Long Panel Data Methods in Econometrics and Its Application on Minimum Wages

Economics Honors Program

By:

Alfredo Nicholas Effendy

Advisor:

Prof. Matthew Harding

Department of Economics

University of California, Irvine

29th May 2018

Table of Contents

1. Introduction 1

2. Models 3

3. Monte Carlo Simulation 5

4. Empirical Application 12

5. Discussion 14

6. Appendix: Data 17

References 18

1. Introduction

Panel data models allow scholars to analyze a group of people over time. However, these applications tend to only work in short panel data. One of the main reason is most of the models developed by economists only work well (unbiased estimators) for short panel data. In analyzing short panel data models, scholars often used the assumption that the experiment subjects behave uniformly across the experimental period. Since most drastic changes require lengthy period, the assumption works well. Unfortunately, when the models are extended to long panel data, which requires longer time frame, this assumption does not always hold. The longer the experiment takes place, it allows the experiment subjects to have more opportunity to behave differently from the previous known patterns. This leads to the tendency of getting biased result when short panel data analysis is applied to long panel datasets.

In the last decade, economists had been trying to evaluate this issue by applying various approaches in addressing the problem. Pesaran (2006) and Bai (2009) stated that the standard fixed effects method on long panel data is be biased and not appropriate. So, they introduced an approach that is called time-varying individual effects. Harding and Lamarche (2011 and 2014) implemented this approach in both Monte Carlo simulation and datasets. The results show that the time-varying individual effects give a better approximation and model compared to the standard fixed effects model. Besides, they also implemented differences-in-differences estimations that gave a similar problem.

One possible empirical application of this long method is the minimum wage problem that arises among teen workers. Neumark, Salas, and Wascher (2014) addresses a critic from DLR (2010) and ADR (2011) who suggest that minimum wage does not influence teen workers' employment rate. They counter this statement by applying fixed effects approach to show that

minimum wage has a negative effect on teen employment. This problem is interesting, because Pesaran and Bai state that standard fixed effects method is biased and inappropriate for long panel data. Therefore, this study is trying to apply the past findings long panel data analysis and apply them to the minimum wage problem among teen workers.

The first step in doing the analysis is to do Monte Carlo simulation to determine the biasness of the model. For example, OLS in a simple model is unbiased (**Figure 1**), so we will know whether OLS is biased for long panel data or not by applying it and compute the biases. Same approaches will be used for the fixed effects estimator and time-varying individual effects. Then, use the different estimators on the same model and dataset used by Neumark, Salas, and Wascher. Afterwards, the consistency and the robustness of each estimator are analyzed. Finally, the results from Monte Carlo simulation are used determine which method is the most sufficient in addressing the minimum wage problem on teen workers.

The minimum wage problem is important and interesting, especially for teen workers employment. Teenagers are highly affected with the minimum wage policy, because employers tend to see their lack of experience and skills as a weakness compared to more experienced workers. However, when the minimum wage is low enough and there is a vacancy in the positions in which teen workers are suitable for, it is beneficial for the employers to hire teen workers. One of the main reason is because the employers can exploit the fact that they can pay the teen workers with the minimum wage compared to hiring more experienced people with higher pay. Nevertheless, when the minimum wage is too high, the employers tend to be reluctant in hiring teen workers. They can get better workers who are willing to get paid with the minimum wage because it is high enough. Concluding all the facts above, minimum wage has a big impact in the employment of teen workers.

2. Models

There are several models that are used in this study, each of them has different purposes. The first model is the basic linear model; it is unbiased under most methods and approaches. So, this is used to show that the methods are unbiased under the linear model.

Model:

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i \quad \textbf{(1)}$$

Initialization:

$$\beta_1 = 1,\ \beta_2 = 2,\ \epsilon_i \sim N(0,1),\ x_i \sim N(1,1)$$

$$Sample\ Size(N) = 50, 100, 1000, 10000\ \ \#Simulation = 1000$$

The second model is the modified panel data model from Harding and Lamarche (2014). It is used as the starting point to investigate biases of the methods used in this study (OLS, Fixed Effects (FE), and Common Correlated Effects (CCE).

Model:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \gamma d_t + \lambda_i' F_t + u_{it} \quad \textbf{(2)}$$

$$x_{it} = \pi_0 + \pi_1 w_{it} + g d_t + li' F_t + a\lambda_i' F_t + v_{it} \quad \textbf{(3)}$$

$$F_{jt} = \rho_f F_{jt-1} + \eta_{jt} \quad \textbf{(4)}$$

$$\eta_{jt} = \rho_\eta \eta_{jt-1} + \epsilon_{jt} \quad \textbf{(5)}$$

Initialization:

$$\beta_0 = \pi_0 = l = 2,\ \beta_1 = \gamma = \pi_1 = g = 1, \rho_f = 0.9, \rho_\eta = 0.25$$

$$\omega_{11} = \omega_{22} = 1, d_t \sim N(0,1), \lambda_{i1}, \lambda_{i2} \sim N(1,0.04), (u_{it}, v_{it})' \sim N(0, \omega),\ \ e, w \sim N(0,1)$$

$$j = 1,2\ , \qquad t = -49, \dots, 5$$

Design I:

$a = 0, \omega_{12} = \omega_{21} = 0$

Design II:

$a = 2, \omega_{12} = \omega_{21} = 0$

Design III:

$a = 2, \omega_{12} = \omega_{21} = 0.5$

In this panel data model, the error terms v and u are the error terms of x_{it} and y_{it} respectively. It is assumed that some components in v are stochastically dependent on u. The x_{it} and d_t represent the individual and common observed regressors. Meanwhile, F_t and λ_i denote the unobserved common regressors and individual loadings. Lastly, the w_{it} represents the vector instrument that satisfied the identification conditions.

The third model is the model used by Neumark, Salas, and Wascher in their minimum wage study. This paper uses the same model to compare the result of the different estimators.

Model:

$$\log(y_{it}) = \beta_0 + \beta_1 \log(x_{1,it}) + \beta_2 x_{2,it} + \beta_3 x_{3,it} + \epsilon_{it} \quad \textbf{(6)}$$

Description:

$y: Employment\ per\ Population$

$x_1: Minimum\ Wage$

$x_2: Unemployment\ Rate$

$x_3: Relative\ Size\ of\ Youth\ Population$

3. Monte Carlo Simulation

The first two models mentioned in the previous section use Monte Carlo simulation as their environment to produce the desired result. Monte Carlo simulation performs repetitive random sampling with the desired amount of replication. The randomized values are based on the given distribution that has a range of value and each of them has different probability. Even though the values are randomized, the average value of the estimated parameter will converge to the actual mean as the number of observations gets higher. This property is very useful in determining the biasness of a model. Given a biased model, the difference between the average value of the estimates and the actual mean will not be 0. Furthermore, this simulation can generate the random values that behave like the desired models, which act like a dataset. Therefore, people can take advantage of Monte Carlo simulation by applying it to the other methods to check the biasness of that methods.

One of the approaches that is used in this study is the OLS Estimator. Applying Monte Carlo simulation to this method would be able to determine its biasness. This method is applied to both the linear model and panel data. The expected outcome from this approach is getting an unbiased result for the linear model and biased estimators for the panel data.

Using the model in **Equation (1)**, an illustration regarding the biasness of the model is conducted by applying the Monte Carlo simulation and OLS. The parameter of interest is β_2 which is the slope of the graph. After running the simulations with different number of repetition in each of them, the following graphs were obtained.

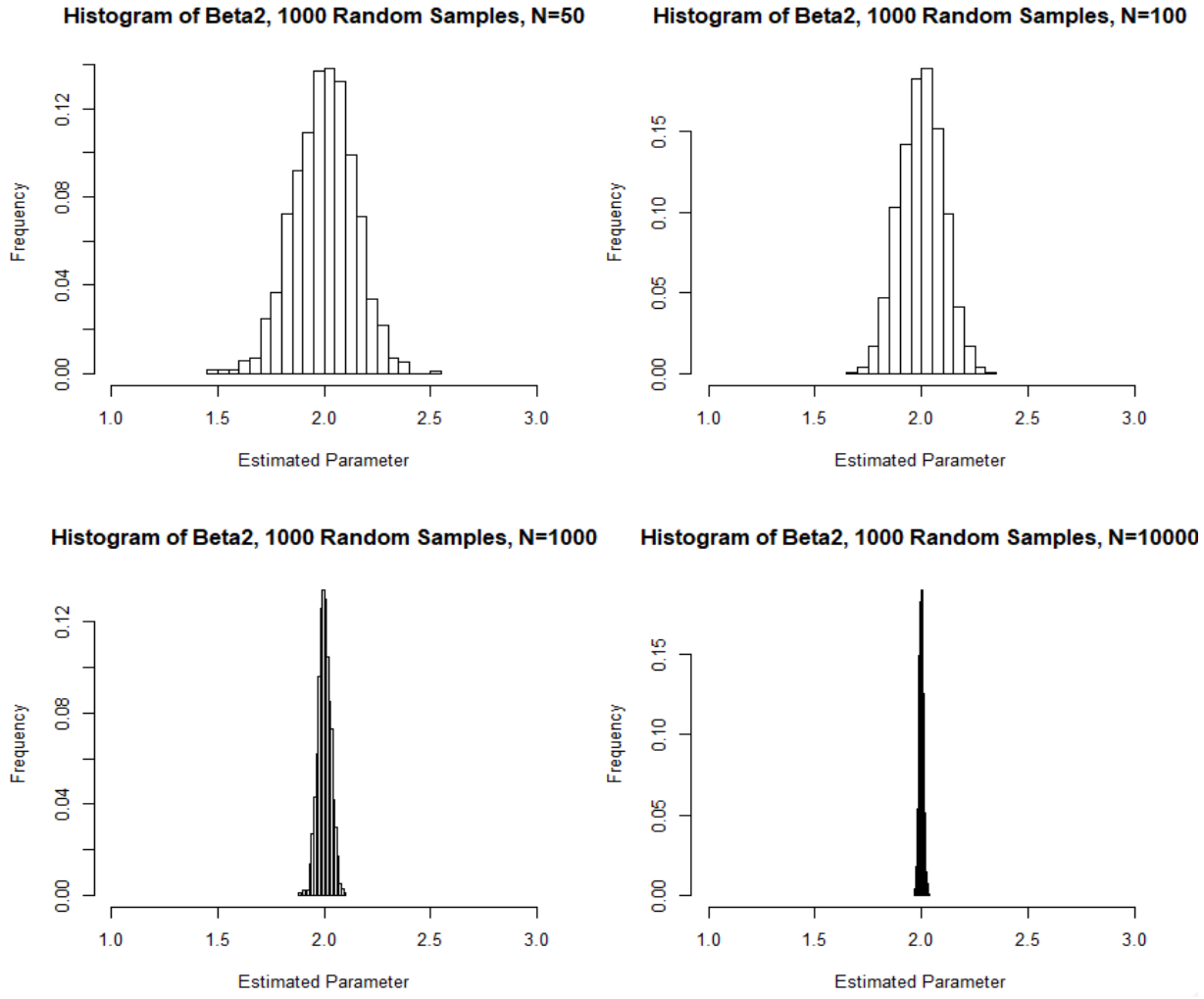

Figure 1. *Monte Carlo Simulation for Linear Model OLS*

From the result, it is observed that as the number of observations gets higher, the estimated parameter converged to the initialized value of the parameter. This shows that the given model is unbiased.

For the panel data model, even though there are many parameters that can be estimated, the main interest lies on β_2 which is the coefficient of x_{it}. It measures the slope of the graph and the relationship between x_{it} and y_{it}. The value of each parameters is initialized in the beginning of the simulation. The biasness of the parameter of interest is important here because it measures how good a method estimates the parameter. It is calculated by subtracting the estimated value of that parameter in each parameter with the initialized value. The average of those biases is then calculated to get the bias of each set-up. Besides, the root mean squared errors (RMSE) of the parameter of interest is also an important aspect to measure robustness of the method; when the value becomes too big, it shows that the method is not robust or not suitable for the given model. The squared errors are obtained by squaring the difference of the estimated value and the initialized value, to get the RMSE, take the square root of the averaged squared errors.

The errors, regressors, and loadings in this experiment are randomized using the Monte Carlo simulation, assuming the timeframe goes from -49 to 5. However, the resulting data from time -49 to 0 are omitted; the purpose is to generate large timeframe so that long panel assumption can be used. There are 50 and 100 samples in each simulation that goes over 1000 times replications, the biases and the MSE of each design are then compared. After conducting the experiment, the following result is obtained for OLS estimator:

Table 1. *Panel Data Monte Carlo Simulation with OLS result.*

Statistics	**N**	**T**	**Design I**	**Design II**	**Design III**
Bias	50	5	0.3139	0.2551	0.2880
RMSE	50	5	0.3344	0.2583	0.2907
Bias	100	5	0.3027	0.2537	0.2869
RMSE	100	5	0.3256	0.2569	0.2892

The result shows that the current model is biased. The pattern of the biases is not expected. As the model gets more complicated by introducing unobserved common regressors and individual loadings on variable x_{it}, from **Design I** to **Design II**, the biases decrease, but when it goes to the most complicated design (introducing correlated error terms), **Design III**, the biases in increase, even though not exceeding the biases in **Design I**. Overall, the values of the biases are reasonably high, and OLS might not be the best approach in this case.

The second estimator used in this experiment is fixed effects (FE) estimator. FE is an enhanced version of OLS, it takes account the individual heterogeneity. In most cases, every individual is assumed to have different complexity and ability in the topic of interest. For example, individual A is better than individual B in problem-solving skill, however, individual B is better than individual A in sports. So, if the type of study conducted is related to having a good problem-solving skill, the expected result of individual A is better than individual B. In the other hand, if the type of study conducted requires a good ability in sports, the expected result of individual B is better than individual A.

In this study, the FE method that is used is not the standard FE method, but the time-varying individual effects. This method was introduced by Bai and Pesaran, it is mainly used when dealing with panel data, because it measures individual heterogeneity in each given period. It takes account the possibility that each individual might have different ability in different period. For example, individual A at 22 years old (time 2) might have a better problem-solving skill compared to individual A at 21 years old (time 1) due to the knowledge that he gains during that one-year period. With including this time-varying individual heterogeneity, it is expected that FE would have less biases, because it takes account this individual effect where OLS does not. The result of this FE estimator is the following:

Table 2. *Panel Data Monte Carlo Simulation with FE result.*

Statistics	N	T	Design I	Design II	Design III
Bias	50	5	0.3297	0.2208	0.2544
RMSE	50	5	0.3494	0.2239	0.2550
Bias	100	5	0.3238	0.2211	0.2558
RMSE	100	5	0.3429	0.2238	0.2562

The result shows that the FE method is still biased. The pattern of the biases is the same with OLS. As the model gets more complicated by introducing unobserved common regressors and individual loadings on variable x_{it}, from **Design I** to **Design II**, the biases decrease, but when it goes to the most complicated design (introducing correlated error terms), **Design III**, the biases in increase, even though not exceeding the biases in **Design I**. Nevertheless, the values of the biases and RMSE meet the expectation since they decrease compared to the values from OLS. This means that by taking account of each time-varying individual heterogeneity, the values of the parameter of interest get closer to the true value.

The next method used is common correlated effects (CCE). This method introduced by Pesaran (2006) and uses the cross-sectional averages to account the unobserved factors in the model. It observes the cross-sectional dependence errors that are often assumed to be independent by earlier literature. The independence assumption works fine under short panel, but, panel literatures have shifted from short panel to long panel these days and the assumption can lead to serious errors. Since the model used in this experiment is long panel data, this method should be able to reduce the biases that might be existed in the previous models due to the independence assumption.

Table 3. *Panel Data Monte Carlo Simulation with CCE result.*

Statistics	N	T	Design I	Design II	Design III
Bias	50	5	0.0010	0.2508	0.3767
RMSE	50	5	0.0621	0.2725	0.3814
Bias	100	5	-0.0005	0.2508	0.3751
RMSE	100	5	0.0420	0.2693	0.3787

The result shows that under **Design I**, the CCE method is unbiased since the value of the bias is very close to 0. However, as the model becomes more complicated in **Design II** and **III**, the CCE method becomes biased. The pattern follows the expected result, because the method is supposed to be more biased under a more complicated model. Some of the added effects might not be taken account be the method that leads to the biased **Design II** and **III**. Overall, the result for **Design I** is the best compared to OLS and FE. However, it might not yield a better result when the problems are similar with **Design II** and **III**.

Lastly, the final method used in this paper is the combination of CCE and FE. The idea behind combining these two methods is to take account the time-varying individual heterogeneity that CCE does not take into account and it is shown from the FE result (**Table 2**) that the time-varying individual heterogeneity does decrease the biases and RMSE. Therefore, by including it to the method, the result should be better compared to the normal CCE.

The result below shows that under **Design I**, the model is still unbiased which agrees with the previous CCE result. The outcomes from **Design II** and **III** are also agree with the expected result, the pattern from CCE still holds. By introducing the FE in doing CCE, the method is less biased compared to CCE. The improvement in **Design II** is very significant, because the resulted bias is the least among the other methods. However, for **Design III**, the result improves from the

original CCE, but is still higher compared to FE alone. Overall, the result from combining CCE and FE is the best method in this paper.

Table 4. *Panel Data Monte Carlo Simulation with combining CCE and FE result.*

Statistics	**N**	**T**	**Design I**	**Design II**	**Design III**
Bias	50	5	0.0005	0.0790	0.2909
RMSE	50	5	0.0528	0.1036	0.2953
Bias	100	5	-0.0015	0.0780	0.2883
RMSE	100	5	0.0380	0.0957	0.2907

4. Empirical Application

The empirical application of this study is on minimum wage, especially in teen workers employment. This is important because teen workers tend to get the minimum wage salary in the society due to the lack of experience that they have compared to other types of workers. Depending on the value of the minimum wage, the employers might decide to hire teen workers or not. Therefore, teen workers employment is highly affected by the value of the minimum wage. This paper is using the same model as Neumark, Salas, and Wascher, which can be seen in **Equation (6)**. The idea of this empirical application is to test the consistency and the robustness of the results from Monte Carlo simulation. The results of the methods compared to the previous study are the following:

Table 5. *Empirical Application in Minimum Wage Compared to Previous Study.*

Statistics	OLS	FE	CCE	Previous Study (FE)
$\widehat{\beta_1}$	-0.437	-0.177	-0.147	-0.165
RMSE	0.017	0.081	0.065	0.041

The results seem robust, because using the same estimator, the values are not that different. The error of each estimator is also reasonably small. Since all the estimators are applicable, the results from Monte Carlo simulation are used to determine the best estimator and most appropriate coefficient. Since CCE seems to have the lowest bias for long panel data, it would be the best estimator for this minimum wage model. However, the consistency and the robustness over different clusters of the datasets also need to be analyzed to confirm the finding.

Table 6. *Clustering the Data to Check Consistency and Robustness.*

Cluster	Statistics	OLS	FE	CCE
Full	$\widehat{\beta_1}$	-0.437	-0.177	-0.147
	RMSE	0.017	0.081	0.065
First Half	$\widehat{\beta_1}$	-0.471	-0.143	-0.170
	RMSE	0.037	0.069	0.185
Second Half	$\widehat{\beta_1}$	-0.140	-0.132	-0.093
	RMSE	0.037	0.050	0.087
First Quarter	$\widehat{\beta_1}$	-0.010	-0.149	-0.228
	RMSE	0.084	0.100	0.415
Second Quarter	$\widehat{\beta_1}$	-0.553	-0.039	-0.303
	RMSE	0.060	0.113	0.197
Third Quarter	$\widehat{\beta_1}$	-0.049	-0.036	-0.052
	RMSE	0.059	0.096	0.287
Fourth Quarter	$\widehat{\beta_1}$	-0.028	-0.066	-0.161
	RMSE	0.070	0.070	0.140

The results show that OLS and FE are robust and consistent for all clusters, because the coefficients are always negatives and the errors are relatively small. CCE is robust and consistent for the full dataset and for half clusters. However, when it is analyzed quarterly, the RMSE starting to grow which shows the non-robustness of the result. This finding leads to the idea that OLS and FE works under any sizes of the datasets, but CCE might only work under big datasets.

5. Discussion

From the Monte Carlo simulation, it was found that the OLS, FE, and CCE are all biased. OLS seems to have the highest consistent increasing bias among the three (**Table 1**). There are some possible explanations why OLS performs the worst here. First, OLS relies heavily on the Gauss-Markov Theorem that say it would be the Best Linear Unbiased Estimator (BLUE) if the linear regression assumptions are hold. Unfortunately, that is not the case here; most of the linear regression assumptions are violated in the model and makes OLS biased. Next, OLS might not have the ability to taking account of some elements in the model due to its complexity.

The Fixed Effects (FE) estimator performs better than OLS (decreases after Model I), but not as good as CCE. It works better compared to OLS because FE takes account of individual time-varying effects. Hence, FE can observe some effects that might not be observed by OLS. However, it does not mean that this observed effects always have a positive result on the biasness of the estimator. Sometimes it can give a higher bias compared to OLS, for example when the simulation is applied on **Design I**. Overall, FE has a decreasing bias as the model gets more complicated, but the biases are still relatively high (**Table 2**).

The Common Correlated Effects (CCE) estimator performs the best compared to OLS and FE, especially in **Design I**. The biases for this specific case are almost 0. However, there are some weaknesses of this model. As the unobserved common regressors and individual loadings are introduced in the variable x for **Design II**, the biases go up significantly. Furthermore, as the error terms are made to be correlated to each other in **Design III**, the biases go up again. In both cases, the biases exceed OLS and FE (**Table 3**). These observations lead to a hypothesis that CCE does not work that well for more complicated model.

As each of the estimators above are used to estimate the coefficient for minimum wage on teen workers' employment, all of them give a robust estimate (**Table 5**). The previous study by Neuman, Salas, and Wascher uses FE as their method in analyzing the data. The FE result in this study is reasonable close to the past study. In choosing the best estimator for the minimum wage problem, the result that is the most analogous to the model is **Design I** from the Monte Carlo simulation. The main reason is that the model is relatively simple, and **Design I** has the simplest model in the simulation. Therefore, the most appropriate method based on the result is CCE, which has the coefficient of -0.147 and error of 0.065. The result is not that different compared to the both FE models, but the OLS result deviates a lot. This follows the pattern of the Monte Carlo Design I simulation where OLS performs the worst, followed by FE and CCE.

Furthermore, the methods above are also applied to the clusters of the datasets. The purpose of this study is to show that the findings are robust and consistent by showing same pattern and small errors. From **Table 6**, as the dataset is clustered into half, all the estimators give the same pattern. The effect is still negative, and the first half coefficient is high than the second half. This indicates that the effect of minimum wage on teen workers' employment decreases from first half to the second half. The finding on these clusters are robust and consistent since the errors are also relatively low.

In the other hand, as the dataset is clustered into quarters, the findings are not that useful. The only consistent thing is that the effects are all negatives. Besides, the patterns as it moves from first quarter to second quarter, second quarter to third quarter, and third quarter to fourth quarter are not consistent. Lastly, the errors all the estimators are also relatively high compared to the previous results. These inconsistency and non-robustness might be caused by the size of the

datasets, because as it got clustered into quarters, the time becoming smaller in addition to the data is not big enough for the result to converge.

Overall, this empirical application shows a good illustration that some methods work better than the others in conducting long panel data analysis. All methods and approaches are robust and consistent given the size of the dataset is big enough (full dataset and clustered into half in this case). CCE performs the best as it agrees with the Monte Carlo simulation **Design I**. The study in minimum wage also supports the result from previous study by Neumark, Salas, and Wascher that minimum wage has a negative correlation with teen workers' employment. This shows that teen workers are being exploited if the minimum wage is low and they are not being hired if the minimum wage is high. Furthermore, this study also finds out that the effect has been getting milder from the first half of the dataset to the second half of the dataset.

6. Appendix: Data

The main dataset in this study is CPS state-level data from first quarter 1990 to second quarter 2011. In total, there are 86-time periods on 51 different states giving the total 4386 observations for each variable. The dataset has 8 variables that consist of census division, state code, date (timeframe), teen employment, minimum wage, overall unemployment rate, ratio of teen populations, and total teen populations. The size of the dataset is sufficient to be categorized as long panel data and hence can be used in this study.

Acknowledgements

I am very grateful to Professor Matthew Harding, my honor advisor, who has introduced me to this study and gives me the supplementary materials for the researches. I also thank Professor Fabio Milani and Professor Gary Richardson for giving the comments and suggestions in addition to providing samples from the previous thesis and meeting room to do presentations and interact with my fellow students. Last, but not least, I also thank my classmates, who paid attention to my presentations and gave helpful feedbacks throughout the quarter.

References

Angrist, Joshua D., and Allan B. Krueger. “Empirical strategies in labor economics.” Handbook of Labor Economics 3 (1999): 1277-1366.

Bai, Jushan. “Panel data models with interactive fixed effects.” Econometrica 77.4 (2009): 1229-1279.

Harding, Matthew, and Carlos Lamarche. “Least squares estimation of a panel data model with multifactor error structure and endogenous covariates.” Economics Letters 111 (2011): 197-199.

Harding, Matthew, and Carlos Lamarche. “Estimating and testing a quantile regression model with interactive effects.” Journal of Econometrics 178 (2014): 101-113.

Neumark, David, Salas, J. M. Ian, Wascher, William. 2014a. Revisiting the minimum wage and employment debate: Throwing out the baby with the bathwater? ILR Review 67(Supplement): 608–48.

Pesaran, M. Hashem. “Estimation and inference in large heterogeneous panels with a multifactor error structure.” Econometrica 74.4 (2006): 967-1012.