

Course Form for PKU Summer School International 2025

Course Title	Title in English: Applied Econometrics
	Title in Chinese: 应用经济计量
Teacher	Dr. Xuezheng Qin, Professor of Economics, Peking University
First day of classes	July 14, 2025
Last day of classes	July 25 2025
Course Credit	2 credits
Course Description	
Objective:	
<p>This course provides students with a comprehensive understanding of the econometric modeling tools that are frequently used in empirical economic research. The topics include linear and non-linear regressions, least squares and maximum likelihood estimation, time series and panel data models, instrumental variable estimation and regression equation systems, models of limited dependent variables, research design and program evaluation methods, etc. The computer programming techniques to implement these models will be taught using the STATA software. To exemplify the above estimation methods in real research settings, a number of published papers are selected into the reading list and will be discussed in the class. In addition, you will get hands-on experience in conducting empirical research by writing and presenting an academic term paper.</p>	
Pre-requisites /Target audience	
<p>This course is intended for the upper undergraduate students in Economics, Business or other social science majors. Graduate students are also welcome. Prior training in Introductory Econometrics or Statistics is required. The main focus of the course is to use econometric tools to solve real-world problems, and thus we will not spend much time on the mathematical derivation of basic models.</p>	
Proceeding of the Course	
<p>The course will be delivered through a mix of lectures, student presentations and computer lab exercises. Students' participation is strongly encouraged.</p>	
Assignments (essay or other forms)	
<p>Reading assignments and computer lab assignments will be given for each lecture. At the end of the course, students will be required to submit an original research paper using econometric models and Stata programs.</p>	
Evaluation Details	

Class Attendance and Participation: 20%
 Group Presentation on Reading Assignments: 10%
 Computer Lab Assignments: 10%
 Final Term Paper: 60%

Text Books and Reading Materials

- 1) *Microeconometrics: Methods and Applications*, by A. Colin Cameron and Pravin K. Trivedi, Cambridge University Press, 2005.
- 2) *Microeconometrics Using Stata (2nd Edition)*, by A. Colin Cameron and Pravin K. Trivedi, Stata Press, 2010.

Academic Integrity (If necessary)

Plagiarism is strictly prohibited in this course. Plagiarism is defined as presenting others' work as one's own. Any ideas or materials taken from another source must be fully acknowledged, unless the information is common knowledge. Any evidence of plagiarism will lead to zero credit to your work submitted and will be reported to university authorities.

CLASS SCHEDULE

(Subject to adjustment)

Session 1: *Introduction*

Date: 7.14

【Description of the Session】

This session provides an overview of the econometric modelling approaches in modern empirical economic research. It starts with a review of the mathematical knowledge (matrix algebra, statistics, etc.) that is needed in deriving the basic econometric models. Then we will discuss the major steps in conducting applied econometric research, the types of data to be encountered, the mainstream modelling approaches and the computer software that are commonly used in empirical research. In the computer lab hours, students will be introduced to the basic programming techniques in Stata.

【Questions】

What are the main steps in conducting applied econometric research? How should we approach an economic problem and write an empirical research paper? What types of data / models / software are frequently used in applied econometrics? What is the difference between parametric and non-parametric models? What are the differences between least squares estimation, maximum likelihood estimation, and method of moments estimation? How to read in a dataset, create variables, plot and tabulate variable values and perform basic statistical regressions in Stata?

【Readings, Websites or Video Clips】

Reading materials will be handed out in the class or uploaded to the course website.

【Assignments for this session (if any)】	
Students are expected to familiarize themselves with the materials covered in the lectures and to present the academic papers in the reading assignments in groups. In addition, computer lab assignments should be completed in the TA recitation session.	
Session 2: <i>Ordinary Least Squares</i>	Date: 7.15
【Description of the Session】	
This session provides an overview of the Ordinary Least Squares (OLS) estimation methods. We will first discuss the purpose of statistical regression and the available regression techniques. Then we will introduce how OLS is carried out in univariate and multivariate settings. Special attention is paid to the following issues within OLS: functional form specifications, dummy variables, time trend and seasonality, model misspecifications and how to avoid them, and lastly, model selection and goodness of fit measures. In the computer lab hours, students will learn to conduct OLS regressions and perform relevant econometric analysis using real world data and Stata software.	
【Questions】	
What are the purposes of statistical regression? What assumptions do we need to make when perform OLS regressions? How to interpret the estimation results when 1) dummy variables are involved; 2) variables are transformed by logarithm; 3) certain variables are omitted from the regressions? How to control for the time trend and seasonality in the data? What are the commonly used goodness of fit measures for model selection?	
【Readings, Websites or Video Clips】	
Reading materials will be handed out in the class or uploaded to the course website.	
【Assignments for this session (if any)】	
Students are expected to familiarize themselves with the materials covered in the lectures and to present the academic papers in the reading assignments in groups. In addition, computer lab assignments should be completed in the TA recitation session.	
Session 3: <i>Generalized Least Squares</i>	Date: 7.16
【Description of the Session】	
This session provides an overview of the Generalized Least Squares (GLS) estimation methods. We will first discuss the definition and implication of non-spherical disturbance, then we will introduce GLS as a solution to the above problem. To exemplify the GLS estimation method, we will look at three specific cases of non-spherical disturbance: (1) heteroscedasticity, (2) serial correlation and (3) spatial dependence; in each cases, we will discuss the diagnostic tests and the Feasible GLS procedures involved in the estimation. In the computer lab hours, students will learn to conduct GLS regressions in the above and more general cases using real world data and Stata software.	

【Questions】

What is non-spherical disturbance, and what impacts does it have on OLS estimation? Why can GLS correct the misspecification of standard errors in presence of non-spherical disturbance? What is the FGLS procedure, and how is it carried out in the cases of 1) heteroscedasticity, 2) serial correlation and 3) spatial dependence? What are the differences between the Breusch-Pagan test and the White test in detecting heteroscedasticity? What are the key differences between the Spatial Lagged Model and the Spatial Error Model in addressing the spatial dependence in geographic data?

【Readings, Websites or Video Clips】

Reading materials will be handed out in the class or uploaded to the course website.

【Assignments for this session (if any)】

Students are expected to familiarize themselves with the materials covered in the lectures and to present the academic papers in the reading assignments in groups. In addition, computer lab assignments should be completed in the TA recitation session.

Session 4: *Maximum Likelihood Estimation and Hypothesis Testing*

Date: 7.17

【Description of the Session】

This session provides an overview of the Maximum Likelihood Estimation (MLE) and the hypothesis testing techniques in applied econometrics. It starts with an introduction to the Maximum Likelihood principle, and continues with a detailed derivation of the MLE procedure and its statistical properties. Then we will compare MLE with other non-linear estimators (such as the Non-linear Least Squares estimator) with respect to their estimation efficiency and sensitivity to error distribution misspecification. In the end, we will discuss hypothesis testing techniques in the context of linear and nonlinear models, as well as linear and nonlinear parameter restrictions. In particular, we will focus on the three MLE-based tests: the Likelihood Ratio test, the Wald test, and the Lagrange Multiplier test. In the computer lab hours, students will learn to conduct MLE regressions and statistical tests using real world data and Stata software.

【Questions】

What is the Maximum Likelihood principle, and how to perform MLE based on given distributional assumptions? What are the key differences between MLE and NLS estimator? How to perform hypothesis testing that involves single / multiple parameter restriction(s)? What are the key differences between the three MLE-based tests: the Likelihood Ratio test, the Wald test, and the Lagrange Multiplier test?

【Readings, Websites or Video Clips】

Reading materials will be handed out in the class or uploaded to the course website.

<p>【Assignments for this session (if any)】</p> <p>Students are expected to familiarize themselves with the materials covered in the lectures and to present the academic papers in the reading assignments in groups. In addition, computer lab assignments should be completed in the TA recitation session.</p>	
Session 5: <i>Time Series Models</i>	Date: 7.18
<p>【Description of the Session】</p> <p>This session provides an overview of the Time Series Models. It starts with an introduction of stochastic processes and the two basic structure in the time series data, i.e. the Autoregressive (AR) structure and the Moving Average (MA) structure. We will then discuss the definition and tests of stationarity and show how a stationary time series process can be decomposed into AR and MA processes. We will then formally introduce the Box-Jenkins approach to the estimation of an ARMA or ARIMA model. The session ends with a discussion of several special topics in time series modelling, i.e. the multivariate structure and the Vector Autoregressive (VAR) models, the stochastic volatility and the (Generalized) Autoregressive Conditionally Heteroskedastic models. In the computer lab hours, students will perform various time series analysis and forecasting using real world data and Stata software.</p>	
<p>【Questions】</p> <p>What is the key feature of time series data and why OLS typically breaks down in time series analysis? What are the key differences between AR and MA processes, and how to transform an AR process into an MA process? How to perform the unit root test for stationarity using (Augmented) Dickey-Fuller test / Phillips-Perron test? What are the four major steps in the Box-Jenkins approach in estimating an ARIMA model? How to perform the Granger Causality test and Impulse Response Analysis in the VAR model? What are the implications of stochastic volatility, and how to use the ARCH model to address this issue?</p>	
<p>【Readings, Websites or Video Clips】</p> <p>Reading materials will be handed out in the class or uploaded to the course website.</p>	
<p>【Assignments for this session (if any)】</p> <p>Students are expected to familiarize themselves with the materials covered in the lectures and to present the academic papers in the reading assignments in groups. In addition, computer lab assignments should be completed in the TA recitation session.</p>	
Session 6: <i>Models of Pooled Cross-sections and Panel Data</i>	Date: 7.21
<p>【Description of the Session】</p> <p>This session provides an overview of the econometric models of pooled cross-sections and panel data. We will first introduce and compare the features of the two data types, then we will discuss in details the Difference-in-Difference (DID) model, the Fixed Effects (FE) model, and the Random Effects (RE) model. Special attention will be paid to the underlying assumptions of each model and how to select the appropriate model based on statistical tests</p>	

and economic intuition. In the computer lab hours, students will learn to conduct various analysis on the pooled cross-sectional and panel data using the Stata software.

【Questions】

What are the differences between the pooled cross-sectional data and panel data? How do we decide whether to pool independent cross-sections together? How to implement the DID model using a step-wise approach / pooled regression approach? Why is the parallel trend assumption crucial for DID analysis? What are the differences between the underlying assumptions of the FE and RE models? How to implement the FE model using the time-demeaned data transformation and the Least Squares Dummy Variable (LSDV) method? How to use the Hausman's specification test to select between FE and RE models?

【Readings, Websites or Video Clips】

Reading materials will be handed out in the class or uploaded to the course website.

【Assignments for this session (if any)】

Students are expected to familiarize themselves with the materials covered in the lectures and to present the academic papers in the reading assignments in groups. In addition, computer lab assignments should be completed in the TA recitation session.

Session 7: *Endogeneity and IV Estimation*

Date: 7.22

【Description of the Session】

This session provides an overview of the endogeneity issue in econometric analysis and the Instrumental Variable (IV) estimation method. We will first look at the sources and consequences of endogeneity, and see how an IV may address these problems. We then discuss how the IV estimation can be implemented using the projection matrix approach and the two-step least squares (2SLS) approach. Several statistical tests associated with the IV model will then be studied, e.g. the Hausman test for exogeneity and the Sargan-Basmann Test for over-identification restrictions. We will then briefly discuss several special IVs, such as binary IVs, multiple IVs, and weak IVs. In the computer lab hours, students will learn to implement the IV models using real world data and the Stata software.

【Questions】

What are the sources and consequences of endogeneity problem in OLS regression? What are the required features of a valid IV, and why may a valid IV be used to address the endogeneity problem? How to conduct the IV estimation using the project matrix / 2SLS method? How to perform Hausman test for the exogeneity of an explanatory variable? What is the purpose and operational procedure of the Sargan-Basmann Test? What implications does a weak IV have on 2SLS results, and how to test for the weakness of an IV?

【Readings, Websites or Video Clips】

Reading materials will be handed out in the class or uploaded to the course website.

【Assignments for this session (if any)】

Students are expected to familiarize themselves with the materials covered in the lectures and to present the academic papers in the reading assignments in groups. In addition, computer lab assignments should be completed in the TA recitation session.

Session 8: *System of Regression Equations*

Date: 7.23

【Description of the Session】

This session provides an overview of the models of regression equation systems. We first discuss the Seemingly Unrelated Regressions (SUR) and how it can be used to control for the contemporaneous correlation between the error terms of different equations. Then we will focus on the Simultaneous Equations Model (SEM) and the available estimation strategies for such models, including the equation-by-equation methods (e.g. 2SLS, LIML) and the system estimation methods (e.g. 3SLS, FIML). Order and rank conditions for the identification of SEM will also be discussed. In the computer lab hours, students will learn how to deal with a system of regression equations using real world data and Stata software.

【Questions】

What is contemporaneous correlation, and why is the SUR model able to address this issue? What are the differences between SUR and SEM? How to use the 2SLS / 3SLS method to estimate an SEM? What are the difference between the equation-by-equation methods (such as LIML) and the system estimation methods (such as FIML)? How to use the order and rank conditions to decide whether an equation in an SEM is identifiable?

【Readings, Websites or Video Clips】

Reading materials will be handed out in the class or uploaded to the course website.

【Assignments for this session (if any)】

Students are expected to familiarize themselves with the materials covered in the lectures and to present the academic papers in the reading assignments in groups. In addition, computer lab assignments should be completed in the TA recitation session.

Session 9: *Discrete Choice Models*

Date: 7.24

【Description of the Session】

This session provides an overview of the discrete choice models. It starts with a discussion of the binary choice models, and introduces the three main estimation methods for such models, i.e. the Linear Probability Model (LPM), the Probit model and the Logit model. Special attention is paid to the limitation of LPM and how the non-linear models may improve on these limitations. ML estimation procedure on the Probit and Logit models are then shown, so is the partial effect calculation method for such models. We then generalize the binary choice model to multinomial models, and discuss two popular models

in this category: the Multinomial Logit (MNL) Model and the Conditional Logit (CL) Model. Comparison and contrast will be made between MNL and CL, with special attention paid to their assumptions of independence of irrelevant alternatives (IIA), which leads to our discussion on the Nested Logit Model (N-logit) and ordered multinomial models. In the computer lab hours, students will learn to carry out the analysis on various types of discrete choice models using real world data and the Stata software.

【Questions】

What are the main limitations of LPM, and why Probit and Logit models can avoid such problems? According to the latent variable approach, how is the probability of the binary dependent variable determined under the Probit / Logit model? How to carry out the ML estimation and the partial effect calculation for the Probit / Logit model? What are the key differences between the MNL and CL models? What is the IIA assumption, and what implications does it have on the MNL and CL predictions? How would the N-logit model and the ordered multinomial models address the IIA problem?

【Readings, Websites or Video Clips】

Reading materials will be handed out in the class or uploaded to the course website.

【Assignments for this session (if any)】

Students are expected to familiarize themselves with the materials covered in the lectures and to present the academic papers in the reading assignments in groups. In addition, computer lab assignments should be completed in the TA recitation session.

Session 10: *Tobit and Selection Models*

Date: 7.25

【Description of the Session】

This session provides an overview of the censored and truncated regression models as well as the various types of sample-selection models. We first introduce the definition of data censoring and truncation, and discuss the sources and implications of such problems. Then we will look at the Tobit model that address the corner solution (or zero censoring) problem, which is followed by a more general treatment of the censored and truncated regression models. In the second part of the session, we will discuss the other types of sample selection models, including the two-part model (2PM), the Heckman selection model, and the Roy model. In the computer lab hours, students will have an opportunity to try the above models using real world data and the Stata software.

【Questions】

What are the differences between the censored and truncated regressions? How to carry out the ML estimation and the partial effect calculation for the Tobit model and the truncated regression model? What are the differences between 2PM, Heckman selection model and Roy model in addressing the sample selection issue? How to estimate the Heckman / Roy model using MLE and the stepwise regression (Heckit) approach?

【Readings, Websites or Video Clips】

Reading materials will be handed out in the class or uploaded to the course website.

【Assignments for this session (if any)】

Students are expected to familiarize themselves with the materials covered in the lectures and to present the academic papers in the reading assignments in groups. In addition, computer lab assignments should be completed in the TA recitation session.

A CV of 250-300 words and a high-resolution personal photo should also be provided