ECONOMETRICS AND DATA SCIENCE
Intuition, Theory and Applications

Instructor: Duncan Thomas
Office: 314 Social Sciences
Email: dthomas@econ.duke.edu
Lectures: Monday and Wednesday 8.30-9.45am, Old Chemistry 116
Office hours: Monday 4:30-6:30 pm or by appointment
Class web site: http://ipl.econ.duke.edu/dthomas/ec204 (or log on through sakai)

Graduate teaching assistants

Name   Email
Rosie Chiang  Nienjune.Chiang@duke.edu
Damien Kim  Dongyoung.Kim@duke.edu
John Shin  Hyunmoh.Shin@duke.edu

Sections   Day/time   Place
01  Wednesday 6:15-7:05 pm  Allen 326
02  Thursday 6:15-7:05 pm  Allen 326

Attendance at lectures is required. Attendance at sections is very strongly recommended. Sections are led by outstanding PhD students who have considerable expertise in applying the econometric methods taught in this class to data and know the material extremely well. In addition to sections, the TAs will hold office hours [TBA].

Undergraduate teaching assistants

The class is also supported by two excellent undergraduate teaching assistants, Hannah O’Sullivan (hannah.o.sullivan@duke.edu) and Tony Wang (yi.jia.wang@duke.edu) who are available to assist you with material from the course and problem sets. They will hold office hours [TBA].

Examinations

There will be two examinations: a midterm and a final. Both examinations are required.

**Midterm:** Monday September 28, 8:30-9:45am (Old Chem 116 and TBA)
The in-class mid-term will be open book. You may bring any materials you need to the exam. You may not communicate with anyone inside or outside the room during the examination.

**Final:** Wednesday December 9, 2:00pm-5:00pm (TBA)
The final will be closed book. You may bring *one two-sided* 8.5”*11” page of formulae and notes that you construct and that you think will be of use to you during the exam. The sheet must be handed in with your exam. The exam will not include questions that require the use of STATA. Please bring a calculator to the exam.

(c) Duncan Thomas 2020
If you have substantive questions about the course material you want to ask me, please see me after class or during my office hours. Emailed questions about the material are very hard for me to answer effectively. It has been my experience that I typically need to know more about the problems you are encountering than you tell me in your email. In addition, I have no way to assess whether my e-mail response has cleared up the problem for you. I will, therefore, not attempt to answer substantive questions sent to me by email. If you cannot attend my office hours, we can make an appointment to meet at a time that is convenient for you.

Course Objectives

This is the second course in the integrated two-course sequence in econometrics and data sciences. The goal of this course is to provide an introduction to both the theory and application of modern econometric and data science methods to address questions in economics, the behavioral and health sciences. Mastery of the material is intended to provide you with sufficient knowledge of econometric and data science methods to enable you to think critically when you evaluate the quality of evidence in support of a claim about how individuals behave, markets work, firms make money, organizations operate or societies co-ordinate. You will develop the skills to be an effective consumer and producer of empirical research in economics and related sciences. Throughout the class, emphasis is placed on intuitive understanding of underlying concepts with more rigorous arguments serving to strengthen the foundation of your knowledge. Central concepts are illustrated with applications.

Course Requirements

You are required to have passed Probability and Statistical Inference (Stat 111) or Econ 104 (Statistical Foundations for Economics and Data Sciences). This class builds on the foundation laid in that class. Ideally you will have taken the foundational class in the semester immediately preceding this class or at least within the twelve months preceding enrolling in this class. It will be assumed that you have a good understanding of the material covered in that class. With my permission, you may substitute Stat 130, 203, 250 Math 230 or 342 for the pre-requisite. In my experience, those classes cover somewhat different material from the pre-requisite; it is incumbent upon you to be sure you have all of the material in the pre-requisite under your belt. In addition, it is assumed that you have a good grasp of calculus.

Building on the foundation laid in the pre-requisite classes, this class will begin with an introduction to the linear regression model. Core statistical concepts that you have covered in your statistics preparation will be applied to the regression model to provide a fuller understanding of the value of these tools to better understand economic phenomena and the world around us. This class will evaluate extensions to the linear regression model that are designed to address real world problems that arise in the study of economic behavior and provide rigorous tests of hypotheses in the economics and related literatures. In addition to understanding the theoretical concepts that underlie modern regression analysis, you should develop the practical skills necessary for good data analysis as well as learn how to interpret the results of your analyses. You will be required to do econometric analysis with real data from actual applications. You may use whatever computer hardware and software you like. Instruction will be provided for using STATA which you are strongly encouraged to learn. The product will serve you well in this class and other classes, research at Duke and beyond Duke. We will use STATA in this class. You may download a copy of STATA/SE for your own Mac, Windows or Linux computers. Instructions are on the class web site, http://ipl.econ.duke.edu/dthomas/ec208d/statalinks.html.

In addition, help using STATA is available in the Data and GIS lab in Perkins and at SSRI West. There are many good on-line resources, some of which are linked from the class web page.
Economics 208D

**Grading and Organization**

Each week there will be two lectures. You are **required** to attend every lecture. Please turn off all phones, tablets, laptops etc. before class starts and do not turn them back on until the class has ended; they may not be used during class. Lectures will be supplemented by a weekly section led by an experienced Ph.D. students. Sections will cover statistical material not covered in this class that was covered in the pre-requisite class. You will be responsible for that material. Sections will also review lecture material, problem sets and provide computer instruction. Each week, one TA will be responsible for all sections and the same material will be covered in every section that week. You may attend the section that is more convenient for you. You are expected to understand and uphold the Honor Code, [http://studentaffairs.duke.edu/conduct/about-us/duke-community-standard](http://studentaffairs.duke.edu/conduct/about-us/duke-community-standard).

**Problem sets**

There will be a problem set every second week. The problem are designed to help you understand important ideas in the theory and application of econometric methods and contribute to your developing a sophisticated understanding of how to interpret empirical evidence in a practical, real-world setting. To underscore the importance of taking the problem sets seriously, they will account for 20% of the final course grade. Completed problem sets must be submitted in person at the beginning of the lecture on the due date. Late problem sets will not be accepted. If you cannot hand in your problem set at the beginning of the lecture, with my prior consent, you can hand it in earlier. Problem sets must be handed in on paper, preferably typewritten. Emailed problem sets will not be accepted.

You may work with other students in this class on problem sets. By collaborating with your peers, my expectation is that you will better learn the material. Explaining an idea, concept, method or result to a peer is one of the best ways to reinforce your own understanding. However, to be sure that you do understand the work, **you must write up your own answers and submit your own work.** Please indicate on your problem set answers with whom you collaborated. You may only refer to the answer keys provided for this year’s problem sets which are posted on the class web page after the problem set is due. Prior year answer keys may not be used in any way or under any circumstances. A grade of zero will be assigned to all problem sets for the course for any student who violates this policy for any problem set.

For each problem set, the grades will be: 3 if you do an absolutely stellar job and your answers are at least as good as the answer key; 2 if, roughly speaking, you get at least 2/3rd’s of the problem set correct; 1 if you get between 1/3rd and 2/3rd’s correct; and 0 otherwise. Answers to questions that call for intuition or interpretation are given more weight than calculations. A grade of 1 is a signal to you that you need help. A grade of 0 indicates a very serious problem and you should seek help from the TAs or me immediately. Problem sets handed in late will get a grade of 0. The final problem set grade will be based on the total score over all problem sets. I cannot overstate the value to you of making a good faith attempt to complete every problem set and submit it on time. Each problem set will be graded by one TA. If you have any questions, that TA will be the best person to review your answers after you have submitted your problem set. All questions about grades should be taken up with the TA who graded that problem set.

Answer keys for each problem set will be posted on the class webpage soon after the submission deadline has passed. Each problem set will be reviewed in section. The section will not only go over the answers but also provide insights into how to think about the problems and thereby strengthen your understanding of the material covered in class. Please carefully review the posted answer keys. There are three reasons this will be a good use of your time. First, it is not possible to provide detailed written individual-specific feedback for each student on his/her answer. Second, it is very difficult for us to identify every instance in which you do not understand something; it is your
Responsibility to make sure you understand all the material covered in each problem as laid out in the answer key. Third, the answer key is designed to provide a succinct, clear answer to each question; following the model used in the answer key will stand you in good stead in this class and beyond.

**In-class quizzes**

In-class quizzes will be administered to those attending lectures at randomly assigned times during the term. The quizzes are intended to assess your understanding of concepts covered in lectures up to that point and to provide feedback to you regarding your grasp of the material covered in the class. The in-class quizzes will account for 15% of the final grade. If you are absent from class, no matter what the reason, your grade for that in-class quiz will be zero.

**Weekly discussion section**

Each weekly discussion section will be led by a TA who is a PhD student. The TAs are experienced teachers who have expertise in econometrics. Sections will provide instruction in STATA, including classes on good programming and data management practices that will be helpful for other classes. Sections will review problem sets and extend ideas covered in the problem sets and review and reinforce material covered in this class or Stats 111. The sections in any week are designed to be identical; you are welcome to attend the section that fits your schedule best in any week.

**Mid-term exam**

There will be one mid-term which will account for 15% of your course grade. The mid-term will be an open book examination. You may bring any books or materials you want to the class. You may not communicate with anyone inside or outside the classroom during the exam; this includes texting, emailing or any other form of communication. There will be no make-up midterm. If you are unable to take the midterm, you must provide me with a written explanation before the mid-term. If the written explanation provides a reason that is clearly beyond your control, and I judge that it is appropriate to do so, then I will substitute your final grade for your midterm grade. In any other instances, your midterm grade will be zero.

**Final exam**

The final exam will be closed book and will cover all the material in the course. You do not need to memorize formulae for this class. You may bring one 8½*11 page of formulae with you for reference during the final exam; you may use both sides of the page. No other reference material is allowed. You may not communicate with anyone inside or outside the classroom during the exam; this includes texting, emailing or any other form of electronic communication. The final exam score will contribute 50% of the final grade for the course. If you miss the final exam for a reason that is outside of your control, with the approval of the Dean and if I judge that it is appropriate, you may be able to take the final exam the next time it is offered. In that case, I will substitute your grade in that exam for the final exam grade for this class after adjusting the grade so that the mean grade in both exams is the same. The date and time of the final exam is determined by the Registrar and it cannot be rescheduled.

**Sharing class materials**

All the material you need for this course will be available on the class web page. The material is copyrighted which means you may not re-distribute any material such as hand outs, slides, problem sets, exams or answer keys that we have created. Re-distributing includes posting material on a website, server, shared drive, file host or similar service or providing material to someone who is not enrolled in the class this term. Please do not distribute any materials from this class. Since I change the class each year, I strongly discourage you from attempting to use material from prior classes; in my experience, that results in confusion and disappointment.

*(c) Duncan Thomas 2020*
Reading

The recommended text for this class is


Access to this book is encouraged. You may borrow the book from the library for 3 hours at a time (see [https://library.duke.edu/course-support/course-reserves/textbooks](https://library.duke.edu/course-support/course-reserves/textbooks)). If you purchase or rent the book, you may buy the paper or electronic version; you do not need access to the on-line resources and an earlier edition of the book will work well. If you find you do not like the presentation in Wooldridge's book, I encourage you to look at one or more of the following:


None of these books covers statistical theory in a comprehensive manner. If during the course you feel you need a statistical reference, use the text from your statistics course or look at:


I will provide handouts throughout the course to supplement the lectures and textbook. There is no shortage of alternative text books and there is a large amount of material on-line that will supplement this course. The class web-site has links to some of the on-line material.

Course Outline and Required Reading

Readings from Wooldridge are intended to complement the lectures and I encourage you to read the appropriate chapters before class.

<table>
<thead>
<tr>
<th></th>
<th>Wooldridge</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Introduction to econometrics</td>
<td>Chapters 1 and 19</td>
</tr>
<tr>
<td>2.</td>
<td>Simple linear regression model</td>
<td>Chapter 2, 9.5, 9.6</td>
</tr>
<tr>
<td>4.</td>
<td>Theory of estimation and inference</td>
<td>Appendices B and C</td>
</tr>
<tr>
<td>5.</td>
<td>Linear regression model: Inference</td>
<td>Chapter 4, 5</td>
</tr>
<tr>
<td>6.</td>
<td>Linear regression model: Interpretation</td>
<td>Chapter 6</td>
</tr>
<tr>
<td>7.</td>
<td>Indicator variables</td>
<td>Chapter 7</td>
</tr>
<tr>
<td>8.</td>
<td>Non-spherical errors: Heteroskedasticity and correlated errors</td>
<td>Chapter 8</td>
</tr>
<tr>
<td>9.</td>
<td>Limited dependent variable models</td>
<td>Chapter 17</td>
</tr>
<tr>
<td>10.</td>
<td>Omitted variables and sample selectivity</td>
<td>Chapter 9</td>
</tr>
<tr>
<td>11.</td>
<td>Instrumental variable estimation and two stage least squares</td>
<td>Chapters 15 and 16</td>
</tr>
<tr>
<td>12.</td>
<td>Panel data methods</td>
<td>Chapters 13 and 14</td>
</tr>
</tbody>
</table>

*(c) Duncan Thomas 2020*
If, after class, you have difficulty with the lecture notes and Wooldridge’s presentation, you should consult one of the alternative texts. They are listed, with chapter references, below. Alternative readings are identified only to assist you and are not required.

**Alternative readings**

**Sections 1, 2 and 3: Introduction to regression model**
- Angrist and Pischke: Chapters 1 and 2
- Goldberger: Chapters 1 and 2
- Gujarati: Chapters 1, 5 and 6
- Hill, Griffiths and Lim: Chapters 1, 2 and 4
- Stock and Watson: Chapter 1

**Section 4: Theory of estimation and inference**
- Goldberger: Chapters 2-4
- Gujarati: Chapters 2, 3
- Hill, Griffiths and Lim: Chapters 1P, 3
- Stock and Watson: Chapters 2-3

**Section 5, 6 and 7: Classical multiple regression model**
- Angrist and Pischke: Chapter 2
- Goldberger: Chapters 6-12
- Gujarati: Chapters 7-9
- Hill, Griffiths and Lim: Chapters 5-7
- Stock and Watson: Chapters 4-7

**Sections 8, 9 and 10: Relaxing assumptions of the regression model**
- Goldberger: Chapters 13-17
- Gujarati: Chapters 10-14
- Hill, Griffiths and Lim: Chapter 8
- Stock and Watson: Chapters 8, 9 and 11

**Sections 11: Unobserved heterogeneity and instrumental variable methods**
- Angrist and Pischke: Chapter 3
- Goldberger: Chapters 18 and 20
- Gujarati: Chapter 15
- Hill, Griffiths and Lim: Chapters 10-11
- Stock and Watson: Chapter 12

**Section 12: Panel data methods**
- Angrist and Pischke: Chapter 5
- Gujarati: Chapters 10, 11 and 12
- Hill, Griffiths and Lim: Chapter 15
- Stock and Watson: Chapter 10