

**Description:** This course is intended as an introduction to statistical data analysis for social scientists. We cover linear regression and as many of the most commonly used non-linear models (e.g., logit/probit, duration models, event-count models) as time allows. Grades derive from problem sets, a final paper, and an open-book, take-home final. Familiarity with material covered in the pre-requisites, PS599 (Statistical Methods I) and PS598 (practical calculus and linear algebra) or their equivalent, is assumed. We are fairly rigorous in this course; however, our primary goal is to develop an adeptness in understanding and applying the statistical analysis of historical data (almost exclusively the only kind we have) to furthering our understanding social-scientific phenomena. As such, the emphasis is on “hands-on” examples and exercises aimed at developing an intuition for statistical procedures, properties, and pitfalls.

**Texts:** I strongly recommend you purchase William Greene, *Econometric Analysis* (despite its ridiculous price tag). Greene is the most thorough and up-to-date text available, and the one you will want to have on your shelf when you have finished the course. However, many students find it too technical at this point, so you may wish to have something like Damodar Gujarati, *Basic Econometrics*, Kmenta, *Elements of Econometrics*, or Goldberger, *A Course in Econometrics*, also. Many students find Peter Kennedy, *A Guide to Econometrics*, an invaluable companion volume. It is a remarkably intuitive, reader-friendly, introduction to some of the concepts (it will not suffice on its own however). For the non-linear models part of the course, we will use Gary King’s *Unifying Political Methodology*, on the utility of which students and instructors generally agree. Also useful are John Aldrich’s and Forrest Nelson’s *Linear Probability, Logit, and Probit Models* and Scott Menard’s *Applied Logistic Regression Analysis*, two of the better Sage volumes on binary dependent variables. Finally, students who would like a helpful math text to accompany should see Alpha Chiang’s excellent and remarkably accessible *Fundamental Methods of Mathematical Economics*. Greene, Kennedy, King, and the two sage volumes are available at campus bookstores. They plus Chiang, Kleppner and Ramsey (see below), De Groot (see below), Mendenhall, Wackerly, and Scheaffer (see below), Gujarati, Kmenta, and Goldberger are also on reserve. If you plan to take 787, I highly recommend you buy Greene and King at least. My own notes are available at *Dollar Bill Copying*. Some students have found them plus one of the main texts sufficient.

**Other Useful Texts and Sources:**

- T. Amemiya, *Introduction to Statistics and Econometrics*, 1994. (An excellent recent treatment, strong on fundamentals.)
- E. Berndt, *The Practice of Econometrics*, 1991. (Limited almost entirely to economic examples, but a very readable, experiential approach to learning how “to do econometrics.”)
- A. Goldberger, *A Course in Econometrics*, 1991. (Very clearly and understandably written, but short on breadth and examples.)
- E. Hanushek and J. Jackson, *Statistical Methods for Social Scientists*, 1977. (A bit dated now, but still one of the clearest and best introductions you’re likely to find.)
- J. Johnston, *Econometric Methods*, 3rd ed., 1984. (Another classic, but still one of the most straight-forward, no-nonsense treatments of the basics. Tightly written: highly recommended if you dislike verbosity and prefer the simple math.)
- G. Judge *et al.*, *Theory and Practice of Econometrics*, 2nd ed., 1990. (The consummate reference text--more than a bit beyond the level of this course though.)
- G. King, *Unifying Political Methodology*, 1989. (MLE made simple. Useful just to build intuition about statistic estimation and inference even if you never use anything more complicated than OLS.)
- G. King, *et al.* *Designing Social Inquiry*, 1994. (Verbal intuition of the important methodological issues given in the context of exploring their application to qualitative research.)
- J. Kmenta, *Elements of Econometrics*. (Treatment of basics is intuitive and strong; limited scope and dated now.)
- G.S. Maddala, *Introduction to Econometrics*, 2nd ed., 1992. (A good addition to your reference list it’s somewhere between Johnston and Kmenta, covering some of the material often omitted elsewhere.)
- R.S. Pindyck and D.L. Rubinfeld, *Econometric Models and Econometric Forecasting*, 3rd ed., 1991. (An eminently readable intro to linear regression, emphasizing time-series issues. Quite limited regarding non-linear models. Entirely economic in examples.)
- S. Weisberg, *Applied Linear Regression*, 2nd ed., 1985. (An applied approach, but limited in scope.)

Also, a series of Sage Papers called “Quantitative Applications in the Social Sciences” provide introductions to specific topics in data analysis, ranging from the most basic to the most sophisticated and specialized. Many are quite good, such as Chris Achen’s *Interpreting and Using Regression Analysis* and the two we use on binary dependent variables, but quality varies. They are good places to begin investigation into some technique you have not used before.

Finally, the formerly-annual and now-quarterly review of the American Political Science Association's Methodology Section, *Political Analysis*, has some excellent articles treating and applying various techniques. Most everything in that publication is worth reading.

For **math help**, see Kleppner and Ramsey, *Quick Calculus*, and, of course, Chiang. (Chiang is a tremendously valuable resource: the most understandable introduction to the math social scientists need I have ever seen. If you are doing any formal theory, I would consider it a must.) For **stats help**, see De Groot, *Probability and Statistics* (THE classic statistics text), or Mendenhall, Wackerly, and Scheaffer, *Mathematical Statistics with Applications* (readable despite the scary title).

### Course Requirements

**The course meets** Tuesdays and Thursdays for 2 hours per session. If you have questions, ask them in class or ASAP of your GSI, Sarah Croco (in whom I have the utmost confidence), or myself. If you are having any trouble, please go to her or my office hours, send her or me e-mail, call for an appointment, send a carrier pigeon; do something! Do *not* simply sit on the problem; it will grow, not go away. We, your GSI and I, are here to help. If you want to understand the material and are willing to keep trying, I *guarantee* we will keep trying too and we will find a way to translate it that makes sense to you.

You will have periodic (almost weekly) **problem sets** distributed in class, due one week after distribution to your GSI, and combining to 30% of your grade.

You will also have a **final exam**: take-home and open-book, open-computer, open-anything-but-another-human-being. The exam is worth 35% of your grade and will be during examination period.

You will, finally, have one **paper** assignment. Find an article that interests you and that applies statistical methodology comparable to or more sophisticated than the material covered in this course. Obtain the data from the author if at all possible. Replicate the published results as nearly as possible. Try to determine why you can't replicate them exactly if you cannot; that is, explain how, statistically, your results could differ from those published in the way they do. Then extend the analysis in some way. You could, for example, (a) suggest a more appropriate functional form for the estimation and re-estimate, (b) argue that one or a set of important variables were omitted and conduct the analysis anew, (c) argue that the results are likely sensitive to sample selection or variable measurement *etc.* and then conduct appropriate analyses to address that possibility, (d) extend the data or use a different data set to test the theory, or (e) any other good idea you might have. Please see me if you have any questions or difficulty regarding your chosen project (including doubts about its applicability to this requirement). This is the final 35% of your grade. **The paper is due in three parts.** Choose a paper to replicate by the end of week 5 (2/6/01). Collect the data and provide descriptive statistics and graphs of them by the end of week 10. The finished paper is due the last day of exam period.

**Summary of Requirements:** Problem Sets: 30% ; Take-Home Final: 35% ; Paper: 35%

### Quick Overview and Outline of the Course

- I. Introduction and Review
  - A. Meeting One: Introduction and Logistics; Begin Math Review (1/6)
  - B. Linear Algebra and Basic Calculus Fundamentals for Econometrics (1/11)
  - C. Probability and Distribution Theory (1/13, XX)
  - D. Statistical Inference Review (1/20)
- II. The Classic Linear Regression Model
  - A. Bivariate (1/25, 1/27)
  - B. Multivariate (2/1, 2/3)
- III. More Inference, Interpretation
  - A. More Hypothesis Testing (2/8, 2/10)
  - B. Functional Form, Non-Linearity, and Specification Issues (2/15)
  - C. Data "Problems" (2/17)
  - D. More Regression Diagnostics (2/22, 2/24)
- E. Closing up some Large Sample Results for the CLRM (3/8)
- IV. Non-Spherical Disturbances
  - A. General Treatment (3/10, 3/15)
  - B. More on Heteroskedasticity (3/17)
  - C. Correlated Disturbances and Time-Series Models (3/22, 3/24)
- V. Advanced Topics
  - A. Pooled Time-Series-Cross-Section (3/29)
  - B. Endogeneity / Simultaneous Equations (3/31, 4/5)
  - C. Dichotomous Dependent Variables: Logit and Probit (4/7, 4/12, 4/14)
- VI. Other Topics (4/19, as time allows)
  - A. Durations and Counts
  - B. Multinomial Logit and/or Probit
  - C. More Time-Series, TSCS Complications

## Weekly Course Schedule

All readings (plus the notes) are “suggested”. You should determine for yourself which texts most help you in grasping the material; start with Greene or Gujarati and turn to Kennedy and/or my notes for a more intuitive (less rigorous) presentation or *vice versa*. Learning to find methodological references that enable you to grasp and employ new techniques is one of the skills you will hopefully learn in this course.

### I. Introduction and Review

A. Meeting One: Introduction and Logistics; Begin Math Review **(1/6)**

B. Linear Algebra and Basic Calculus Fundamentals for Econometrics **(1/11)**

Greene App A.1-7; Gujarati App. B; Kmenta App. A-B; Goldberger 17; Chiang, 4-5; Weisburg 2A.1-2

Greene App A.8; Kleppner & Ramsey -- work on what you need; Chiang, 6-8

C. Probability and Distribution Theory **(1/13, xx1/18xx)**

Greene App B; Gujarati App. A.1-6; Kennedy 2; Kmenta 3-4; Goldberger 1-7; King 1-3

D. Statistical Inference Review **(1/20)**

Greene App C-D; Gujarati App. A.7-8; Kennedy 4; Kmenta 1-2,5-6; Goldberger 8-12; King 4

### II. The Classic Linear Regression Model

A. Bivariate **(1/25, 1/27)**

Greene 2-4, 5.1-2, 6.4; Gujarati 1-4; Kennedy 3; Kmenta 7; Weisburg 1; Goldberger 13

1. Assumptions (Greene 2)

2. Least-Squares Regression (Greene 3.2; skim 3.3-3.4)

3. Goodness of Fit and Analysis of Variance (Greene 3.5; skim 3.6)

4. Statistical Properties in Finite Samples (Greene 4.1-4.8)

5. Asymptotic Properties (Greene 5.1-5.2)

6. Stochastic X and Non-Normality (Greene 4.5; 6.4)

B. Multivariate **(2/1, 2/3)**

Greene 2-4, 5.1-2, 6.4; Gujarati 7, 9.1-5; Kennedy 3; Kmenta 10.1-2; Goldberger 14-16,17; Weisburg 2

1. Assumptions (Greene 2)

2. Least-Squares Regression (Greene 3.2; skim 3.3-3.4)

3. Goodness of Fit and Analysis of Variance (Greene 3.5; skim 3.6)

4. Statistical Properties in Finite Samples (Greene 4.1-4.8)

5. Asymptotic Properties (Greene 5.1-5.2)

6. Stochastic X and Non-Normality (Greene 4.5; 6.4)

### III. More Inference, Interpretation, and Prediction

A. More Hypothesis Testing **(2/8, 2/10)**

Greene 6-7, 8.3; Gujarati 5, 8, 9.6-11; Kennedy 4; Kmenta 7.4, 10.2; Goldberger 20-22

1. Confidence Intervals (Gujarati 5.1-5.4; Kennedy 4.4)

2. Restrictions on Coefficients (Greene 6.1-6.2; Gujarati 5.5-8, 8.1-7, 9.6-9; Kennedy 4.2-3)

3. General Testing Procedures (Greene 6.3-6.5; Gujarati 8.11-12; Kmenta 11.2:491-497; Kennedy 4.5)

4. Structural Change (Greene 7.4; Gujarati 8.8)

5. More Tests of the Model (Greene 7.5-7.6, 8.3; Kmenta 11.10)

6. Prediction (Greene 6.6; Gujarati 8.10, 9.9)

7. Interpretation and Evaluation (Gujarati 5.9-13)

B. Functional Form, Non-Linearity, and Specification Issues **(2/15)**

Greene 4.9, 5.6, 7.1-7.3, 8-9; Gujarati 6, 15; Kennedy 5-6, 14; Kmenta 10.4

1. Dummy Variables (Greene 7.1-7.2; Gujarati 15; Kennedy 14; Kmenta 11.1; Weisberg 7.2-4)
  2. Non-Linearity in Variables (and interactions) (Greene 7.3; Gujarati 6; Kennedy 6.3; Kmenta 11.7)
  3. Specification; Inclusion and Omission of Variables (Greene 8.1-8.2; Gujarati 13.2-4; Kennedy 5, 6.2; Goldberger 24:190)
  4. Non-linear Regression Models (Greene 9; Gujarati 6; Kennedy 6.3)
- C. Data “Problems” (2/17)
- Greene 4.9, 5.6, ; Gujarati 10, 13.5
1. Multicollinearity/Micronumerosity (Greene 4.9; Gujarati 10; Kennedy 11; Kmenta 10.3; Goldberger 23; Weisberg 8.2)
  2. Measurement Error and Proxy Variables (Greene 5.6; Gujarati 13.5; Kmenta 9.1)
  3. Missing Observations and Grouped Data (Greene 4.9; Kmenta 9.2-3; Weisburg 3.3)
- D. More Regression Diagnostics (2/22, 2/24)
- Jim DeNardo’s Regression-Diagnostic notes; Kennedy 18; Weisberg 5, 6.1-5
- E. Closing up some Large Sample Results for the CLRM (3/8) (if necessary)
- Greene 5; Kmenta App. C; Goldberger 18-19
- IV. Non-Spherical Disturbances
- A. General Treatment (3/10, 3/15)
- Greene 5.3, 10; Gujarati 11, 12; Kennedy 8; Kmenta 12.1; Goldberger 27-28; Weisburg 4.1
1. Consequences for OLS Estimation of Non-Spherical Disturbances (Greene 10.1-2; Gujarati 11.2, 11.4, 12.2, 12.4; Kennedy 8.2)
  2. Robust Estimation of Asymptotic Covariance Matrices (Greene 10.3)
  3. Efficient Estimation by GLS (Greene 10.5.1; Gujarati 11.3, 12.3)
  4. Feasible GLS when  $\Omega$  Unknown (Greene 10.5.2; Gujarati 11.6, 12.6;)
  5. ML Estimation (Greene 10.6)
- B. More on Heteroskedasticity (3/17)
- Greene 11; Gujarati 11; Kennedy 8.3; Kmenta 8.2; Goldberger 28.2
1. Robust OLS Estimation (Greene 11.2)
  2. Testing for Heteroskedasticity (Greene 11.4)
  3. GLS (=WLS) and FGLS (FWLS) (Greene 11-5)
  4. Conclusions, Extensions (ARCH) (Greene 11.6-9)
- C. Correlated Disturbances and Time-Series Models (3/22, 3/24)
- Greene 12, 19-20; Gujarati 12, 17; Kennedy 8.4; Kmenta 8.3; Goldberger 28.3; King 7; Beck, Political Analysis
- V. Advanced Topics
- A. Pooled Time-Series-Cross-Section (3/29)
- Greene 13; Gujarati 15.12; Kmenta 12.2-3; Beck & Katz, Political Analysis
- B. Endogeneity / Simultaneous Equations (3/31, 4/5)
- Greene 14-15; Gujarati 18-20; Kennedy 10; Kmenta 13; Goldberger 30-4; King 8.2; Bartels, AJPS
- C. Dichotomous Dependent Variables: Logit and Probit (4/7, 4/12, 4/14)
- Greene 17, 21.1-21.6; Gujarati 16; DeMaris, Logit Modeling; Kennedy 15.1; Kmenta 11.5; Weisberg 12.2; King 5.1-3, 6
- VI. Other Topics (4/19, as time allows)
- A. Durations and Counts (Greene 22.5, 21.9-21.10; Kennedy 15.4; King 5.7-9)
  - B. Multinomial Logit and/or Probit (Greene 21.7-8; Kennedy 15.2-3; Kmenta 11.6; King 5.4-5)
  - C. More Time-Series Complications (Greene 19-20; Gujarati 21-22; Kennedy 16-17)