has been shown that several kinds of estimators under GMM framework are found to be better in this case. Many researchers, such as Anderson and Hsiao (1982), Arellano and Bond (1991), Ahn and Schmidt (1995, 1997) etc. explored various GMM estimators, which are shown to perform well in dynamic panel models, while during the GMM estimation, the number of instruments is one of the crucial parameters to choose. Hahn (1997) also showed that the GMM estimator with

an increasing set of instruments attains the semiparametric efficiency bound of the model.

Does this mean that the more instruments there are, the more efficient GMM estimators are? No, not really. Tauchen (1986) found that with $T = 50$ or 75, there is a tradeoff between efficiency and bias as the number of moment conditions increase and thus he concluded that short lag lengths are preferred to longer ones in small samples. Andersen and Sϕrensen (1996) argued GMM using too few instruments is just as bad as using too many instruments. Ziliak (1997) then asks whether the tradeoff is still binding in large samples, say, larger than 500 and he performs an extensive set of Monte Carlo experiments for a dynamic panel data model and finds that the bias/efficiency tradeoff still exists. Alonso-Borrego and Arellano (1999) are also motivated by the finite sample bias in panel data instrumental variable estimators when the instruments are weak. They found that the dynamic panel model generates many overidentifying restrictions even for moderate T.

So it seems crucial to select the proper size of instruments and also that the test of overidentification should be considered. The most commonly used test of overidentifying restrictions is initiated by Sargan (1958), then improved by Hansen (1982) and it is also called the J test. According to them, the test statistic is the number of observations multiplied the minimized GMM criterion, which asymptotically has a χ^2 distribution with the degree of freedom equal to the number of overidentifying restrictions. However, under the panel framework, while this test is imposed, only the GMM minimand is used as the test statistic is supposed to have the same asymptotic distribution. One of the reasons why the number of observations in the panel data model is absent is that usually NT is so large that it will make the test statistic always large, thus increasing the size of the test. However, Bowsher (2002) found that in AR(1) dynamic panel data models, Sargan test based on the full instrument set as Blundell and Bond (1998) behaves so badly that it never rejects the null when the number of

moment conditions becomes too large for a given value of N even for moderate values of T. So the test can be very undersized and possesses extremely poor power properties.

Bootstrap is a good consideration as an alternative to inference based on parametric assumptions when those assumptions are in doubt. Just as Hall stated in his book, "Bootstrap methods are potentially superior to large sample techniques for small sample sizes."

In the framework of the panel data model, Kapetanios (2008) suggested resampling the whole cross-sectional units with replacement which is proved to provide asymptotic refinements and outperforms block bootstrap in case where there is no cross sectional dependence but temporal dependence. Bischof (2009) showed cross-sectional bootstrap is a powerful alternative for calculating standard errors when the variance is inadequate for addressing difficult pattern of serial correlation, unless $T >> N$ or in datasets where contemporaneous correlation is the most pressing issue.

There is another crucial parameter to choose during GMM estimation which is the weighting matrix. While heterogeneity and autocorrelation are one of the typical characteristics for most macroeconomic and some financial series. Economic theory often provides information about the specification of moment conditions, but not necessarily about the dependence structure of the moment conditions. This has prompted the development of the class of Heteroscedasticity and Autocorrelation Consistent (HAC) Covariance estimators which are consistent under relatively weak assumptions on the dependence structure of the process. Additionally, Inoue and Shintani (2005) showed that the overlapping block bootstrap based on HAC covariance matrix estimators provides asymptotic refinements. In this paper a cross-sectional bootstrap based on HAC covariance matrix estimator will be conducted in a dynamic panel DGP and only linear moment conditions are considered.

The chapter is organized as follows: in section 2 I will first introduce a general dynamic

panel data model and the HAC covariance matrix GMM estimator, then make a claim for the overidentifying test in models without cross-sectional dependence; section 3 mainly deal with the bootstrap, describing the procedure and showing the asymptotic properties of it; Monte Carlo simulations are conducted in section 4 and section 5 will conclude.

1.2 The Model and Overidentifying Restriction Test

1.2.1 The Model and HAC Covariance Estimator

Consider a general panel model

$$y_{it} = x_{it}'\theta + u_{it}, \qquad i = 1, ..., N; t = 1, ..., T \tag{1.1}$$

where $x_{it} = (x_{1it}, ..., x_{pit})$ is the explanatory variables of dimension p, u_{it} is the error term and θ is the parameter of our interest. For GMM estimation we should also introduce a $q-$dimensional instrument variable z_{it} which satisfies

$$E(z_{it}u_{it}) = 0, \tag{1.2}$$

In dynamic panel data models x_{it} is comprised of lagged values of y_{it} and exogenous variables, z_{it} is mainly lagged value of y_{it} and q should be larger than p for overidentification.

We also define $\mathbf{Y} = (y_1, ..., y_i, ..., y_N) = (y_{.1}, ..., y_{.t}, ..., y_{.T})'$, $\mathbf{X} = (x_1, ..., x_i, ..., x_N) =$

$(x_{.1}, ..., x_{.t}, ..., x_{.T})'$, $\mathbf{u} = (u_1, ..., u_i, ..., u_N) = (u_{.1}, ..., u_{.t}, ..., u_{.T})'$, $\mathbf{Z} = (z_1, ..., z_i, ..., z_N) = (z_{.1}, ..., z_{.t}, ..., z_{.T})'$, where $y_i = (y_{i1}, ..., y_{it}, ..., y_{iT})'$, $x_i = (x_{i1}, ..., x_{it}, ..., x_{iT})'$ and $u_i = (u_{i1}, ..., u_{it}, ..., u_{iT})'$.

Equation (2) is also named sample moment condition in GMM estimation and in this paper only linear moment conditions are considered. The 2-step GMM estimation of θ using HAC covariance matrix is based on the sample moment condition in equation (1). We first obtain the first-step estimator $\tilde{\theta}$ by minimizing

$$\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}u_{it}\right]' W_{NT} \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}u_{it}\right] \tag{1.3}$$

where W_{NT} is a $qNT \times qNT$ positive semi-definite matrix, which may be a simple but not optimal weighting matrix, for example, the identity matrix I. And also estimate $\hat{S}_{HAC}$ based on the residuals as

$$\hat{S}_{HAC} = \hat{\Gamma}_0 + \sum_{j=1}^{\mathfrak{l}} \omega_{j,\mathfrak{l}}(\hat{\Gamma}_j + \hat{\Gamma}_j'),$$

where $\mathfrak{l}$ denotes the lag truncation parameter, $\omega_{j,\mathfrak{l}}$ is the kernel function which weights down autocovariance estimates with large j. $\omega_{j,\mathfrak{l}} = 0$ for $j > b(\mathfrak{l})$ with the bandwidth $b(\mathfrak{l}) \to \infty$ as $\mathfrak{l} \to \infty$. Secondly, estimate the second-step estimator $\hat{\theta}$ by minimizing

$$\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}u_{it}\right]' \hat{S}_{HAC}^{-1} \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}u_{it}\right] \tag{1.4}$$

and the estimator can be written as

$$\hat{\theta} = \left(X'Z\hat{S}_{HAC}^{-1}Z'X\right)^{-1} X'Z\hat{S}_{HAC}^{-1}Z'Y.$$

The properties of several HAC estimators are analyzed and compared in the literature,

such as White (1987), White and Domowitz (1984) and Donald and Andrews (1991). They have established the consistency of such estimators, while in Andrews (1991), the consistency applies to more general conditions with respect to the class of kernels with unbounded support and to allowable rate of increase of bandwidth parameter sequences.

1.2.2 The Overidentifying Restriction Test: J test

In non-panel data models, the J test is based on the sample moment conditions, which is also the null hypothesis

$$E[z_i u_i(\hat{\theta})] = 0, \qquad i = 1, ..., N.$$

And the J test statistic is

$$\mathbf{J} = NQ(\hat{\theta}) = N^{-1}\mathbf{u}(\hat{})'\mathbf{Z}W\mathbf{Z}'\mathbf{u}(\hat{})$$

where $Q(\hat{\theta})$ is the GMM minimand and under the null hypothesis, $\mathbf{J}$ has a chi-square distribution with the degree of freedom equal to the number of overidentifying restrictions.

While in a panel data model, the dimensions of $\mathbf{u}$ and $\mathbf{Z}$ are $NT \times 1$ and $NT \times q$ respectively. In order to get nice asymptotic property according to the central limit theory and its lemma, the J test statistic should be set as

$$\mathbf{J} = NTQ(\hat{\theta}).$$

However, in almost all panel literatures such as Baltagi(2004), when J test is imposed, only the GMM minimand is tested, i.e. without NT in $\mathbf{J}$ above. And those J test

statistics behave usually too badly to obtain nice size or power property.

I would claim for a $N \times T$ cross-sectional independent panel data set, the J test statistic should be

$$\mathbf{J} = NQ(\hat{\theta}) = N^{-1}\mathbf{u}'(\hat{})\mathbf{Z}W\mathbf{Z}'\mathbf{u}(\hat{}). \tag{1.5}$$

To show the asymptotic properties of $\mathbf{J}$, several assumptions should be made as follows

- ASSUMPTION 2.1 The process $(x'_{it}, y_{it}, z'_{it})' \in \mathbb{R}^{p+1+q}$ is strictly stationary and ergodic;

- ASSUMPTION 2.2 The regressors x_{it} are covariance stationary, independent of the errors, u_{is} for all t and s and X_i are independent across i;

- ASSUMPTION 2.3 The instruments z_{it} with absolutely summable autocovariances satisfies $E(z_{it}u_{it}) = 0$ and $E(u_{it}^2 z'_{it} z_{it})$ is finite. $E(z_{it}x'_{it})$ has full column rank;

- ASSUMPTION 2.4 The regressors x_{it} have finite fourth moments. $E(u_{it}) = 0$, $E(u_{it}u_{js}) = 0$, for all $i \neq j, t \neq s$, $E(u_i u'_i) = \Sigma_i$.

- ASSUMPTION 2.5 The slope individual-specific coefficients, θ_i are restricted to be equal to the panel slope coefficient, θ.

REMARK 2.1 The above assumptions are relatively mild and the main restrictions addressed are independence across cross-sectional units and the exogeneity which are mainly needed for underlying GMM estimation.

With those assumptions, the following theorem can be obtained.

THEOREM 2.1 Under ASSUMPTION 2.1-2.5, for fixed T and $N \to \infty$, the J test statistic, as defined in equation (5), $\mathbf{J} \to \chi^2_{N(q-p)}$. Further more, as $T \to \infty$ and

$N \to \infty$ sequentially, $T\mathbf{J} \to \chi^2_{N(q-p)}$.

PROOF The GMM minimand is

$$Q(\theta) = \sum_{i=1}^{N} N^{-1} u_i'(\theta) Z_i W_i N^{-1} Z_i' u_i(\theta)$$

Plug $u_i(\theta) = y_i - x_i'\theta$ into it, we can obtain

$$Q(\theta) = N^{-2} \sum_{i=1}^{N} \{y_i' Z_i W_i Z_i' y_i + \theta' X_i' Z_i W_i Z_i' X_i \theta - 2 y_i' Z_i W_i Z_i' X_i \theta\}$$

The customary first-order condition for minimizing is

$$\frac{\partial Q(\theta)}{\partial \theta} = 0$$

for $\hat{\theta}$, it yields the equation

$$\sum_{i=1}^{N} \left(N^{-1} X_i' Z_i \right) W_i \left(N^{-1} Z_i' u_i(\hat{\theta}) \right) = 0$$

with the asymptotic counterparts as

$$\sum_{i=1}^{N} \left(E(x_{it} z_{it}') \right) W_i \left(E(z_{it} u_{it}) \right) = 0$$

Define $F_i = W_i^{1/2} E(z_{it}' x_{it})$ with $W_i^{1/2}$ a "root" of W_i such that $W_i = W_i^{1/2'} W_i^{1/2}$ and to ensure identification, $rank\{F_i\} = p$, then re-write the population moment condition as

$$F_i' W_i^{1/2} E(z_{it} u_{it}) = 0,$$

multiplying these p equations from the left by the projection matrix yields

$$F_i(F_i'F_i)^{-1}F_i'W_i^{1/2}E(z_{it}u_{it}) = 0$$

These are q equations but only p are linearly independent, as $rank\{F_i(F_i'F_i)^{-1}F_i'\} = rank\{F_i\} = p$, then it follows immediately that

$$\left(I_q - F_i(F_i'F_i)^{-1}F_i'\right) W_i^{1/2}E(z_{it}u_{it}) = 0 \tag{1.6}$$

which constitutes a set of $(q-p)$ linearly independent equations. And what's more, according to ASSUMPTION 2.5, $Y_i = X_i'\theta + u_i$, we can obtain

$$\begin{aligned}\hat{\theta} &= \left(\sum_{i=1}^{N} X_i'Z_iW_iZ_i'X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i'Z_iW_iZ_i'Y_i\right) \\ &= \theta + \left(\sum_{i=1}^{N} X_i'Z_iW_iZ_i'X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i'Z_iW_iZ_i'u_i\right)\end{aligned}$$

So the error term can be represented

$$\begin{aligned}u(\hat{\theta}) &= u + \sum_{i=1}^{N} X_i'\theta - \sum_{i=1}^{N} X_i'\hat{\theta} \\ &= u - \sum_{i=1}^{N} X_i(F_i'F_i)^{-1}F_i'W_i^{1/2}N^{-1}Z_iu_i\end{aligned}$$

Remember the J test statistic here is $\mathbf{J} = N\sum_{i=1}^{N} u_i'(\hat{\theta})Z_i'W_iZ_i'u_i(\hat{\theta})$ and let $\mathbf{J_i} = Nu_i'(\hat{\theta})Z_i'W_iZ_i'u_i(\hat{\theta})$, so

$$\begin{aligned}\mathbf{J}_i^{1/2} &= N^{1/2}W_i^{1/2}Z_i'u_i(\hat{\theta}) \\ &= N^{1/2}W_i^{1/2}Z_i'u_i - N^{1/2}W_i^{1/2}Z_i'X_i(F'F)^{-1}F'W_i^{1/2}Z_i'u_i \\ &= N^{1/2}(I_q - F_i(F_i'F_i)^{-1}F_i')W_i^{1/2}Z_i'u_i\end{aligned}$$

Following CLT together with equation (6), for fixed T when $N \to \infty$,

$$\mathbf{J}_i^{1/2} \to N(0, S);$$

where S is a certain semipositive definite matrix. And thus $\mathbf{J}_i \to \chi^2_{q-p}$. Together with Assumption 2.2, $\mathbf{J} \to \chi^2_{N(q-p)}$. ■

1.3 Bootstrap Procedure and Asymptotic Theory

1.3.1 Bootstrap Procedure

For a panel dataset there are in general three resampling schemes, temporal resampling, cross-sectional resampling and a combined resampling. The method used in the paper is cross-sectional bootstrap as suggested by Kapetanios (2008). As is usual in bootstrap literature, we will also use the superscript star notation in all resampled data.

Definition 3.1 For a $T \times N$ matrix of random variables $\mathbf{B}$, cross-sectional resampling is defined as the operation of constructing a $T \times N^*$ matrix $\mathbf{B}^*$ where the columns of $\mathbf{B}^*$ are a random resample with replacement of blocks of the columns of $\mathbf{B}$ and N^* is not necessarily equal to N.

And dealing with the non-parametric bootstrap, for the panel data model defined in equation (1), cross-sectional resampling will draw columns of $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{Z}$ with replacement, that is, $\mathbf{Y}^* = (y_{i_1}, ..., y_{i_i}, ..., y_{i_N})$ with the elements of vectors of indices $(i_1, ..., i_i, ..., i_N)'$ obtained drawing replacement from $(1, ..., N)'$. And the same scheme

is used to obtain $\mathbf{X}^*$ and $\mathbf{Z}^*$. We use the following cross sectional bootstrap procedures:

1. Let $N_1, N_2, ..., N_B$ be iid uniform random variables on {0, 1 ,..., N} and let $(x_b^{*'}, y_b^*, z_b^{*'})'$ of sample size of $N_b \times T$, $b = 1, ..., B$ be the resampling from $\{x_i, y_i, z_i\} \mid_{i=1}^{N}$ with replacement.

2. Calculate 1-step bootstrap GMM estimator $\tilde{\theta}^*$ by minimizing bootstrapped (2) and calculate $\hat{S}_{HAC}^*$ based on the residuals;

3. Obtain 2-step bootstrap GMM estimator $\hat{\theta}^*$ by minimizing bootstrapped (3) using $\hat{S}_{HAC}^*$;

4. Obtain the bootstrapped J test statistic as

$$\mathbf{J}^* = N \sum_{i=1}^{N} u_i^{*'}(\hat{\theta}) Z_i^{*'} \hat{S}_{i,HAC}^{*-1} Z_i^{*'} u_i^*(\hat{\theta}).$$

By repeating 1-4 B times, where B should be sufficiently large, one can approximate the finite sample distributions of the J test statistic by the empirical distributions of their bootstrap version.

1.3.2 Asymptotic Theory

The bootstrap improvement can be explained in terms of Edgeworth expansions, just as done by Singh (1981), Hall (1986) and Beran (1988).

Inoue & Shintani (2006) have shown that the empirical distribution of bootstrapped J test statistic based on HAC covariance matrix estimators can be approximated by their Edgeworth expansions. As the asymptotic properties of bootstrapping is independent of the size of bootstrapping blocks, we can rely on the proof that they

made in the paper. Define k as the characteristic exponent of the kernel function ω, i.e. the largest number such that $lim_{x\to\infty}\frac{1-\omega(x)}{|x|^k} \in [0,\infty)$. After making regularity assumptions of the moment conditions, the lag of truncation l, the kernel functions and the characteristic function of (z_{it}', u_{it}) satisfying Cramer's condition, they obtained the following results.

1. The approximation errors made by the first-order asymptotic theory for the J test statistics are of order $O(lN^{-1}) + O(l^{-k})$;

2. The bootstrapped J test statistics can be approximated by Edgeworth expansions and the errors are of order $o_p(lN^{-1}) + O_p(l^{-k})$;

3. The approximation error $O_p(l^{-k})$ is due to the bias of HAC estimation and bootstrap can provide asymptotic refinement only if $O(lN^{-1}) = o(l^{-q})$, together with the assumption that $l = O_p(T^{1/3})$, proposed by White and Domowitz (1984) to solve the problem of consistency, it implies that $k \geq 2$ [1], i.e. in the estimation of HAC covariance matrix, the characteristic exponent of the kernel function should be larger than 2, for example, the truncated kernel, the trapezoidal kernel and the Parzen (b) kernel;

4. Once $k \geq 2$ is satisfied, the approximation error of bootstrap, $o_p(lN^{-1})$ is smaller than that of the first-order asymptotic, $O(lN^{-1})$;

5. The non-positive semidefiniteness of HAC covariance matrix is still a very troubling issue to solve.

Based on the results above, the following theorem can be obtained,

THEOREM 3.1 For fixed T, when $N \to \infty$, the cross sectional bootstrap estimation of the distribution of $\mathbf{J}$ is $o_p(lN^{-1}) + O_p(l^{-k})$. What's more, with kernel functions

[1] The proof can be found in the appendix.

with characteristic exponent larger than 2, the cross-sectional bootstrap can provide asymptotic refinement. And with $N \to \infty$ and $T \to \infty$ sequentially, the bootstrapped errors become $o_p(lN^{-1}T^{-1})$.

PROOF Just as the second part in the proof of THEOREM 2.1 $T\mathbf{J}^* \to \chi^2_{(q-p)}$. Then the results just follows as the case with fixed T.

1.4 Monte Carlo Simulations

In this section, I conduct experiments to examine the small sample properties of the cross-sectional bootstrap J test. An individual fixed effect AR(1)dynamic panel data generating process(DGP) is defined as

$$\begin{aligned} y_{it} &= \gamma y_{i,t-1} + \beta x_{it} + \alpha_i + u_{it} \\ x_{it} &= \rho x_{i,t-1} + v_{it}, \qquad i = 1, ..., N; t = 1, ..., T \end{aligned}$$

where the error terms $u_{it} \sim NID(0,1)$ and $v_{it} \sim NID(0,1)$, α_i is the individual fix effect which introduces the individual dependence and later in the second experiment a similar time fix effect will be also considered to check the performance of my bootstrap with cross-sectional dependence. The sample size is quite small with $N = 20$ and $T = 10$. The instruments z_{it} are chosen as $(y_{i,t-1}, y_{i,t-2})$, thus $q - p = 1$. In order to get a consistent estimator of HAC covariance matrix the kernel function and the bandwidth b_t should be chosen carefully. In my first experiment, I will compare the performance of my bootstrap with different kernels. As for the bandwidth, people argued it is a much more important determinant of the finite sample properties of S_{HAC}. Andrews (1991) shows that the asymptotic mean square error is minimized by setting $b_T = O(T_{1/3})$ for the Bartlett weights and $b_T = O(T_{1/5})$ for both the Parzen and Quadratic Spectral weights, which unfortunately provides little practical guidance.

Table 1.1: Empirical size of bootstrap J test with different kernels

Nominal level	Bootstrap test
	Bartlett Kernel
0.1	0.092
0.05	0.048
0.001	0.0018 *
	Truncated Kernel
0.1	0.08*
0.05	0.052
0.001	0.002*
	Parzen Kernel
0.1	0.096
0.05	0.044*
0.001	0.008*

Newey and West (1994) propose a nonparametric method for selecting the bandwidth and show it minimizes the asymptotic mean square error criterion, which is now widely used in different packages and I will also select the bandwidth via this procedure.

Table 1 describes the results of the cross-sectional bootstrap using HAC covariance matrix estimated by different kernels, Bartlett kernel, truncated kernel and Parzen kernel. There are 200 bootstraps and 500 Monte Carlo, the auto-regression coefficient of y_{it}, γ is set to be 0.1, $\beta = 0.9$, $\rho = 0.1$. The results show that Bartlett kernel outperforms the other two, which are also not extremely bad, just as Newey and West (1994) conclude that the choice between the kernels is not particularly important estimation of S_{HAC}.

Secondly, I conduct different experiments to examine the sensitivity of my bootstrap procedure against different parameter settings and compare the performance to Arellano and Bond (A&B) GMM estimation, the results are illustrated in Table 2. Generally my bootstrap is superior to that with A&B estimation where all the asymptotic levels are significantly different from the nominal level at 10%, just as people detected

Table 1.2: Size properties sensitivity analysis

Nominal level	HAC GMM Bootstrap test	GMM bootstrap test
	200 Bootstraps, 500 Monte Carlo	
0.1	0.092	0.16*
0.05	0.048	0.07*
0.001	0.0018*	0.026*
	$\gamma = 0.9$	
0.1	0.064*	0.034*
0.05	0.042*	0.073*
0.001	0.004*	0.006*
	Twoway fix-effect DGP	
0.1	0.102	0.17*
0.05	0.048	0.07*
0.001	0.01	0.016*
	400 Bootstraps, 800 Monte Carlo	
0.1	0.0925	0.14*
0.05	0.0475	0.069*
0.001	0.0075*	0.005*

before, the J test statistic with A&B estimation is seriously distorted, according to my results, quite oversized. As for the sensitivity analysis,firstly I increase the auto-regression coefficient γ from 0.1 to 0.9, and the results become worse, exactly speaking very oversized at significant level 10% and 1% and undersized at 5%. Then I add a time fix effect in the DGP to create cross sectional dependence. To my surprise, the results are not so bad as I imagine and even better than that of the one way effect DGP. I would the reason for that may be that the way of introducing the cross sectional dependence is so weak that it is not dominated or may easily diminish during the process of estimation. At last, I increase the number of bootstraps and Monte Carlo to 400 and 800 to check whether the results are improved, while the improvement can be seen very clearly.

1.5 Conclusion

In this paper, I suggest in cross-sectional independent dynamic panel data model, the overidentifying test, J test statistic as the GMM minimand multiplied by the number of cross section units, which is shown to have a χ^2 asymptotic distribution with the degree of freedom of the number of overidentifying restrictions for fixed T. As usually the dependence structure of the moment conditions is unknown, the HAC covariance matrix is chosen as the weighting matrix in GMM estimation. Inoue and Shintani(2006) has shown the bootstrap with HAC GMM estimator can provide asymptotic refinement but only for kernel functions with characteristic exponent larger than 2. I also propose a cross sectional bootstrap scheme based on HAC covariance matrix estimators and by Monte Carlo simulations, the proposed test statistic is shown to have nice size properties and the Bartlett kernel function seems to perform best. It is suggested that this scheme of J test can be used in dynamic panel data where there is weak or even no cross-sectional dependence and relatively small auto regression coefficients.

Chapter 2

Utility Payments *In*discipline and Its Determinants in Russia

2.1 Introduction

The issue of poor payment discipline for utilities in Russia and some former Soviet-Union countries has been discussed particularly in Russian media. The problem appeared after the collapse of the Soviet Union, before which the utilities sector had been heavily subsidized and households had to pay relatively low fees. After support from the government to the sector faded and the infrastructure has become shabby, fees have started to increase but the industry particularly in Russia came across the problem of low payment discipline, which allows neither to invest in infrastructure renewal nor to keep the sector viable. Currently, Russian government is trying to attract private investments to the sector but this has been difficult to implement due to low payment discipline and subsequent lower than necessary payment collection rates.

In the current study we use detailed panel data on Russian households to analyze

incidence of non-payments for household utilities and try to find out if households with payment arrears have special systematic characteristics. As far as the issue of type of households with payments arrears in Russia, it has not been analyzed in detail so far. Additionally, benefits and discounts on housing payments and on the contrary punishment policies can be better targeted if features of households having relatively systematic difficulties with timely payment are identified and better understood.

The paper proceeds as follows. First, we will briefly review the literature that addressed various issues related to the utilities sector in transition countries. Then, we provide description of the data used in the study and discuss characteristics of the sample. After econometric specification is defined, we discuss our findings and conclude the study.

2.2 Utility Payments in Russia and CIS

As it is explained in detail in the World Bank report on infrastructure in transition countries[1] the problem of poor payment discipline for utility services appeared in most of transition economies after the practice of subsidizing operations of state-run utility service providers was stopped following the collapse of Soviet Union. Since the problem of poor payment discipline emerged, governments of newly independent states have resorted to various measures to increase collection rates and turn provision of utility services into a self-sustaining industry. It is hard to state with confidence if this has been achieved so far, since according to recent research on affordability in Russian and transition countries, poor payment discipline still appears to be an issue to various extent in these countries (Fankhauser and Tepic (2005)).

The issue of poor payment discipline for utilities, affordability of services and other

[1] *Infrastructure in Europe and Central Asia Region – Approaches to Sustainable Services.* World Bank, June 2006.

related issues in Russia and former communist countries in general have been discussed by several researchers. For instance, Fankhauser and Tepic (2005) analyze current and future affordability for utility services in a number of transition countries including Russia. They conclude that households in transition economies on average are able to pay utility bills without much difficulty. Affordability is still a problem in certain countries and particularly for low-income groups, though they admit that this is a complex issue and straightforward conclusions about affordability are hard to make. Both Fankhauser et al. (2008) and Bashmakov (2004) suggest that affordability may get worse in coming future, since attempts are being made to transform utility services sector into a financially viable one and large scale investments are necessary for refurbishment and improvement of current poor infrastructure.

The issue of collection rates, when private investments are interested in the sector is certainly of foremost importance. According to Bashmakov (2005) collection rates should be kept at the level of 95%, so that costs are covered, but average collection rates over the whole Russia have been around 85-90% between 1999 and 2004, which is apparently insufficient. Bashmakov (2004) determines threshold level of growth rates for utility payments in Russia, after passing which collection rates start to fall. This question is mainly important for private sector businesses entering the sector, which will be expecting return to their investments in infrastructure improvement. He determines critical growth rates for utility payments, which start affecting collection rates and claims that they are stable across the population and for various countries.

Hamilton et al. (2008) analyze relevant issue of household subsidies in Russia using a nationally representative survey, NOBUS, conducted in 2003. They find that currently housing allowances are not targeted efficiently. They also establish that availability of housing subsidies vary across regions. Additionally, urban households were found to be more likely to receive housing subsidies than rural households.

A similar study analyzing housing subsidies and utility payments but in neighbor-

ing to Russia Ukraine, which also suffers from low payment discipline, was done by Fankhauser et al. (2008). They find that affordability of utility services improves with income and size of a household. As in case of Hamilton et al. (2008) for Russia, they also find that subsidy targeting used currently in Ukraine does not appear efficient. In Ukraine, interestingly, industrialized regions in Central and Eastern part of the country including well-off capital city Kiev, in addition to poor regions of the country, have high incidence of arrears in utility payments. Additionally, households in urban areas, families with older members and households, where the first respondent is a woman, are more likely to have problems with timely payment for utilities.

Lampieti et al. (2001) study the effect of increase in electricity tariffs on payments by poor and non-poor households in late 1990s in another former Soviet Union country, Armenia. After tariff rise, electricity consumption fell both among poor and non-poor households, while the average monthly bill increased. Consequently, collection rates significantly fell due to both increase in the size of arrears and rise in the number of households not paying bills on time. After tariff rise, number of households classified as "non-poor" but at the same time carrying arrears increased from 22 to 37 percent of the total number of non-poor households, while number of poor households carrying arrears increased from 27 to 46 percent of all poor households. This indicates that households in Armenia are quite sensitive to rise in tariffs, which on the other hand might be crucial for further stable functioning of the utility services sector.

Our contribution in current work is to provide more understanding to what type of families are actually prone to have problems with timely payment for utilities in Russia by analyzing data on non-payments and various household characteristics. Considering repeated findings that housing subsidy programs in transition countries are not efficient, our findings may provide additional help to better tailoring future adjustments to assistance programs. Moreover, results may suggest what type of household categories are under pressure, as far as punishment policies for systematic non-payments

are concerned.

2.3 Data and Sample Characteristics

We use data from three rounds, Rounds 9, 10 and 11, of Russian Longitudinal Monitoring Survey (RLMS) for years 2001, 2002 and 2003 respectively. The survey collects a vast array of data at both household and individual level on income, expenditures, health, employment, consumption, living conditions, land and time use, qualitative variables on the head of a household and data on wage arrears. Table 2.3 provides the list of variables used in the analysis. The data is mostly household based but additionally individual characteristics of persons identified as a head of a household are used. Individual characteristics used in the analysis are labor market characteristics and his or her marital status.

Only households surveyed during round 9 with complete information were included, thus new families added to the sample during rounds 10 and 11 are not in the analysis. Originally round 9 of RLMS included data on 4006 households but the final sample used in the analysis consists of a panel on 2335 households for three years totaling in 6756 observations. Except the data on access to various type of household utilities and wage arrears, variables overall do not contain missing values.

On average about 23% of households (average over rounds) in the sample indicated that they had unpaid household bills. Interestingly, share of households, which indicated that they had unpaid housing bills is almost the same both in urban and rural settlements, though access rates to utilities certainly differ. Table 2.1 shows percentage of households, which reported existence of unpaid housing bills in rural and urban areas as well as poor and non-poor households[2]. Years correspond to survey rounds

[2]Country level poverty threshold was used for identifying poor households.

Table 2.1: Existence of unpaid bills in the sample households

Settlement type	Poverty status	2001	2002	2003
Urban	Poor	42.2	40.1	34.7
	Non-poor	21.2	22.6	20.1
Rural	Poor	36.4	34.6	28.4
	Non-poor	22.8	20.6	18.8

and we can see that non-payment rate has been falling among both urban and rural households, though existence of unpaid bills still seems substantial, especially among poor households in urban areas.

In the sample most households (65%) live in urban areas, while the rest of the households reside in rural areas and so-called city-type settlements. In the analysis city-type settlements were classified as rural settlements. Observations are also classified based on eight geographical regions with distinct dummy variable for major metropolian areas, Moscow and Saint Petersburg[3].

About 60% of households are indicated as male-headed, while about 11% are female-headed just as shown in Table 2.3. As for employment status of household heads, around 8% of family heads reported themselves as unemployed. Additionally, over the survey period around 29% of household heads indicated that at the time of an interview, their place of work owed him or her money, thus wage arrears incidence was quite significant. Marital status of a household head has also been included. The latter might be considered as an indicator of the family structure of a given household. Over 60% of household heads are indicated as married, over 7% as "living together, but not registered", while the rest being divorced (about 11%), widowed (over 15%)and never married (about 10%).

[3]The other regions are Northern and North Western, Central and Cenral Black-Earth, Volga-Vyatski and Volga Basin, North Caucasian, Ural, Western Siberian and Eastern Siberian and Far Eastern.

Table 2.2: Access to utilities in urban and rural areas (in percent)

	Urban areas	Rural areas	Overall
Central heating	87	29	67
Central water	91	52	77
Hot water	72	17	53
Central sewerage	86	25	65
Metric gas	88	47	74

Data on poverty status of a household is also provided. Poverty status is shown by a dummy variable, which equals one if a total household income is less than a threshold level and zero if it is not. During the analysis we refer to a poverty threshold level determined for the whole Russia rather than threshold levels used for a particular region of the country. Threshold level along with corresponding dummy variable are provided in the original data set. Based on a country level poverty threshold, we see that on average over the survey period about 28% of households may be qualified as "poor", though if separate rounds are considered this percentage has declined from 36.4% in 2001 (round 9) to 21.5% in 2003 (round 11).

As for access to utilities Table 2.2 presents access rates to various utilities in our sample overall and separately for urban and rural households. As expected, access rates to utilities differ substantially among urban and rural households.

2.4 Econometric Specification

Our dependent variable is a dummy variable indicating whether a household at a time of the survey had unpaid housing and utility bills and, thus, we use limited dependent variable models to estimate such an equation.

First, we have performed Hausman-type test to test for fixed individual effects, which is based on the difference between conditional fixed-effects logistic regression formulated by Chamberlain (1985) and the usual logistic estimation, where Chamberlain's estimator is consistent whether the null hypothesis that the fixed individual effects are not present is true or not, while usual logistic estimation is consistent and efficient under the null hypothesis (see (Baltagi, 2005)). Though, the test has been performed not on the whole sample and complete set of variables due to the following reason.

A particular feature of Chamberlain's estimator is that it uses only those groups of data, or households in our case, for which the value of dependent variable has changed from zero to one or vice versa. This resulted in that 1575 groups or households, which accounted for 4511 observations have been dropped from the estimation. Total number of observations used by Chamberlain's estimator was 2245, which is substantially less than our initial sample. Obtained value of Hausman's χ^2 statistic equals 18.43, which means that we cannot reject the null hypothesis that the fixed individual effects are not present.[4] Moreover, geographical variables, which we are interested in, were also dropped by the estimator due to no within-group variation. Thus, random-effect models were preferable. We implement random-effects probit estimation methods to verify the results. And it is worth mentioning that in a probit estimation results might be dependent on the number of quadrature points. We performed quadrature check procedure and established that increase or decrease in the number of quadrature points in random effects probit estimation does not affect results significantly. Thus, probit estimation has been performed with 12 quadrature points.

Additionally, we can not ignore the dependence between past non-payment history and the probability of non-payment at present, then a dynamic model that takes into account the household's past experience is more appropriate. However the estimation of dynamic binary choice panel data model is far more complicated than the static

[4]Results can be provided upon request.

one. In probit or logit models, all explanatory variables should be strictly exogenous, otherwise the parameter estimates, especially for short panels (i.e. with small temporal dimension), jointly estimated with the individual effects can be seriously biased. Arellano and Carrasco (2003) propose a GMM estimator which is shown to be consistent and asymptotically normal for fixed T and large N. They present a class of binary choice panel data models with the following features: The explanatory variables are predetermined but not strictly exogenous, including lagged dependent variables; individual effects are allowed to be correlated with the explanatory variables. By means of Monto Carlo simulation in a binary choice model with a single lagged dependent variable, they compare the performance of the two-step GMM estimator with other estimators, such as maximum likelihood (ML) estimator etc. and conclude that the GMM estimator performs as well as the ML estimator, but the biases of GMM estimation increase with T for fixed N and tend to be higher for data with higher persistence. In our case, there are total 3 periods of survey, thus in a dynamic model, $T = 2$ and N is substantially large. So the GMM estimator will be more efficient here.

We estimated two specifications of the dynamic models: with and without the variable on wage arrears. The reason for estimating twice is that the wage arrears happen only for those who were employed and thus taking it into account will drop out those households whose heads are unemployed in the sample. Then in the specification with wage arrears, the employment status of the household head is dropped for collinearity, however the use of this variable reduces included observations to 2717 (from initial 4506). Nevertheless, we included this variable in a separate estimation, since wage arrears had been a serious issue in transition countries. Though, due to reduced sample size, results of estimation with wage arrears should be considered with additional caution. Thus, estimation has been done with and without this variable to see if it has effect on utility payment arrears.

2.4.1 Explanatory variables

As mentioned above Table 2.3 provides the list of explanatory variables used in the analysis with their corresponding mean and standard deviation. Two types of variables have been used: those characterizing certain features of a household head and those providing information on a household itself. We use information on employment status of a household head, marriage status, whether a household head is a pensioner and whether an employee of a household head owes him or her money. We also include a dummy variable, which equals one if a family is female-headed.

As for household-level variables, we use size of a household, number of children less than six years old and of school-age, real housing benefits in 1992 rubles, and a poverty indicator, which equals one if a household is considered to be 'poor' by an income threshold defined for the whole Russia. We also include variables indicating whether a house or an apartment a household lives in is its private, or as it is sometimes put "privatized", property. Dummy variables defining access by a household to various utilities, such as central heating, central water, hot water, central sewerage and metric gas, are also included.

Moreover, there is one more crucial household variable directly affecting the collection rates of utility, the affordability, which is a problem for low-income consumers in most countries, in particular in the CIS countries. Fankhauser and Tepic (2005) find that there is a positive relation between the affordability and the probability of payment. Simply speaking, affordability is defined as the share of monthly household income that is spent on utility services, such as electricity, district heating and water. In order to assess whether it is problematic, many governments and international financial institutions have developed *ad hoc* rules on what constitutes an acceptable level of utility expenditures. That is, affordability becomes problematic if the utility bills account for more than 25 percent of total outgoings on average over a year. We calculate the

Table 2.3: Variables used in the analysis

Variable	Details	Mean	St.dev.
HH† non-payment(*hunpaid*)	(1=utility unpaid)	0.23	0.42
HH head unemployed (*unem-s*)	(1=unemployed)	0.08	0.28
HH head married (*marstat2*)	(1=married)	0.60	0.49
HH head pensioner (*pensio*)	(1=pensioner)	0.34	0.47
HH head is owed money at work (*pjowed*)	(1=owed money currently)	0.29	0.45
Female headed (*fh*)	(1= HH female headed)	0.11	0.31
HH size (*nfmem*)	-	2.81	1.43
HH affordability(*aff*)	-	0.09	0.35
Number of children <6 years (*ncat1*)	-	0.19	0.47
Number of children 6-18 years (*ncat2*)	-	0.48	0.74
Savings (*savingr*)	in 1992 rubles	359.55	2051.77
Lendings (*lendr*)	in 1992 rubles	316.46	7080.55
Real housing benefits (*aptbenr*)	in 1992 rubles	46.54	125.92
Poverty indicator (*a-pind*)	(1=below threshold)	0.28	0.45
Private property (*hhpp*)	(1=owning private)	0.72	0.45
Central heating (*hcheat*)	(1=available)	0.67	0.47
Central water (*hcwatr*)	(1=available)	0.77	0.42
Hot water (*hhwatr*)	(1=available)	0.53	0.49
Central sewerage (*hcsewr*)	(1=available)	0.65	0.48
Metric gas (*hmgas*)	(1=available)	0.75	0.43
Settlement type (*sett-typ*)	(1=urban)	0.65	0.47
Moscow and St. Petersburg (*r1*)	-	0.04	0.20
Northern and North Western (*r2*)	-	0.07	0.26
Central and Central Black-Earth (*r3*)	-	0.21	0.40
Volga-Vyatski and Volga Basin (*r4*)	-	0.21	0.40
North Caucasian (*r5*)	-	0.13	0.33
Ural (*r6*)	-	0.16	0.36
Western Siberian (*r7*)	-	0.06	0.25
Eastern Siberian and Far Eastern (*r8*)	-	0.08	0.27

Notes:†household

affordability as the proportion of utility expenditure in the total expenditure of the household and the mean of it is 9%, which is lower than the bechmark, but they are scattered distributed with large standard deviation. So in this sample, the affordability may be problematic for some families and it is supposed to be negatively correlated with the probability of payment arrears.

As for geographical identifiers, dummy variable indicating whether a household lives in an urban or rural area has been included. The survey provides information on three types of settlements – urban, rural and so-called settlements of urban type. In our analysis, we have identified the latter as rural households as well. Additionally, regional identifiers have also been included indicating separately largest metropolitan areas as Moscow and Saint Petersburg and seven other regions of Russia.

Additionally, we have included such expenditure categories as household savings and lendings to control for their effect on the payment behavior.

People may ask whether those dummy variables especially those for the financial status of the household could be highly correlated, for example, usually households having lendings or owning private properties are much less likely to be poor. Moreover, Guzanova (1998) who analyzes privatization patterns in Russia, finds that pensioners are an active group of population, who privatized their housing along with well-off households. This finding raises concerns about possible dependence problem arising from certain variables in our analysis – dummy variable on retirement status of a household head, poverty status and a variable specifying ownership of a housing. Thus we also check the correlation between them, just as shown in Table 2.4. However none of them are highly correlated with others. The largest correlation is between the marriage statement of the household head and whether the household is female headed, which is -0.37. So according to this sample, the female heads are more prone to be unmarried or divorced.

2.4.2 Estimation Results

Table 2.5 provides estimation results. The first column provides estimates from random-effects probit regression. The second and third columns present results of

Table 2.4: Correlation of the variables

Variables										
HH† head unemployed	1.00									
HH head married	.0086	1.00								
HH head pensioner	-.21	-.22	1.00							
Female headed	.018	-.37	-.17	1.00						
HH affordability	.074	-.0127	-.018	.01	1.00					
Lendings	-.011	.0186	-.0214	-.0057	.0013	1.00				
Real housing benefits	-.042	-.048	.14	.047	-.011	-.01	1.00			
Poverty indicator	.25	.073	-.29	.06	.12	-.02	-.052	1.00		
Private property	.034	-.017	.14	-.071	-.016	.009	-.052	-.048	1.00	
Settlement type	-.12	-.037	.025	.08	.051	-.014	.14	-.15	-.23	1.00

Notes: †household

dynamic models. Column three presents results for the subsample of the employed, and thus with a variable on wage arrears, which was included once for reasons specified above, while columns one and two present results without this variable but for the full sample. We will focus on random-effect estimation methods, since Hausman χ^2 statistic indicated that fixed effects are not present and, additionally, fixed-effects estimator drops important geographical identifiers.

In those two AR(1) binary choice models, the first-order lagged variables are both highly positively significant, which demonstrates that those households who have experienced non-payment in the last year are much more likely not to pay for the utility this term. So we cannot ignore the influence of the past and the dynamic models are more credible. However we also find that most of the coefficients are consistent across these two types of estimations. According to the first two columns, the affordability is highly significant and, just as expected, negative. So the smaller the proportion of expenditure on utilities in the total outlay is, the smaller is the probability that the household won't pay. And unemployment of a household head is a significant determinant of existence of non-payments in a household for the full sample. Interestingly, dummy variable for pensioners appears highly significant and negative, indicating that

Table 2.5: Estimation results

Explanatory Variables	R.e.Probit	GMM(Full)	GMM(Empl.)
HH unpaid util. last year		0.35**	0.38**
		(0.016)	(0.019)
HH affordability	-0.80**	-0.043**	
	(0.26)	(0.016)	
HH head unemployed	0.57**	0.051*	
	(0.20)	(0.024)	
HH head pensioner	-1.08**	-0.046**	-0.064**
	(0.18)	(0.012)	(0.021)
Owed money at work			0.066**
			(0.018)
No. children <6 years	0.18*		
	(0.08)		
No. children 6-18 years	0.53**	0.032**	0.028**
	(0.12)	(0.009)	(0.011)
Lendings		9.48e-07**	6.94e-07**
		(2.08e-07)	(2.67e-07)
Real housing benefits	-0.002**	-0.0002**	
	(0.00)	(0.00)	
Poverty indicator	0.43**	0.054**	
	(0.15)	(0.016)	
Private property	-0.79**	-0.061**	
	(0.15)	(0.014)	
Central heating	0.64*		0.045*
	(0.27)		(0.022)
Settlement type			0.042*
			(0.021)
Northern and North Western	1.35*	0.051*	
	(0.53)	(0.023)	
Volga-Vyatski and Volga Basin	1.40**	0.037*	
	(0.49)	(0.029)	
Ural	1.34**	0.054**	
	(0.49)	(0.016)	
Western Siberian	1.67**	0.056*	
	(0.49)	(0.025)	
Constant	-2.47**	0.15**	0.064**
	(0.55)	(0.016)	(0.018)
Number of Observations	4483	4506	2717

Notes: Single and double "*" indicate 5% and 1% significance levels respectively. Standard errors are in parentheses.

pensioner-headed households have better payment discipline. This might seem unlikely at first sight, since pensioners in transition countries are typically classified as in low-income part of population. This finding suggests that retired people might be more responsible as far as payments for services are concerned. Additionally, as the sample shows pensioner-headed households are typically small, which may have an effect on their payment discipline. In the sample almost 45% of pensioner-headed households are single person, while almost 41% are two-person households.

Variables for a number of children appeared significant, especially the number of children of $6-18$ years old. This might suggest that in some families with school-age children expenditures for products and services other than utilities are higher (like food, clothing, school related expenses, pocket money for children and etc.), which leads to households having problem with timely payment of utilities. This could also be the reason why the dummy on little children, i.e. smaller than school age, is not significant in the dynamic model.

Lendings of households could not be neglected even with the rather little value, since lending appears significant. Housing and apartment benefits on the other hand are highly significant in all specifications and with expected negative sign, though the value of the coefficient is relatively low. Those small values are derived from the scaling of those covariates, as shown in Table 2.3. Poverty indicator is highly significant with positive sign in all specifications. This might suggest that to a big extent non-payments for utilities in Russia among other potential reasons is a problem of poverty rather than negligence or simple refusal to pay higher than previous fees as it is sometimes suggested.

Dummy variable indicating whether an apartment or house is a private property is highly significant in all specifications and has negative sign. This indicates that households living in a not privatized housing are more inclined to have payment arrears. The issue of ownership seems to be worth analyzing further but in current setting it

is difficult to identify how exactly this factor affects payment discipline. Interestingly, according to Guzanova (1998), there are few incentives for many Russian households to buy a private housing and a sense of "occupational rights" still plays a significant role in people's attitude towards housing in Russia. Apparently, this sense of "occupation rights" might have effect on attitude towards utility payments as well. It is important to mention that according to the sample, households living in a not self-owned housing have substantially higher than average access rates to all utility services with, as the estimation shows, relative higher inclination to have payment arrears. Therefore, these households may be one of key groups of interest to advance understanding of non-payments.

As for regional dummy variables four out of eight regions have appeared highly significant in the full sample estimation. All these four regions are underdeveloped compared to other metropolitan, such as Moscow and St. Petersburg. For example, Western Siberian region, according to Kolenikov and Shorrocks (2005), Siberia is a region with one of the highest incidence of poverty. And the three insignificant regions are the most developed areas in Russia, where just as expected, the collecting rates should be higher than others. However, there is one exception, north Caucasian region, which has gone through long-lasting military conflict, and thus might have contributed to low payment discipline in this region. However in our specification, it turns out to be insignificant.

As we clarified above, due to the potential significance of wage arrears, estimation has been performed separately with this variable as well for the subsample of the employed. The descriptive statistics of the employed is very similar with the full sample except with better income and financial status, such as affordability, savings and lendings etc.. Estimation results for this subsample are presented in the last column with some differences from coefficients for the full sample. Not surprisingly, the new variable indicating existence of wage arrears appeared highly significant with positive sign. This

suggests that wage arrears also have an impact on existence of utility payment arrears. As for the rest of the variables, the unpaid experience in the last year, dummies for whether a pensioner is the household head, whether having children between 6-18 years old and whether the household has lendings remain significant with the same sign and close value of coefficients as in the full sample estimation. However covariates about finance status of a household, such as the affordability, poverty indicator and whether they own private property and all the regional dummies turn to be insignificant. On the other hand, the access to central water and settlement type show as significant factors. Significance of variable on settlement type has appeared with positive sign. This might be an indicator of the fact that access rates to utilities are higher in urban settlements, thus non-payments are more frequent there. Access to central water is positively significant here, which is consistent with Fankhauser et al. (2008), who also find a negative correlation between the accumulation of arrears with connection rates to centralized gas and sewerage in Ukraine. As Fankhauser et al. (2008) explain the relationship between access to utilities and payment arrears, the positive relationship may be explained by difficulties in disconnecting households for non-payments. Variables for access to other public utilities, on the contrary, are insignificant.

Despite the fact that significance has not changed between estimates of static and dynamic models, value of coefficients differ with coefficients obtained from static estimation being approximately ten times as large as those from a dynamic model. In the specifications with wage arrears variable, variable on unemployment was omitted, since data on wage arrears is provided only to employed household heads. As far as signs and significance levels of variables are concerned, results appear uniform. Nevertheless, it would be interesting to look at estimation results not only for three rounds as in this study but for longer time periods as well.

2.5 Conclusions

Using panel data set on thousands of households for three consecutive years we analyzed existence of utility payment arrears in Russia. Results show that incidence of payment arrears is highly influenced by the past record of the non-payment and affordability of the utility of the households and also the employment and retirement status of the head, existence of wage arrears, number of children in a family and housing benefits. Additionally, households living in non-privatized housing appear to have higher probability of carrying payment arrears. Regionally, households in Northern and North Western, Volga-Vyatski and Volga Basin, Ural and Western Siberian regions appear to be more inclined to have arrears.

Overall, non-payment for utility services largely appears to be a problem of poverty. Nevertheless, one potential group for further analysis are households, who have not yet owned private housing and appear to be more inclined not to pay utility bills on time. This effect is highly significant and might be important for understanding the problem better considering the fact that these households have higher than average access rates to all types of utilities considered in this study.

Interestingly, considering the subsample of the employed only, the factors affecting the probability of non-payment change slightly. Especially the covariates about poverty are no longer significant, instead the dummy for wage arrears appears to be a good explanatory variable. Additionally, regional dummies do not provide further explanation of utility arrears in this subsample.

Due to limited access to the data, only three rounds of the survey have been included in the analysis. Therefore, it would be interesting to find out if obtained results persist in longer time periods.

Chapter 3

Is the labor market in China segmented? Evidence from semi-parametric analysis

3.1 Introduction

The International Labor Organization reported (see ILO (2002)) that informal employment makes up 48% of non-agricultural employment in North Africa, 51% in Latin America, 65% in Asia, and 72% in sub-Saharan Africa. If agricultural employment is included, the percentage rises, moreover in some countries like India and many sub-Saharan African countries it exceeds 90%. Estimates for developed countries are around 15%. The dualism of labor markets especially in least developed countries (LDC) has been discovered and debated for long. According to the dualistic view, there are barriers to entry into formal sectors so that some workers are forced to accept informal jobs which are usually characterized by inferior earnings and working conditions. Many existing studies of labor market segmentation underline different

wage function specifications across formal and informal sectors and find that there are always wage premiums in the formal sector within developing countries, such as Mexico (see Magnac (1991); Gong and van Soest (2002); Roberts (1989) for Guadalajara), Malaysia (see Mazumdar (1981)) and Turkey (see Tansel (1999)).

Along with economic restructuring and urban economic reform, China has witnessed a surging unemployment at the end of last century: between 1995 and 2001, about 45% of all employees in the state-owned enterprises, totally 45 million, lost their jobs, which is so called "Xiagang". Informal employment in China's urban residents was formed in this process and the Chinese government also showed an increasing interest in expanding other forms of ownership of enterprises to address the great issue of unemployment. Along with the economic reform, the informal economy is growing as a more and more important growing engine of economic development. Additionally besides the mass redundancy by government and collective owned enterprises, heavy migration from rural areas also increases sharply the supply of informal employment. According to the definition and estimates of informal employment by Wu and Cai (2006), in 2002 urban informal employment reached between 40.3% and 45.2% in the total employment. By 2008, this proportion remained at 41.0%. The average growth rate of informal employment between 1994 and 2004 was about 12.5% annually (see Hu and Zhao (2006)).

Does such a large-scale informal employment mean that the labor market in China is seriously distorted along the formal-informal line? Does informal employment in the industrialized environment with low wages indicate that they are blocked in the inferior market and can not move? Either by the labor market segmentation theory or finding different wage functions across sectors, Li (2009), Xu (2008) and Jin (2006) claim that the labor market is segmented along formal and informal line. However, Heckman and Hotz (1986) pointed out that earning functions could be different for a variety of reasons, so people should not simply conclude that the labor market is

segmented because factors affecting the earnings are different. Hu and Yao (2011) set up a logit model using Chinese urban labor market data in 2001 and find that wages increase when people enter the informal sector from the formal one, then they claim that people are not forced to accept informal jobs. A similar study on Mexican data by Maloney (1999) also concludes that both earnings differentials and patterns of mobility indicate that much of the informal sector is not inferior but a desirable destination and that the distinct modalities of work are relatively well integrated. Chen and Hamori (2009), based on the China Health and Nutrition Survey (CHNS) questionnaire (2004 and 2006 pooling data), estimate the hourly wage equations for the formal and informal employment separately and decompose the wage differentials by the method of Blinder and Oaxaca. The results indicate that differences in the characteristics between formal and informal employment account for a much higher percentage of the hourly income differential than do discrimination for the job type.

From a methodological standpoint, in observational studies an important problem of causal inference is how to estimate treatment effects, as like in an experiment a group of units is exposed to a well-defined treatment, but unlike an experiment there are no systematic methods of experimental design to maintain a control group, just as stated by Dehejia and Wahba (2002). So the estimate of a causal effect obtained by comparing a treatment group with a nonexperimental comparison group could be biased because of problems such as self-selection or some systematic judgment by the researcher in selecting units to be assigned to the treatment. Propensity score matching (PSM), initiated by Rosenbaum and Rubin (1983), is an important subset of matching methods. It corrects for sample selection bias due to observable differences between the treatment and comparison groups, i.e. the observable differences of the worker and job characteristics across sectors. The basic idea of propensity score matching is to find in the group of nonparticipants those who are similar to the participants in all relevant pretreatment characteristics. It matches participants with certain nonparticipants who get similar propensity scores. The propensity score employs a predicted probability of

group membership and summarizes all the background information about treatment selection into a scalar. Therefore, the propensity score matching controls for the differences of individuals and makes the groups receiving treatment and non-treatment more comparable. Lechner (2000) investigates the question whether it really matters for microeconometric evaluation studies to take account of the fact that the programmes under consideration are heterogeneous and applies propensity score matching to the analysis of the active labor market policy in a Swiss region. Ham et al. (2005) use propensity score matching to measure the effect of migration on the wages of those who move. Brand and Halaby (2006) adopt this matching method to identify and estimate the average treatment effect of attending an elite college on educational and career achievement. Pratap and Quintin (2006) find that after controlling for selection by propensity score matching, no wage premium remains in the formal sector for Argentinean data. Badaoui et al. (2008) provide three estimators of the informal wage penalty in South Africa by ordinary least squares, differences-in-differences(DID) estimator and combined propensity score matching and DID estimator. They find that the informal wage penalty is only marginally higher than for the simple DID estimator but much lower than for the OLS estimator and by taking the tax avoidance into account, the penalty becomes insignificant.

This paper, using the data of the China Health and Nutrition Survey (CHNS), investigates whether the unregulated or informal labor market in China is integrated into the formal one. It is organized as follows: Section 2 first defines informality in this study and then describes the data; Section 3 discusses the propensity score matching methods combined with the difference in difference estimator and the robustness to its assumption and model specification and finally; Section 4 concludes.

3.2 Data

The China Health and Nutrition Survey (CHNS) was designed to see how the social and economic transformation of Chinese society is affecting the health and nutritional status of its population. The study population is drawn from 9 provinces including Guangxi, Guizhou, Heilongjiang, Henan, Hubei, Hunan, Jiangsu, Liaoning and Shandong. This sample is diverse, with variation found in a wide-ranging set of socioeconomic factors including income, employment, education and other related health, nutritional and demographic measures. It is pooling data for the 2004, 2006 and 2009 waves in this paper.

Due to data availability, formal employment in this paper is identified according to the type of contract, i.e. only employees with long-term contracts are considered as formally employed, which is in accordance with the specification by Chen and Hamori (2009). Individuals who *"work for another person or enterprise as long-term employees"* are considered as the formally employed and those who are *"self-employed"* or *"contract workers with other people or enterprise"* or *"temporary workers"* or *"paid family workers"* are considered to be informally employed. There are in all 2056 individuals identified as informal employed accounting for 46.9% of the total sample which is about 10 percentage points more than the population informal proportion reported in China Statistical Book (2005)[1]. The reasons may be that the agricultural population is excluded from our informal employment and the excess part of informal employment might be derived from our specification of informal employment employed here. We can not ignore that even some short-term employees or self-employed are entitled to the social protections and should be categorized as formal employment. Henley et al. (2006) also look into the degree of congruence between different practical proxies of informality using Brazilian household survey data for the period 1992 to 2001. They

[1]Definition of informal employment in this paper is not completely consistent with the one in the official annual statistical book.

find that in the self-employed group, there are roughly 99% individuals not having labor cards, and in the temporary employees, the percentage is about 90 %, which means that 10% of the temporary employees are incorrectly identified as informal workers according to our proxy of informal employment. Although the exact amount of misspecification depends on countries, their results indicate that our categorization of formal and informal employment overestimate the amount of informal employment in China.

ILO (2002) characterizes informal employment as follows: Under the expanded concept, informal employment is understood to include all remunerative work - both self-employment and wage employment - that is not recognized, regulated, or protected by existing legal or regulatory frameworks as well as non-remunerative work undertaken in an income-producing enterprise. As China has experienced the economic transition from a totally planned economy, all the employment was permanent before the economic reform. According to this specific conditions, Hu and Yang (2001) give their definition of the Chinese informal employment as: The informal sector in China mainly refers to small-scale business units outside of the legally established independent corporate units (enterprises, institutions, government agencies, community groups and social organizations). Informal employment in China mainly refers to the employment existing in the informal sector and sometimes those in the formal sector, including temporary employment, part-time employment, dispatching labor force, subcontracting workers etc. They also set up an empirical proxy of the definition of Chinese urban informal employment as the self-employed or family or temporary workers and claim that the large amount of rural migrants workers cannot be easily precisely counted up, so that they are not included in the informal employment.

The descriptive statistics of the variables used in the paper are stated in Table 3.1. People are surveyed for the income of last month, including wages, bonuses and subsidies (grocery subsidy, health allowance, bath and haircut allowance, book and newspaper

Table 3.1: Individual and job characteristics of formal and informal employment

	All		Formal		Informal		Diff	*"Contract*
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*t-test*	*workers"*
Hourly wage	5.28	8.83	5.22	9.21	5.35	8.23	-.08*	5.42
Age	37.55	13.55	38.23	12.06	36.44	14.61	1.79*	32.13
Experience	26.04	13.87	26.44	12.19	25.37	16.23	1.07*	20.77
Married	.64	.48	.72	.45	.54	.50	.12*	.57
East	.34	.47	.34	.47	.34	.47	.003*	.38
Hukou	.74	.44	.77	.42	.69	.46	.08	.66
Gender	.51	.50	.56	.49	.42	.49	.26*	.40
Health Cond.	.96	.23	.97	.18	.95	.18	.003*	.97
Education								
University	.11	.29	.14	.33	.08	.22	.12*	.13
Vocational	.19	.33	.28	.37	.07	.26	.11*	.07
High school	.29	.40	.27	.41	.37	.38	.09	.18
Middle school	.31	.45	.22	.44	.37	.46	-.05*	.31
Primary school	.10	.30	.09	.28	.11	.31	-.02	.08
Establishment size								
< 20	.17	.29	.09	.27	.39	.32	-.02*	.15
$20 - 100$	.32	.32	.21	.41	.38	.29	.12*	.26
> 100	.51	.45	.69	.48	.23	.38	.29*	.59
Type of work unit								
government	.52	.49	.70	.30	.30	.34	.70*	.45
collective	.18	.28	.06	.21	.37	.35	-.07*	.17
private	.29	.31	.25	.21	.34	.39	-.06*	.38
Type of occupation								
Technological	.24	.38	.42	.40	.07	.33	.07	.14
Management	.12	.27	.23	.30	.10	.21	.06*	.06
Craft	.55	.48	.38	.49	.74	.47	.07	.44
Service	.10	.25	.07	.25	.19	.25	.01*	.12
2004	.256		.276		.234			
2006	.296		.324		.264			
2009	.448		.400		.502			
Observations	4385		2329		2056			372

Notes: Age is measured in years and real hourly wage in RMB yuan. Other entries give the friction of employees in each category.

allowance, housing and other subsidies) and also the working hours. The hourly wage engaged in our analysis is calculated as income standardized by the total work hours

in that month. The informal employees can earn slightly more than the formal in our data. One reason might be that the self-employed are included in the informal employment which will enhance the average income of informal employment.

The sample average age is around 37 years, while formal employees are in general older than the informal employees and 72% formal employees are married, compared to 54% informal workers. Education levels here are categorized based on academic degrees with five dummy variables for *"professional school (3-year college) or higher"*, *"technical or vocational degree"*, *"upper middle school degree"*, *"lower middle school degree "* and *"primary school"*. Most of the individuals are not highly educated, especially in the informal sector, as their average age is about 40 years, for whom the nine-year compulsory education was not implemented in China. *Experience* is then calculated as age diminished by years of schooling, which are 15 for those with professional school (3-year college) or higher, 14 for technical or vocational degree, 12 for upper middle school degree, 9 for lower middle school degree and 6 for primary school. The formal employees are slightly more experienced than the informal. *East* represents whether people are settled in the eastern area, i.e. in Shandong or Jiangsu province, where economies are more developed. There are in total 34% of those surveyed settled in the east and the formal employees concentrate more in the eastern provinces than the informal. *Hukou* denotes whether the employee is registered as rural or urban resident. This also matters a lot in China, as in the countryside, people do not have access to some social securities, such as free education for the children and health insurance. Formal employment also outperforms informal one in this measure, although overall there are about 74% people from the rural areas. It is interesting to notice that more than half of the surveyed people work in large firms, i.e. firms with more than 100 employees. Moreover, about 70 % formal employees are in large enterperises in this survey. This might be caused by the methods of categorization, as in China we can hardly call a firm with 100 employees large. According to the conventional classification of firm scales in CSB (2010), enterprises with more than 2000 employees can

be considered as large and those with 300 workers or less are classified as small firms. According to the type of work unit and occupation, there are more informal workers in state owned (government owned or collective owned) enterprises and more technological and managing employees, while informal workers concentrate doing craft and service work and in private sectors.

The second last column in the Table 3.1 is the results of the test for the mean difference between the formal and informal groups. Most of the differences are significantly different from zero, which means the formal and informal employees are very distinct with each other according to the human and job characteristics. It is very interesting to find that the informal employees tend to work in smaller establishments and do manual labor as craftsmen or service workers and they are in general younger, less educated than the formal employees, but can earn more than the others. This can be explained as the following: paying personal income tax and the social insurance is compulsory for the formal employees, which reduces the income level of them. Moreover, the formal employees, especially those work in the large or state-owned enterprise can obtain much in-kind and other hidden income.

The classification of informal employment in this paper is consistent with that in Hu and Yang (2001), except for the *"contract workers with other people or enterprise"*. Contract workers, which are defined as short-term employees recruited by enterprises or institutions by signing an employment contract,[2] are different from either long-term or temporary workers, as they have to buy the social insurance by themselves but sometimes they can also get some allowance either from the company or the government. In this context, they should be classified as informal employment. And it would be interesting to look into details of the contract workers, the descriptive statistics of which are shown in the last column of Table 3.1. There are in all 372 out of 4385 surveyed people signed as contract workers, and they are earning slightly more than

[2]See *www.baike.baidu.com/view/136312.htm#1* for the definition of Chinese contract workers.

the average informal employees, and younger and higher educated, but in general, they are not the outliers among the informal employees.

To keep in accordance with other studies for segmentation of labor market by informality, agricultural households are excluded in our analysis. According to Chinese labor law, we only keep employees between 16 (school leaving age) and 59 (state retirement age) years old.

3.3 Combined Diff-in-diff Propensity Score Matching Estimator

Matching, as a strategy to control for covariates, has attracted increasing empirical interests in the past decade or two. One common feature of the structural regression estimators is to assume the linearity of the relationship to underlie the effects of treatment, which can lead to comparing very different treated- and control-group individuals. Thus one would ideally like to estimate the treatment effects of individuals with similar characteristics, in our case, to compare the wage differences for each groups of similar employees, in which there are some working in the formal sector while others are in the informal sector. One difficulty in ensuring that one can have such comparable groups is the reduction of the comparison dimensionality, as usually there is a set of characteristics that determine the selection into the formal sector and the wages, hence people face the problem of matching individuals on multiple dimensions. The section in the follows will present one possible solution to this problem by combining the differences-in-differences and propensity score matching estimator.

3.3.1 Propensity Score (PS) Matching

The propensity score, introduced by Rosenbaum and Rubin (1983), is defined as the conditional probability of being treated, i.e.

$$p(x) = Pr(z = 1|x),$$

where x is the observed covariates and z is the treatment assignment with 1 for being treated and 0 for without treatment. It is shown by them as the coarsest balancing score, $b(x)$, which is defined as a function of x, such that the conditional distribution of x given $b(x)$ is the same for treated $(z = 1)$ and controlled $(z = 0)$, meaning that given the covariates, propensity score matching can be used to balance the covariates in the two groups, and therefore reduce the selection bias.

In econometric literatures, the value of interest, and most commonly studied amount i.e. the *Average Treatment Effect for the Treated* (ATT), is then defined by Heckman et al. (1999) as:

$$\alpha_T = E(r|x, z = 1) - E(r|x, z = 0),$$

where r is the outcome. However, the latter item is counter-factual, as for any treated individual, people cannot observe their outcomes for being untreated. In order to be able to calculate the latter item, a key assumption is required for identification, which was first articulated by Rosenbaum and Rubin (1983) as "ignorable treatment assignment", or the "conditional independence assumption (CIA)" by Lechner (2000)

$$(r_1, r_o) \perp\!\!\!\perp z|x,$$

which requires that the treatment assignment z should be conditionally independent of the response (r_1, r_o) given x. Lechner (2002) calls this assumption a "data hungry" identification strategy and points out that all that is needed is to identify mean effects

is conditional mean independence. So he claims in most empirical studies it would be difficult to argue why conditional mean independence should hold and CIA might nevertheless be violated. However in the following empirical study of training effect on the wage levels, he just assumes the data base is fairly informative because it contains all the information, thus the independence assumption should be fulfilled. Brand and Halaby (2006) make a similar assumption that their observable exogenous covariates capture pre-treatment characteristics of units that determine selection into treatment and control status, into elite and non-elite colleges. They also indicate that all frameworks for causal inference make a similar assumption at some point.

There has been some controversy about the plausibility of the CIA in economic setting, arguing that agents' optimizing behavior precludes their choices being independent of the potential outcomes, whether or not conditional on covariates. However, Imbens (2004) provides three arguments against those queries. First, statistically the logical step of evaluation of any program is to adjust any difference in average outcomes for differences in exogenous background characteristics, so that the violation of such an assumption would seem difficult to rationalize in a serious attempt to understand the evidence regarding the effect of the treatment; secondly, that assumption merely asserts that all variables that need to be adjusted for are observed by the researcher, which is an empirical question and not one that should be controversial as a general principle; finally, even without invalidating that assumption, two agents with the same values for observed characteristics may differ in their treatment choices, as sometime the difference in their choices is driven by differences in unobserved characteristics that are unrelated to the outcomes of interest. In general, the validity of the conditional independence assumption is an empirical question, depending on the exact nature of the optimization process faced by the agents.

There is one way to release the strong conditional independence assumption, when people are only interested in the ATT. As there, only the independence of the controls

is needed. In this case, we only need to assume that

$$r_o \perp\!\!\!\perp z|x.$$

In this paper, denote the wage in formal and informal sector as ω^F and ω^I respectively, the treatment $z = 1$ for formal employment and $z = 0$ for informal one and X and Y as the characteristics of individual and employer. The CIA is then as follows

$$(\omega^F, \omega^I) \perp\!\!\!\perp z|(X, Y). \tag{3.1}$$

It is a strong assumption assuming that our data is so informative that the choice of be formally or informally employed only depends on the observable characteristics of the employees and firms, but not the formal and informal wages they might obtain. It also indicates that in this model, the offer for the job type is predetermined by the characteristics, so that bargaining is almost impossible. Imagine that before posting a job offer, the firm has already decided to hire a certain person either formally or informally. This may be restricted by the social regulation or their budget. Then the candidates are already aware of the type of job before they bargain for the possible wage.

The average effect of the treated, i.e. the wage premium of the formal employment in our study, is then defined as

$$\alpha_T = E(\omega^F|X, Y, z = 1) - E(\omega^I|X, Y, z = 1).$$

With the fulfillment of the CIA assumption, the formal wage premium becomes

$$\alpha_T = E(\omega^F|X, Y, z = 1) - E(\omega^I|X, Y, z = 0).$$

Rosenbaum and Rubin (1983) also establish that conditioning on the propensity score is equivalent to conditioning on the covariates themselves, which means that

$$\alpha_T = E(\omega^F | P(X,Y), z = 1) - E(\omega^I | P(X,Y), z = 0).$$

In our case, to estimate the after-matching wage premium α^M, write $i \in F$ if worker i is formally employed, and $j \in I$ if worker j is informally employed. $p_i = Pr(z = 1|X_i, Y_i)$ and $p_j = Pr(z = 0|X_j, Y_j)$ denote their propensity scores. Then we can compute α^M as

$$\alpha_T^M = \frac{1}{N} \sum_{i \in F} \left(\omega_i^F - \sum_{j \in I} \phi(|p_i - p_j|) \omega_j^I \right),$$

where N is the number of informal employees in the sample and the function $\phi(|p_i - p_j|) \in [0,1]$ accounts for the weight assigned to informal worker j matched for the formal worker i and it should be a decreasing function of the difference of p_i and p_j.

3.3.2 Differences-in-Differences(DID) PSM Matching Estimator

The assumption in Equation (3.1) might be violated is one concern of propensity matching estimation. This occurs when the selection into the formal sector may rely on the unobserved heterogeneity which affects the wages. A differences-in-differences matching strategy, as defined by Heckman et al. (1997), allows for temporally invariant differences between the treated and control groups, thus is less demanding of the data. However, it is also more demanding because it requires pre-programme data. Heckman et al. (1997) verify the validity of those two assumptions. The CIA assumption similar with our Equation (3.1) are decisively rejected in the data, while the weaker assumption that justifies the conditional DID method is not rejected for any group. Similarly,

Smith and Todd (2005) show that the CIA assumption in Equation (3.1) does not hold using a case study, instead, they combined a PSM estimator with a DID estimator and find that it performs the best in a nonexperimental context. Thus in this section we proceed along this line.

Let t represent a time period after the programme start date and t' a time period before the programme. The combined DID estimator compares the conditional before-after earnings of programme participants with those of non-participants. In the context of our study, this estimator compares the change in wage for workers who move from the informal to the formal sector to the wage change of workers who remain in the informal sector from one period to the next. So the effect of the treatment on the treated, in this context, the wage premium of the movers compared to the stayers turns to

$$\alpha^{MDID} = \frac{1}{N} \sum_{i \in I \to F} \left((\omega_{it}^{F} - \omega_{it'}^{I}) - \sum_{j \in I \to I} \phi(|p_i - p_j|)(\omega_{jt}^{I} - \omega_{jt'}^{I}) \right),$$

where $i \in I \to F$ denotes the individual who moves from being informally employed to formally and $j \in I \to I$ denotes those who stay in the informal status.

3.3.3 Empirical Results

We begin with estimating the propensity scores by a probit specification. As suggested by Dehejia and Wahba (2002), we start with a parsimonious model with most of the covariates, and in order to go through the balancing test, i.e. for all covariates, differences in means across treated and comparison units within each stratum are not significantly different from zero, the model is modified by adding some interaction terms. Results in Table 3.2 show that the significant independent variables are gender, hukou dummies for education levels, establish sizes, types of work unit, living in eastern provinces and the indicators which takes the value 1 if people are doing technological work. Addition-

ally, the interaction terms of dummies for working in the government-owned enterprise and large firms, registered as urban residence and living in eastern provinces, working in government owned and living in the east are significant. By caculating the marginal effects of the covariates, we can observe that the propensity score increases if people are male, relatively highly educated, if people work in the large enterprises and engaged in a technological job. Working in private and government-owned firms are both positively significant, so working in the collective group has a negative impact relative to other firms. However it is surprising to notice that registered as urban has a negative effect on the probability of being formally employed. It can be explained as that working formally will help the rural people obtain permanent city resident certificate, which is supposed to bring them many advantages. So the rural residence are much more prone to seize the opportunity to be employed formally. Among the interaction terms, if people work in a large and government-owned enterprise, the probability of being formally will rise a lot and similarly for the eastern urban residence. However working in the eastern government-owned firms will reduce this probability. The result of the Hosmer-Lemeshow goodness-of-fit test indicates that we cannot reject the good fit of the model. Thus we can say that our data is acceptably informative.

Figure 3.1 describes the histogram of estimated propensity scores for the two types of job. It is obvious that the critical condition of common support, which is imposed to ensure that propensity scores used from the control group fit within the minimum and maximum propensity scores from the treatment groups, is satisfied.

As we employ the propensity score matching method to account for the differences of characteristics between the two groups, it is necessary to check whether the characteristics distribute similarly. Figure 3.2 plots the distribution of each covariates for the treated and the control groups in the common support, where the distributions are very similar for each group and there are always very large overlaps. So the propensity

Table 3.2: Results of Probit estimation of propensity scores

	Coefficient	Marginal Eff.	p-value
Experience	-.002	.00	0.63
Gender	.14*	.04	0.03
Married	.13	.04	0.13
East	.20*	.06	0.19
Hukou	-.32*	-.11	0.03
Education			
university	.38*	.11	0.00
vocational degree	.40*	.11	0.00
high school	.31*	.09	0.00
middle school	-.01	-.003	0.87
Establishment size			
< 20	.04	.04	0.00
> 100	.94*	.23	0.00
Type of work unit			
government	3.03*	.87	0.00
private	.44*	.12	0.00
Type of occupation			
Technological	.46*	.09	0.00
Management	-.03	-.01	0.76
Service	-.12	-.03	0.31
Interaction Terms			
I(gov.large)	1.3*	.46	0.00
I(hukou.east)	.27*	.003	0.00
I(hukou.exp)	.01	.00	0.06
I(small.private)	-.37	-.12	0.08
I(gov.east)	-.67*	-.28	0.00
Constant	-1.58*		0.00
Common Support	4086	93.18%	
Hosmer-Lemeshow goodness-of-fit test			0.56

Notes: The dependent variable is 1 if the individual works formally. Asterisks denote significance at 5% level of t-test.

score is a good proxy of the characteristics, i.e. conditioning on similar propensity score is equivalent to conditioning on similar characteristics of employees and firms.

However in principle, the probability observing two units with exactly the same propen-

sity score is zero since the propensity score is a continuous variable. In order to identify the informal workers (control) who are acceptably matched to any given formal worker (treated), there are several matching algorithms after knowing the propensity scores. The most widely used are Nearest Neighbor Matching, Caliper Matching and Kernel Matching. Nearest neighbor matching reports for each treated unit a control unit with the closest propensity score and is usually used with replacement, that is certain control unit can be matched for several treated units. Then the difference between the outcome of the treated units and the outcome of the matched control units is computed. The ATT of interest is then obtained by averaging these differences. However there is pitfall of this matching method, as the closest matched control units can have a very different propensity score with the given treated. The caliper matching and kernel matching can properly address this problem. With caliper matching, each treated unit is matched only with the control units whose propensity score falls into a predefined neighborhood of the propensity score of the treated unit. A tolerance level is used on maximum propensity score distance to avoid risk of bad matches, and the smaller the caliper, the better the quality of the matches. On the other hand, there

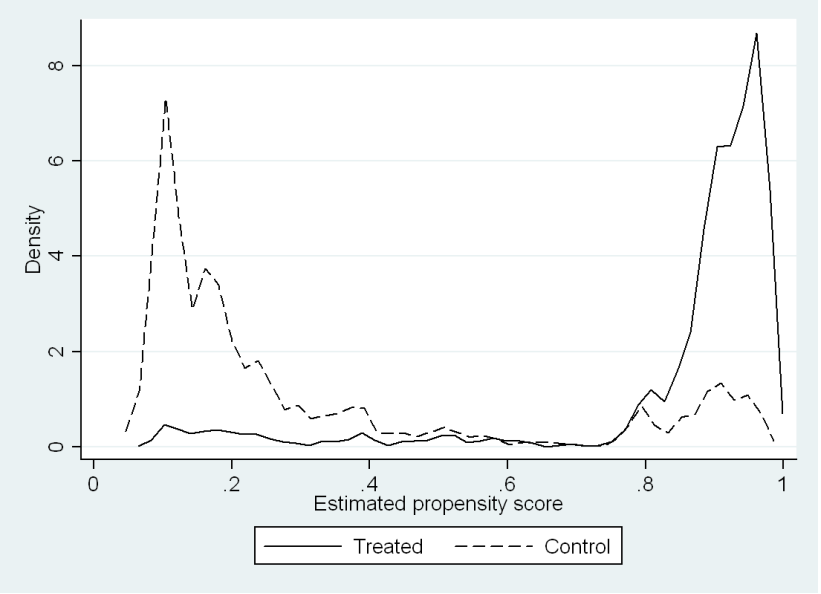

Figure 3.1: Distribution of Estimated Propensity Scores

may be some treated not matched for very small caliper. Kernel matching uses the weighted averages of all control units for each treated. The weights are computed as inversely proportional to the distance between the propensity scores of treated and controls. Three kernel functions are widely used in applied work: Gaussian, Triangular and Epanechnikov. Obviously there is a tradeoff between quality and quantity of matches among those three matching algorithms, and none of them is superior to the others. So people usually report the results by all three methods.

There is a not so large but acceptable amount of mobility in the informal employment sector in our data: 87 workers moved from informal employment to formal and 219 workers stayed. Table 3.3 reports the estimated average treatment effect for the treated, in our case, the wage premium for the movers. We choose a caliper of 10^{-3} for caliper matching and we also tried 10^{-4}, finding that there is little effect on the results. Kernel matching with Epanechnikov kernel function is stated in the table and the other two kernel functions mentioned above are also employed, while the results are similar. The wage premium of the movers, α^{MDID} is not significant. This suggests that workers who move to the formal sector do not see a change, especially an increase, in their wage than those who stay as informal employees. The result is also consistent with the study of China's labor market by Hu and Yao (2011). They, by introducing the lagged variables in the regression of wages, find that there is no significant wage decrease after the transition from informal to formal employment happens. Pratap and Quintin (2006)) at first only use the PSM method to estimate the wage premium of formal employment and then combine the PSM with DID estimator. The estimated formal premium prove either small and insignificant, or negative for their Argentine data.

However one concern of the insignificance is the possibility of decreasing the size of sample substantially. Just as pointed out by Smith and Todd (2005), results from a "matched" sample cannot necessarily be interpreted as representative of the whole

Table 3.3: Wage premium α^M estimated by DID PSM matching

Matching algorithms	α^{MDID}	St. Err.
Caliper ($\gamma = 10^{-3}$)	.809	.765
Nearest neighbor	.430	.679
Epanechnikov kernel	1.06	.699

Notes: The number of neighbors we specify here is 2. We also tried other kernel functions, such as Gaussian, and the results are very similar. None of the estimates are significant at the 5% significance level of t test.

sample.

3.3.4 Robustness Checks

The CIA assumption for our semi-parametric methods is very strong and thus the results might be under much suspicion. So in this section, we first flank the results we got from the DID PSM estimator by conducting other two conventional estimators: linear regression estimated by OLS and the panel data model with fixed effect. If we can obtain similar results from these estimators, then the matching method is more likely to be credible. Furthermore, we also evaluate the sensitivity of our estimator of the wage premium against various econometric specifications.

Linear Regression

We start by supposing the relationship between the wages and all the observed covariates is linear. In this section, a saturated model with interaction terms of the wage equation is implemented

$$\omega_i = \alpha_i + \beta_1 Z_i + \beta_2 X_i + \beta_3 Y_i + \beta_4 I_i + u_i,$$

where i represents the individual units; ω_i is the hourly wage; Z_i is the dummy variable taking a value of 1 if the individual is formally employed and 0 for informal employment; X_i and Y_i are the observed characteristics of the employees and jobs respectively and I_i denotes the interaction terms, which are the product of those dummies in X_i, Y_i and Z_i. The coefficient β_1 can be explained as the main effects of the type of jobs on the wages, which is of our greatest interest.

Table 3.4 presents the results estimated by OLS, which is the best linear unbiased estimation for a linear regression. The first three columns are the results for the whole sample, while the last two are those of the subsamples for the single and married. Model I controls for all observed covariates of the employees and jobs. The coefficient of the job type is insignificant there, but age, gender, whether the individual is settled in eastern provinces and registered as urban resident and works in small company are positively significant, and the years of experience and doing job of management are negatively significant. Model II additionally controls for interaction terms, which are the products of the dummies. The coeffient of a interaction term tells us how the main effects of one covariate differs by the other covariate. For example, the coefficient of the interaction term of job type and *Hukou* indicates how the formal-informal job effect differs by the type of the individual's residence, rural or urban. The results show that in this specification the main effect of formal-informal job is negatively significant, that is to say, there is wage premium of the informal employees, which is quantitatively about 50%. This result is a little out of line with studies of other countries. For example, Pratap and Quintin (2006) run OLS to estimate the wage differential between formal and informal jobs in Argentina and after controlling for individual and firm characteristics, the formal wage premium is 25 %. Badaoui et al. (2008) find that by controlling for human and job characteristics there is still a 37 % informal job wage penalty in South Africa. However, in China, similarly Xin (2001) finds the informal employees can earn 33.2 % more than the formal in big cities. Among the interaction terms, most of which are the products of the job type

and the characteristics, there will be positively significant formal job effect on wage if the individual is highly educated and married and negative effect if he works in large enterprise or doing management work. Additionally, if people work in a large and government-owned enterprise, there will be a positive effect on their income.

I have tried many different specifications of the model, especially with different interaction terms. It is very interesting to find that when the interaction between the job type and marital status is not controlled for, the main effect of formal-informal job turns to be insignificant, just as the result of the Model III for example. So I look into more details to the subsamples according to the marital status and see what happens to the job type effect on the wage, the results of which are shown in the last two columns. For the married group, the job type effect is still insignificant, while for the single, informal employees can earn significantly more than the informal. Additionally, we can observe a quite different wage equation between the two groups. So marital status is an essential consideration when people investigate the wage determinations for Chinese. Mindful of this result, we can conclude that the formal-informal job effect on the wage is sensitive to the model specification.

However one cannot ignore the assumption that we made before we run the OLS regression, especially the independence assumption of the error terms with the dependent variables conditional on all the regressors, i.e. in our case

$$Z_i \perp\!\!\!\perp u_i | X.$$

It requires that all the covariates explaining the wages could be observed. The R^2 of a model can tell us how much of the dependent variable can be explained by the model and in those models above, the R^2s are generally about 50%, which indicates that our data is acceptably informative in determining the wages.

Table 3.4: Linear Regression of Logarithm Hourly Wage Equations

	Model I	Model II	Model III	Single	Married
job	.08	-.49*	.04	.56*	.07
Experience	-.02	-.01	-.01	-.02	-.01
Gender	.24*	.07	.02	.28	.09
East	.34*	.37*	.37*	.41*	.35*
Hukou	.25*	.12	.10	.75*	-.01
Married	-.03	-.45*	-.04		
Health Cond.	.08	.12	.11	-.29	.15
Education					
university	.72*	.45*	.46*	.11	.69*
vocational degree	.69*	.44*	.46*	-.33	.61*
high school	.51*	.54*	.53*	.02	.68*
middle school	.06	.11*	.09	-.15	.19*
Type of work unit					
government	-.12*	-.35*	-.35*	-.57	-.31*
private	.23*	.25*	.25*	.32*	.26*
Establishment size					
< 20	.02	-.01	.01	.01	-.05
> 100	.38*	.22*	.17*	.11	.26*
Type of occupation					
Technological	.05	.05	.04	-.49*	.13*
Management	-.36*	-.28*	-.29*	.44*	-.33*
Service	.04	.05	.05	.58*	.01
Interaction Terms					
I(job.gender)		-.05	-.04	-.38	-.04
I(job.East)		.01	.01	-.25	.05
I(job.Hukou)		.18	.23*	-.31	.23*
I(job.HE)		.36*	.33*	.36	.35*
I(job.Married)		.65*			
I(job.Management)		-.09	-.08	-1.26*	.05
I(job.government)		.06	.05	-.11	.06
I(job.large)		-.23*	-.21	-.21	-.32*
I(gov.large)		.45*	.47*	.80*	.44*
I(female.exp)		-.01	-.01	.00	-.01
2004	.11	-.06	.01	.02	.00
2006	-.02	-.07	.13	-.05	.17
Constant	.34*	.83*	.57*	1.08*	.23
R^2	41.52%	43.58%	42.75%	53.57%	48.97%
N of Observations	4385			1578	2807

Notes: "HE" is the dummy for the highly-educated, which is calculated as the sum of "University"' and "Vocational Degree". Dependent variable is the logarithm hourly wages. Asterisks mean significance at 5% level of t tests.

Fixed-Effect Panel Data Model

Although our data is acceptably rich in employee- and job-level characteristics, there may still be some other unobserved factors both influential to the wages and the selection into formal or informal sector. One example could be the individual productivity uncorrelated to the education and the non-pecuniary benefit of a job. The failure to account for such characteristics could lead to biased estimation of the job type effect on the wage. As our data is longitudinal with three surveyed waves, we can also run a panel data model. Fixed effect models can wipe out the unobserved individual-specific effects if they are time invariant. Another possibility would be to use an instrumental variables approach, however just as argued by Badaoui et al. (2008), it is notably difficult to find instruments that determine selection into sectors but are convincingly independent of the wage. The two-way fixed effect model implemented in this paper is as follows

$$\omega_{it} = \alpha_i + \lambda_t + \beta_1 Z_{it} + \beta_2 X_{it} + \beta_3 Y_{it} + \beta_4 I_{it} + u_{it},$$

where i and t are the individual and time indexes, α_i and λ_t are the individual and time-specific effects respectively and other parameters are kept consistent with the one-dimensional linear regression. Treating the individual effects as parameters to be estimated is algebraically the same as estimation in deviations from means. That is to say, we need first calculate the individual means

$$\bar{\omega_i} = \alpha_i + \bar{\lambda} + \beta_1 \bar{Z_i} + \beta_2 \bar{X_i} + \beta_3 \bar{Y_i} + \beta_4 \bar{I_i} + \bar{u_i}.$$

Then the deviations from means are estimated

$$\omega_{it} - \bar{\omega_i} = \lambda_t - \bar{\lambda} + \beta_1 (Z_{it} - \bar{Z_i}) + \beta_2 (X_{it} - \bar{X_i}) + \beta_3 (Y_{it} - \bar{Y_i}) + \beta_4 (I_{it} - \bar{I_i}) + u_{it} - \bar{u_i}.$$

So the time invariant individual characteristics, no matter observed or unobserved, are purged by the deviations. The results are provided in Table 3.5, where the formal-

informal job effect is not significant. Additionally, people with more experience and doing technological work in government-owned enterprises can expect higher income and in contrast, working in small and private firms has a negatively significant on the level of income. Surprisingly, we also observe a negative effect of working in government-owned companies. This can be explained as, we actually investigate the wage difference among the job movers,either from formal to informal or the other way round, and in China the pecuniary income in the government agencies is not competitive, in return the workers there can obtain much more in-kind or gray income.

As the fixed effect panel data model also controls for the unobserved effects, the estimates of those coefficients should be smaller than those in the cross section equations as we did in the last section, especially when the time-invariant unobservables are an important factor behind the formal or informal wages. Badaoui et al. (2008) find that in their South African data, the informal wage penalty falls from 37 % to 17.9 % after taking the unobservables into account. However, just as we showed previously, fixed effect panel models focus on the effects when transition happens. In our case, this model only estimates the formal informal job effects for the job movers, or in other words, it actually only shows the formal informal job effects on wage for the job movers, while in our data the job stayers account for a larger percentage. There are in total 2802 employees but the data is not balanced panel data, i.e. 530 individuals are surveyed consistently for 3 waves, 565 individuals for 2 waves and 1707 are reported only once, so there are altogether 4385 observations in the data. Through those three waves,the movers only account for less than 10 %, including 87 workers moving from informal employment to formal employment and 132 from formal to informal. Then the small sample size might be another reason for the insignificance of the coefficients.

Table 3.5: Fixed-Effect Panel Data Estimation of Logarithm Hourly Wage

	Coeff.	p-value
job	-.11	.09
Experience	.39	.00
East	.10	.19
Hukou	.17	.25
Health Cond.	.01	.95
Type of work unit		
government	-.22	.03
collective	-.11	.18
Establishment size		
< 20	-.01	.96
> 100	.03	.71
Type of occupation		
Technological	-.16	.07
Management	-.08	.46
Craft	-.06	.43
Interaction Term		
I(job.East)	-.09	.48
I(job.Hukou)	-.04	.61
I(job.Management)	.11	.27
I(job.government)	.13	.15
I(priv.small)	-.21	.03
I(gov.large)	-.06	.51
I(tech.gov)	.28	.00
I(female.experience)	-.05	.57
Constant	1.91	.00
Number of Individuals	2802	
Number of Observations	4385	

Robustness to Different Model Specifications

Table 3.6 shows the estimation results of the formal wage premium after dropping the covariates of education levels, establishment size and type of work unit respectively. The wage premium stays insignificant at 5% level without education in the estima-

Table 3.6: Wage premium by kernel matching for different specifications

	α^{MDID}	St. Err.	t-statistic
No Education	1.26	.66	1.92
No Establish. Size	.840	.627	1.34
No Type of Work Unit	.953	.673	1.42
No Establish. Size & Type of Work Unit	1.46*	.634	2.30

Notes: "No ." means "." is dropped in estimation of wage premium. Asterisks mean significance at 5% level of t test.

tion.[3] This can be explained as that people with different education levels can earn similarly across the sectors. Those variables significantly could affect the wage level for both formal and informal employment but add no information in explaining the wage difference between the two sectors after controlling for other characteristics of jobs and employees.

It is more interesting to find that dropping either only establishment size or only the type of work unit does not change the insignificance of wage premium, while cutting both of them from the model, the premium becomes very significant and positive. This may be caused by the dependence between the firm size and working unit type, the correlation of which turns to be 0.35. So in China, usually the government-owned enterprises are in much larger scale than the private ones. Peng (2007) also concludes that wage returns to firm size in those enterprises is much larger. Ye et al. (2011) find that monopoly of some government owned enterprises plays a significant positive effect on the wage difference with the private firms. Actually both of these two dummy variables can capture the information of firm sizes. This could be the reason why dropping only one of those variables makes no difference in the final results. However without any information about the firm size, a significant wage premium emerges. It is commonly viewed that larger firms or organizations offer more attractive wage and better social security in most countries. And size of establishment in the formal sector

[3]It is significant at the level of 1%. To keep along with the other tests in this paper, we hereby do not explain the education effect in details.

tends to be larger. So it follows that without considering the establishment size in estimation, the wage premium should be significantly positive.

People could interpret the size wage premium as evidence that the labor market is segmented according to the firm size. Just as mentioned in Section 2 that some studies also define informal employment as employment in small firms, although this definition is in our opinion inappropriate and not yet supported by evidence. There are a variety of papers studying the relation between firm size and wage differentials and many of them find there is a size effect in explaining the earning gaps. Oi and Idson (1999) find that in the U.S. men employed in firms with 1,000 or more employees earn 45 % more on average than men employed in firms with fewer than 25 employees. For women, the ratio is 1.30. And in Japan in 1988, the earnings ratio between firms with 1,000 employees or more and firms with 10 to 99 employees was 1.69 for men, 1.68 for women. Oosterbeek and van Praag (1995) after comparing wages in large (more than 500 employees) and medium-sized firms (between 20 to 500 employees) in Netherlands show that the wage difference is one-third due to the fact that large companies are in a position to attract workers with more favorable endowments, and two-thirds due to differing wage regimes. Chen et al. (2010) also find that firm size is still significant in explaining the wage differential after controlling for the externality of China's labor market, using data of firms in Zhejiang province in China. Overall our finding of size-wage premium by semi-parametric method is in line with those in other countries, but it is seldom considered as evidence of segmentation of labor markets, as the reason for its existence of this premium is still sparking much debate.

A drawback of this study is the identification of the informal employment. Due to the data availability, we only took the employees with permanent contracts as formal. However just as reported by Henley et al. (2006), the probabilities of the temporary employees and self-employed working formally are not zero. For the short-term employees, only those with high wage offers are willing to submit taxes under the social

regulation. So the informal wage penalty could be underestimated by our specification.

3.4 Conclusion

We ask whether individuals formally employed can significantly earn more than those working informally in China. Using the data from the China Health and Nutrition Survey (CHNS), we applied the differences-in-differences propensity score matching estimator to compare the wages between the formal and informal employees with similar human and job characteristics. The results tell us that we cannot find any evidence that the Chinese labor market is segmented along the formal-informal line.

From the descriptive statistics of our data, we can observe that the Chinese formal and informal employees vary widely in human and job characteristics and there is no apparent wage difference between the two types of employees. Additionally the Chinese informal workers on average earn slightly more than the formal ones, just as found by Xin (2001). In order to make our estimation more plausible against the CIA assumption, we proceed to using the combined estimator of the propensity score matching and differences-in-differences estimator in the main analysis. The results show that for the workers who move from the informal employment to formal employment, there is no significant increase or change in wages compared to those who stay as informal employees. It is worth noting that the estimation of the propensity score by probit model passes the goodness of fit test, which indicates that the included covariate is acceptably rich to explain the selection of formal-informal jobs and thus adds some weight on the plausibility of our causal effects.

In order to check the robustness of our estimator, I also carry out two other methods, linear regression estimated by OLS and panel data model with fixed effect, the results of which are similar to those from DID PSM estimator. Together with the results

by Chen and Hamori (2009), we could conclude that the wage differences can be mostly explained by the divergence of the human and job characteristics rather than the different types of job. We also test the sensitivity of the results to difference model specifications in estimating the propensity scores and conclude that controlling for the establishment characteristics, such as type of ownership, establishment size, is important. Larger firms seem to pay more in China and also most countries, and government-owned enterprises are usually in very large scales, so these two variables both capture the size effect of wage difference, then at least one of them must be considered while comparing the wage difference across sectors. In general, as almost all the results are consistently arguing that we cannot find a significant formal wage premium for our Chinese data, the strong CIA assumption is somehow fulfilled in our model and thus the DID PSM estimator is credible.

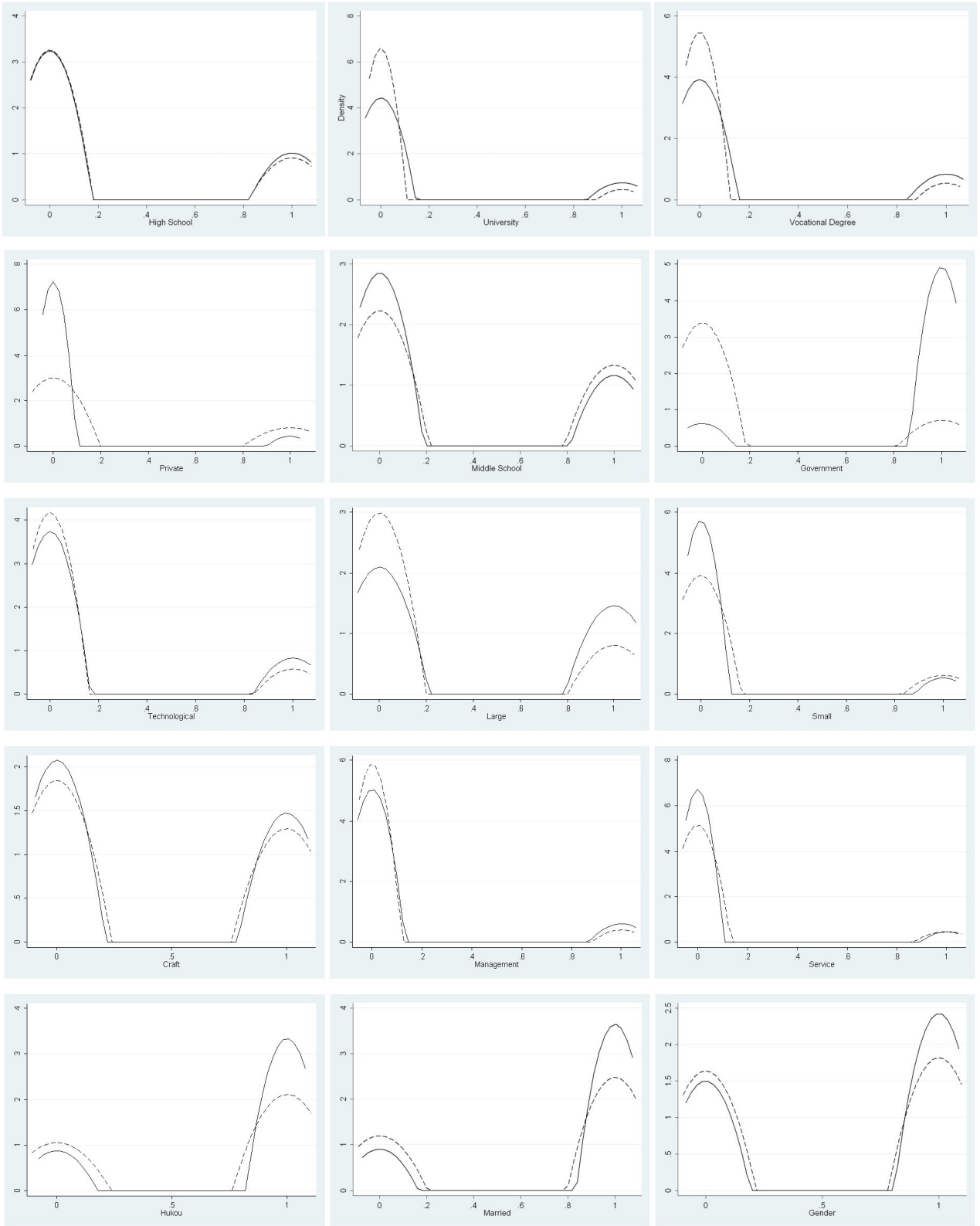

Figure 3.2: Kernel Density Estimates

Proof for Chapter 1

LEMMA Suppose:

$$l = O_p(T^{\frac{1}{3}}) \tag{A.1}$$

$$O_p(l^{-q}) = o_p(lT^{-1}) \tag{A.2}$$

we have

$$q \geq 2. \tag{*}$$

PROOF From my understanding, l must be a sequence of random variables, rather than a single random variable. Because the definition of "the order in probability" concerns about the convergence of a sequence of random variables. Therefore, I would like to use l_N henceforth. (A.1) and (A.2) then become

$$l_T = O_p(T^{\frac{1}{3}}) \tag{A.1'}$$

$$O_p(l_T^{-q}) = o_p(l_T T^{-1}) \tag{A.2'}$$

From the definition, (A.1') could be interpreted as: for any $\epsilon > 0$, there exist a pair of M_{ϵ_1}, N_{ϵ_1} such that for all $T > N_{\epsilon_1}$

$$P\{\frac{\mid l_T \mid}{T^{\frac{1}{3}}} > M_{\epsilon_1}\} < \epsilon$$

We then immediately claim that there exist a $N_{b_1} > N_{\epsilon_1}$ such that for all $T > N_{b_1}$

$$P\{\log_T(\mid l_T \mid) > \frac{1}{3}\} < \epsilon \tag{A.3}$$

(A.3) is indeed true and could be seen by assuming the opposite.

(A.2') implies that: if H_T is a sequence of random variables and $H_T = O_p(l_T^{-q})$, $H_T = o_p(l_T T^{-1})$ then must hold. Analogously, $H_T = O_p(l_T^{-q})$ could be interpreted as: for any $\epsilon > 0$, there exist a pair of M_{ϵ_2}, N_{ϵ_2} such that for all $T > N_{\epsilon_2}$

$$P\{\frac{\mid H_T \mid}{l_T^{-q}} > M_{\epsilon_2}\} < \epsilon \tag{A.4}$$

We then immediately claim that the above is true is and only if there exist a $N_{b_2} > N_{\epsilon_2}$ such that for all $T > N_{b_2}$

$$P\{\log_{|l_T|}(\mid H_T \mid) > -q\} < \epsilon \tag{A.5}$$

When (A.4) holds, (A.5) is indeed true and could be seen by assuming the opposite. Let's now take a look at the opposite direction. When (A.5) holds, $\frac{|H_T|}{l_T^{-q}}$ must be bounded in the limit in probability, i.e., (A.4) is also true. Therefore, (A.4) and (A.5) are equivalent. (A.5) is also equivalent to

$$P\{\frac{\log_T(\mid H_T \mid)}{\log_T(\mid l_T \mid)} > -q\} < \epsilon$$

or

$$P\{\log_T(\mid H_T \mid) > -q \log_T(\mid l_T \mid) \text{ and } \log_T(\mid l_T \mid) > 0\} < \epsilon \tag{A.6}$$

$H_T = o_p(l_T T^{-1})$ could be interpreted as: for any $\eta > 0$, it holds that

$$P\{\lim_{T\to\infty} \frac{\mid H_T \mid}{l_T T^{-1}} > \eta\} = 1$$

We then immediately claim that for any $\epsilon > 0$, there exist a N_{b_3} such that for all $T > N_{b_3}$

$$P\{\log_T(\mid H_T \mid) \geq \log_T(\mid l_T \mid) - 1\} < \epsilon \tag{A.7}$$

(A.7) is indeed true and could be seen by assuming the opposite. Now define

$$N_b = \max(N_{b_1}, N_{b_2}, N_{b_3})$$

Therefore, (A.3), (A.6) and (A.7) must apply for all $T > N_b$.

The fact that if H_T is a sequence of random variables and $H_T = O_p(l_T^{-q})$, $H_T = o_p(l_T T^{-1})$ then must hold, implies that if (A.6) holds, (A.7) is true. We then claim for all $T > N_b$,

$$-q \log_T(\mid l_T \mid) < \log_T(\mid l_T \mid) - 1$$

i.e.,

$$\frac{1}{1+q} < \log_T(\mid l_T \mid) \tag{A.8}$$

Combining (A.8) and (A.3), we have for all $T > N_b$,

$$P\{\frac{1}{1+q} > \frac{1}{3}\} < \epsilon$$

i.e.,

$$P\{2 > q\} < \epsilon \tag{A.9}$$

We know that q is a constant rather than a random variable. Thus (9)implies (*). ■

Biography

Bibliography

Arellano, M. and Carrasco, R. (2003). Binary choice panel data models with predetermined variables. *Journal of Econometrics*, 115(1):125–157.

Badaoui, E. E., Strobl, E., and Walsh, F. (2008). Is there an informal employment wage penalty? Evidence from South Africa. *Economic Development and Cultural Change*, 56:683–710.

Baltagi, B. (2005). *Econometric Analysis of Panel Data*. John Wiley and Sons, 3rd ed. edition.

Bashmakov, I. (2004). Thresholds in residents? ability and willingness to pay for housing and utility services (russian). *Voprosy Ekonomiki*, 4.

Bashmakov, I. (2005). Payment discipline - integral parameter of success of reforms in the russian utility services sector (russia). Technical report, Center for Effective Energy Utilization.

Brand, J. E. and Halaby, C. N. (2006). Regression and matching estimates of the effects of elite college attendance on educational and career achievement. *Social Science Research*, 35(3):749 – 770.

Chamberlain, G. (1985). Heterogeneity, omitted variable bias and duration dependence. In Heckman, J. J. and Singer, B., editors, *Longitudinal Analysis of Labor Market Data*, chapter 1, pages 3–38. Cambridge University Press.

Chen, G. and Hamori, S. (2009). Formal employment, informal employment and income differentials in urban China. MPRA Paper 17585, University Library of Munich, Germany.

Chen, L., Li, H., Xiong, Y., and Zhou, L. (2010). The influence of firm size on the wage of employees: Evidence from the Chinese competitive labor market(in Chinese). *Journal of Financial Research*, 2:18–30.

CSB (2010). China statistical book (in Chinese).

Dehejia, R. and Wahba, S. (2002). Propensity Score-Matching Methods For Nonexperimental Causal Studies. *The Review of Economics and Statistics*, 84(1):151–161.

Fankhauser, S., Rodionova, Y., and Falcetti, E. (2008). Utility payments in ukraine: Affordability, subsidies and arrears. *Energy Policy*, 36(11):4168–4177.

Fankhauser, S. and Tepic, S. (2005). Can poor consumers pay for energy and water? an affordability analysis for transition countries. Working Papers 92, European Bank for Reconstruction and Development, Office of the Chief Economist.

Gong, X. and van Soest, A. (2002). Wage differentials and mobility in the urban labour market: a panel data analysis for Mexico. *Labour Economics*, 9(4):513–529.

Guzanova, A. (1998). The housing market in the russian federation: Privatization and its implications for market development. Working Papers 1891, World Bank Policy Research.

Ham, J. C., Reagan, P. B., and Li, X. (2005). Propensity score matching, a distance-based measure of migration, and the wage growth of young men. FRB of New York staff report no. 212, IEPR Working Paper No. 05.13.

Hamilton, E., Banerjee, S. G., and Lomaia, M. (2008). Exploring housing subsidies to households in russia. *Journal of International Development*, 20(3):257–279.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1999). Characterizing selection bias using experimental data. *Econometrica*, 66(5):1017–1098.

Heckman, J. J. and Hotz, V. J. (1986). An investigation of the labor market earnings of panamanian males evaluating the sources of inequality. *Journal of Human Resources*, 21(4):507–542.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64(4):605–54.

Henley, A., Arabsheibani, G. R., and Carneiro, F. G. (2006). On defining and measuring the informal sector. IZA Discussion Papers 2473, Institute for the Study of Labor (IZA).

Hu, A. and Yang, Y. (2001). The paradigm shift of employment: formal to informal? analysis of China's urban informal employment (in Chinese). *Management World*, 02.

Hu, A. and Zhao, L. (2006). The informal employment and informal economy in urban China during transition (in Chinese). *Journal of Tsinghua University*, 21(2).

Hu, F. and Yao, X. (2011). Urban employee's behaviors of choosing the informal employment and labor market segmentation: An analysis based on a panel data (in Chinese). *Journal of Zhejiang University (Humanities and Social Sciences)*, 2:191–199.

ILO (2002). Women and men in the informal economy: A statistical picture. *Geneva, Switzerland: International Labor Organization.*

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29.

Jin, Y. (2006). Female irregular and unstable employment: Current situation and

countermeasure (in Chinese). *Journal of Hehai University (Philosophy and Social Sciences)*, 1:6–10.

Kolenikov, S. and Shorrocks, A. (2005). A decomposition analysis of regional poverty in russia. *Review of Development Economics*, 9(1):25–46.

Lampieti, J., Kolb, A., Gulyani, S., and Avenesyan, V. (2001). Utility pricing and the poor: Lessons from armenia. Technical Paper 497, World Bank.

Lechner, M. (2000). Programme heterogeneity and propensity score matching: An application to the evaluation of active labour market policies. Econometric Society World Congress 2000 Contributed Papers 0647, Econometric Society.

Lechner, M. (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society Series A*, 165(1):59–82.

Li, P. (2009). The change and characteristics of current employment situation (in Chinese). *Social Science Review*, 2:97–103.

Magnac, T. (1991). Segmented or competitive labor markets. *Econometrica*, 59(1):165–187.

Maloney, W. (1999). Does informality imply segmentation in urban labor markets? evidence from sectoral transitions in Mexico. *World Bank Economic Review*, 13(2):275–302.

Mazumdar, D. (1981). The urban labor market income distribution: A study of Malaysia. *Oxford University Press*.

Oi, W. Y. and Idson, T. L. (1999). Workers are more productive in large firms. *American Economic Review*, 89(2):104–108.

Oosterbeek, H. and van Praag, M. (1995). Firm-size wage differentials in the Netherlands. *Small Business Economics*, 7(3):173–82.

Peng, Z. (2007). Firm size, ownership and wages (in Chinese). *Journal of Shanxi Finance and Economics University*, 03.

Pratap, S. and Quintin, E. (2006). Are labor markets segmented in developing countries? A semiparametric approach. *European Economic Review*, 50(7):1817–1841.

Roberts, B. (1989). Employment structure, life cycle, and life chances: Formal and informal sectors in Guadalajara informality revisited. *The informal economy: Studies in advanced and less develped countries, The Johns Hopkins University Press.*

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Smith, A. J. and Todd, E. P. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2):305–353.

Tansel, A. (1999). Formal versus informal sector choice of wage earners and their wages in Turkey. *Economic Research Forum Working Paper*, 9927.

Wu, Y. and Cai, F. (2006). Informal employment in urban China: Size and characteristics (in Chinese). *China Labor Economics*, 2.

Xin, M. (2001). The informal sector and rural-urban migration-a Chinese case study. *Asian Economic Journal*, 15(1):71–89.

Xu, L. (2008). The effects of labor market segmentation on the rural labor supply (in Chinese). *Reform of Economic System*, 3:36–39.

Ye, L., Li, S., and Luo, C. (2011). Industrial monopoly, ownership and enterprises wage inequality: An empirical research based on the first national economic census of enterprises data (in Chinese). *World of Management*, 04:26–37.

Abstract

The first chapter considers the issue that in dynamic panel data models where there is no cross-sectional dependence, the overidentifying test statistic should be the number of cross sections times the GMM minimand. Dealing with the unknown autocorrelation structures of moment conditions, the heteroscedasticity and autocorrelation consistent (HAC) covariance matrix GMM estimator is incorporated, and it is shown that the bootstrap approximation outperforms the first order approximation. As suggested by Kapetanios (2008), I construct bootstrap samples by resampling the whole cross sectional units with replacement instead of block bootstrap. Monte Carlo simulations show that this bootstrap has good size properties which outperform bootstrap with Arellano and Bond GMM estimators in all cases based on different settings.

In the second chapter, based on three rounds of Russian Longitudinal Monitoring Survey (RLMS), incidence of non-payments for household utilities in Russia has been analyzed by a number of variables. Taking the influence of the past record of non-payment into account, a dynamic binary choice panel data model estimated by GMM methods indicated that the categories of households that are more likely to have non-payments are those with unpaid records, higher proportion of utility expenditure, unemployed household head, school-age children, and those living not in a private apartment or house. Moreover, wage arrears and poverty status of a household appear to be significant factors affecting existence of utility payment arrears. Estimation suggests that existence of non-payments for utilities might be both an issue of poverty

and low willingness to pay higher bills for services of the sector, which has been stopped being subsidized substantially after the collapse of the Soviet Union.

The last chapter tests the labor market segmentation hypothesis, i.e. whether workers formally employed can earn significantly more than those informally employed, using data from the China Health and Nutrition Survey (CHNS). Instead of parametric approach which suffers from some econometric problems, the propensity score matching approach is implemented in estimating the formal employment wage premium. Comparing individuals with similar employee and firm characteristics, it can not be concluded that formal jobs are superior to informal ones in terms of income. We also test the robustness of our results for different model specifications and sensitivity to unobserved covariates.

Zusammenfassung

Das erste Kapitel befasst sich mit dem Thema, dass in dynamischen Paneldaten-Modellen, wo es keine Querschnitts-Abhängigkeit gibt, die overidentifying Teststatistik das Produkt aus der Anzahl der Querschnitte und des GMM minimand sein sollte. Auf die unbekannten Autokorrelation-Strukturen der Moment-Bedingungen eingehend, werden die Heteroskedastizität und der Autokorrelations-konsistente (HAC) Kovarianzmatrix GMM Schätzer aufgenommen, und es wird gezeigt, dass die Bootstrap Angleichung die Näherung erster Ordnung übertrifft. Wie von Kapetanios (2008) vorgeschlagen, konstruiere ich Bootstrap-Samples durch Resampling der gesamten Querschnittseinheiten mit Austausch anstelle von Block Bootstrap. Monte-Carlo-Simulationen zeigen, dass dieser Bootstrap gute Grösseneigenschaften aufweist, welche den Bootstrap mit Arellano und Bond GMM Schätzern in allen auf unterschiedlichen Parametern basierenden Fällen übertrefen.

Im zweiten Kapitel wurde basierend auf drei Runden der russischen Longitudinal Monitoring Survey (RLMS) das Auftreten von Nichtzahlungen für Haushaltsartikel in Russland anhand einer Reihe von Variablen analysiert. Unter Berücksichtigung des Einflusses des vergangenen Zahlungsverhaltens, zeigte ein durch GMM Methoden geschätztes dynamisches binäres Auswahl-Panel-Daten-Modell, dass die Kategorien der Haushalte, die eine höhere Wahrscheinlichkeit von Nichtzahlungen aufweisen, diejenigen mit Aufzeichnungen von Nichtzahlungen in der Vergangenheit, einem höheren Anteil von Versorgungsausgaben, einem arbeitslosen Haushaltsvor-

stand, mit Kindern im Schulalter sowie Bewohner von nicht privatisierten Wohnungen oder Häusern sind. Darüber hinaus erscheinen Lohnrückstände und Ausmass der Armut eines Haushalts signifikante Faktoren für das Auftreten von Zahlungsrückstanden bei Versorgungsausgaben zu sein. Die Schätzung legt nahe, dass die Existenz von Nichtzahlungen für Versorgungsausgaben sowohl eine Frage der Armut als auch der niedrigen Bereitschaft ist, höhere Rechnungen für Dienstleistungen des Sektors, dessen umfangreiche Subventionierung nach dem Zusammenbruch der Sowjetunion gestoppt wurde, zu zahlen.

Das letzte Kapitel testet die Hypothese der Segmentierung der Arbeitsmärkte, das heisst, ob formell angestellte Arbeitnehmer deutlich mehr als informell Beschäftigte verdienen, basierend auf Daten aus der China Health and Nutrition Survey (CHNS). Anstelle eines parametrischen Ansatzes, welcher unter einer Reihe ökonometrischer Probleme leidet, wird der Propensity-Score-Matching-Ansatz für die Schätzung des Lohnvorteils einer formellen Beschäftigung angewendet. Aus dem Vergleich von Beschäftigten mit ähnlichen persönlichen- und Firmen-Eigenschaften kann nicht die Schlussfolgerung gezogen werden, dass formell Angestellte den informell Beschäftigten hinsichtlich der Einkommenshöhe überlegen sind. Ausserdem testen wir die Robustheit unserer Ergebnisse für verschiedene Modell-Spezifikationen und die Sensibilität für unbeobachtete Kovarianten.

Curriculum Vitae

Personal Data

Name:	Miaomiao Yan
Address:	Linzerstrasse 283/8, 1140, Wien
Email:	miaomiao.yan@univie.ac.at
Date of Birth:	July 1, 1981

Education

Since 10/2007	**PhD Program Initiativkolleg(IK) in economics**
	University of Vienna (Expected graduation in summer 2012)
09/2005-08/2007	**Postgraduate Program in economics**
	Institute for Advanced Studies (IHS), Vienna
09/2003-07/2005	**Master program in economics**
	Harbin Institute of Technologies, China (diploma with distinction)
09/1999-07/2003	**Bachelor in economics**
	Harbin Institute of Technologies, China (diploma with distinction)

Scholarships

03/2009	**National Award, China Scholarship Council**
	for Outstanding Self-financed Chinese Students Study Abroad
09/2005-08/2007	**Austrian National Bank Fellowship**
	Insitute for Advanced Studies(IHS), Vienna

Academic Positions

10/2007-06/2011	**Department of Economics,** University of Vienn
	College Assistant

Working Experience

08/2012-10/2012	Internship at the United Nations Industrial Development Organization(UNIDO)
03/2008-07/2008	tutor of **Applied Econometrics**, University of Vienn
	course for graduate students in economics

Languages

Chinese	**Mother tongue**
English	**Excellent**
	Primary language in postgraduate studies
German	**Intermediate (B1)**

Research Interests

Applied and Theoretic Econometrics, Panel econometrics, Empirical Labor Economics