

## Political Science 699

### Statistical Methods II, Winter 2008

MW 1-3, G228 Angell; Make-up Sessions: Th 11-1, 216 Dennison

LAB (GSI: Bryce Corrigan, becorrigan@umich.edu)

**Professor Robert (Rob) J. Franzese, Jr.**

Office Hours: Mon 3:15-4:45 (& by appt), 4246 I.S.R.

franzese@umich.edu

<http://www-personal.umich.edu/~franzese>

**UPDATE: 3/25/08**

**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 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* to start. Kmenta's *Elements of Econometrics* and Goldberger's *A Course in Econometrics* are two other classics that you may find accessible, and that are currently available; I've placed them on reserve so you may decide which of those or Gujarati you most prefer if Greene is too daunting to start. Many students find Peter Kennedy, *A Guide to Econometrics*, an invaluable companion volume also. It will not suffice on its own, but it is a remarkably intuitive, reader-friendly, introduction to and explanation of some of the concepts. For the qualitative or limited dependent-variables parts of the course, we will use Gary King's *Unifying Political Methodology*, on the utility of which most generally agree. We will also use some smaller pedagogical manuscripts on specific topics: Cindy Kam's and my UM Press volume on interaction terms and 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. For (relatively) accessible additional statistics and probability texts (good to have as reference, for example) consider De Groot & Schervish, *Probability and Statistics* (which Barnes & Noble seems to have in paperback for only \$25!). Finally, students who would like a helpful math text to accompany the statistics and econometrics may want to see Alpha Chiang's excellent and remarkably accessible *Fundamental Methods of Mathematical Economics*. Simon and Blume's *Mathematics for Economists* is the more-current standard, but it is also likely less accessible and, if anything, too broad and thorough for our needs. Gill's recent *Essential Mathematics for Political and Social Research* is newer and so less-tested, but perhaps its greater topicality and accessibility may aid comprehension. If you want more of a starter on basic calculus, Kleppner and Ramsey's *Quick Calculus: A Self-Teaching Guide* actually works as titled, and so is quite worth getting and following, especially given its low price, even though it does not go very far, wide, or deep. Greene, Gujarati, Kennedy, King, the two sage volumes, and my and Cindy's book are available at campus bookstores. They, everything else recommended in this paragraph, and few articles and other materials assigned or recommended below, are also on reserve. If you plan to take 787, I highly recommend you buy Greene and King at least. My lecture notes are available at *Dollar Bill Copying*. Some students have found them plus one of the main texts sufficient.

### **Books Ordered &/or on Reserve:**

1. John Aldrich & Forrest Nelson, *Logit and Probit Models*, Sage, 1985.
2. William Greene, *Econometric Analysis*, 6<sup>th</sup> ed., Pearson Education, Prentice Hall, 2007.
3. Damodar Gujarati, *Basic Econometrics*, McGraw-Hill, 4<sup>th</sup> ed., 2002.
4. Cindy Kam & Robert Franzese, *Modeling & Interpreting Interactive Hypotheses in Regression Analysis*, U Michigan Press, 2007.
5. Gary King, *Unifying Political Methodology*, U Michigan Press, 1998.
6. Scott Menard, *Applied Logistic Regression Analysis*, Sage, 2001.
7. Alpha Chiang & Kevin Wainwright, *Fundamental Methods of Mathematical Economics*, McGraw-Hill, 4<sup>th</sup> ed., 2005. (Recommended)
8. Jeff Gill, *Essential Mathematics for Political and Social Research*, Cambridge UP, 2006. (Recommended)
9. Arthur Goldberger, *A Course in Econometrics*, Harvard U Press, 1991. (Recommended)
10. Peter Kennedy, *A Guide to Econometrics*, 5<sup>th</sup> ed., MIT Press, 2003. (Recommended)
11. Daniel Kleppner & Norman Ramsey, *Quick Calculus: A Self-Teaching Guide*, Wiley, 1986. (Recommended)
12. Jan Kmenta, *Elements of Econometrics*, 2<sup>nd</sup> ed., U Michigan Press, 1997. (Recommended)
13. Carl Simon & Lawrence Blume, *Mathematics for Economists*, W.W. Norton, 1994. (Recommended)

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.

### Course Requirements

**The course meets** Mondays and Wednesdays for 2 hours per session. If you have questions, ask them in class or ASAP of your GSI, Bryce Corrigan (in whom I have the utmost confidence), or myself. If you are having any trouble, please visit his or my office hours, send us e-mail, call for an appointment, send carrier pigeons; 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** available online, due as announced in class to your GSI, in the manner Bryce determines, 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 (to be distributed Friday, April 18, 15:00, and due Monday, April 21, 12:00).

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 6 (2/15/08). Collect the data and provide descriptive statistics and graphs of them by the end of week 10 (3/21/08). The finished paper is due the morning (8:00am) of the last day of exam period (4/24/08).

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

### Outline of the Course

#### I. Introduction and Review

- A. Introduction, Logistics, Math Refresher (***1/7, 1/9, 1/14***)
- B. Probability- and Distribution-Theory Review (***1/16, 1/23***)
- D. Statistical-Inference Review (***1/24, 1/28, 1/30***)

#### II. The Classic Linear Regression Model (***2/4, 2/6, 2/11, 2/13***)

- A. Assumptions
- B. Least-Squares Regression
- C. Goodness of Fit and Analysis of Variance
- D. Statistical Properties in Finite Samples
- E. Stochastic X and Non-Normality
- F. Asymptotic Properties

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

- A. Estimate & Express Certainty & Test Hypoth's (***2/18, 2/20, 3/3***)
- B. Functional Form, Non-Linearity, Specification Issues (***3/5, 3/6***)

#### C. Data "Problems" & Regression Diagnostics (***3/6, 3/10***)

#### IV. Non-Spherical Disturbances

- A. General Treatment (***3/12, 3/20***) P-Set (6F and) 7: 3/12
- B. More on Heteroskedasticity (***3/20, 3/24***) P-Set 8: 3/20
- C. Correlated Disturbances, Lagged Dependent-Variables, & Time-Series Models (***3/27***) P-Set 9: 3/31

#### V. Advanced Topics

- A. Qualitative & Limited Dep-Var Models (***3/31, 4/2, 4/3***)
- B. FURTHER TOPICS: (***4/7, 4/9, 4/10***) P-Set 10: 4/9
  - 1. Endogeneity, Simultaneous Equations, & Strategies for Causal Analysis
  - 2. Models for Multilevel, Panel, and Time-Series-Cross-Section Data

### Detailed Course Schedule

The italicized and underlined dates in the outline above and in the detailed schedule below are **make-up sessions**. Those four sessions are on **Thursdays, 11-1, in Dennison 216**. All readings (plus the notes) in the detailed schedule below are "suggested". The notes and lectures will most closely, but still sometimes only rather roughly, follow the italicized readings. 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. First Meeting: Introduction, Logistics, Math Refresher (**1/7**)

Linear-Algebra and Basic-Calculus Fundamentals: (**1/7, 1/9, 1/14**)

*Greene Appendix A*; Gujarati Appendix B; Kmenta Appendix A-B; Goldberger 17; Chiang 4-8; Kleppner & Ramsey; Gill 3-6; Simon & Blume 2, 4.1, 5-9, 11.1, 13.3, 14.1, 14.4-5, 14.7-8, 16, 19, 21, 23, 26, A4

### B. Probability- and Distribution-Theory Review (**1/16, 1/23**)

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

### D. Statistical-Inference Review (**1/24, 1/28, 1/30**)

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

## II. The Classic Linear Regression Model (**2/4, 2/6, 2/11, 2/13**)

*Greene 2-4*; Gujarati 1-4, 7, 9.1-5; Kennedy 3; Kmenta 7, 10.1-2, Appendix C; Goldberger 13-16,17-19

### A. Assumptions (*Greene 2*)

### B. Least-Squares Regression (*Greene 3.1-2, skim 3.3, 3.4*)

### C. Goodness of Fit and Analysis of Variance (*Greene 3.5-6*)

### D. Statistical Properties in Finite Samples (*Greene 4.1-4, 4.6, 4.8*)

### E. Stochastic X and Non-Normality (*Greene 4.5*)

### F. Asymptotic Properties (*Greene 4.9*)

## III. Inference, Interpretation, and Prediction

### A. Estimating and Expressing Certainty and Testing Hypothesis (**2/18, 2/20, 3/3**)

*Greene 4.7, 5, 6.4, 7.3*; Gujarati 5, 8, 9.6-11; Kennedy 4; Kmenta 7.4, 10.2; Goldberger 20-22

#### 1. Confidence Intervals (*Greene 4.7.2-3*; Gujarati 5.1-5.4; Kennedy 4.4)

#### 2. Restrictions on Coefficients (*Greene 4.7.1, 4.7.4-5*; Gujarati 5.5-8, 8.1-7, 9.6-9; Kennedy 4.2-3)

#### 3. General Testing Procedures (*Greene 5.1-5*; Gujarati 8.11-12; Kmenta 11.2:491-497; Kennedy 4.5)

#### 4. Structural Change (*Greene 6.4*; Gujarati 8.8)

#### 5. More Tests of the Model (*Greene 7.3*; Kmenta 11.10)

#### 6. Prediction (*Greene 5.6*; Gujarati 8.10, 9.9)

#### 7. Interpretation and Evaluation (*Greene 5.7*; Gujarati 5.9-13)

### B. Functional Form, Non-Linearity, and Specification Issues (**3/5, 3/6**)

*Greene 6.1-3, 7.1-2, 7.4-6, 11*; Gujarati 6, 13-15; Kennedy 5-6, 14; Kmenta 10.4

#### 1. Dummy Variables (*Greene 6.1-6.2*; Gujarati 15; Kennedy 14; Kmenta 11.1)

#### 2. Interactions and Nonlinearity in Variables (*Greene 6.3*; *Kam & Franzese*; Brambor, Clark, & Golder Political Analysis; Gujarati 6; Kennedy 6.3; Kmenta 11.7)

#### 3. Specification; Inclusion and Omission of Variables (*Greene 7.1-7.2*; Gujarati 13.2-4; Kennedy 5, 6.2; Goldberger 24:190)

#### 4. Model Selection (*Greene 7.4-6*; Gujarati 6, 13; Kennedy 6.3)

#### 5. Non-linear Regression Models (*Greene 11*; Gujarati 14)

### C. Data “Problems” & Regression Diagnostics (**3/6, 3/10**)

*Greene 4.8.1, 12.5, 4.8.2*; Gujarati 10, 13.5

#### 1. Multicollinearity/Micronumerosity (*Greene 4.8.1*; Gujarati 10; Kennedy 11; Kmenta 10.3; Goldberger 23; Weisberg 8.2)

#### 2. Measurement Error and Proxy Variables (*Greene 12.5*; Gujarati 13.5; Kmenta 9.1)

#### 3. Missing Observations and Grouped Data (*Greene 4.8.2*; Kmenta 9.2-3)

#### 4. More Regression Diagnostics (*Jim DeNardo’s Regression-Diagnostic notes*; Gujarati 5.12, 10.7-9, 13; Kennedy 18)

## IV. Non-Spherical Disturbances

### A. General Treatment (**3/12, 3/20**)

*Greene 8.1-3*; Gujarati 11, 12; Kennedy 8; Kmenta 12.1; Goldberger 27-28; Weisburg 4.1

1. Consequences for OLS Estimation of Non-Spherical Disturbances (*Greene 8.1-8.2.2*; Gujarati 11.2, 11.4, 12.2, 12.4; Kennedy 8.2)
2. Consistent Estimation of Asymptotic Covariance Matrices (*Greene 8.2.3*)
3. Efficient Estimation by GLS (*Greene 8.3.1*; Gujarati 11.3, 12.3)
4. Feasible GLS when  $\Omega$  Unknown (*Greene 8.3.2*; Gujarati 11.6, 12.6;)
5. ML Estimation (*Greene 16.9.2, 16.9.2.a*)

**B. More on Heteroskedasticity (3/20, 3/24)**

*Greene 8.4-9*; Gujarati 11; Kennedy 8.3; Kmenta 8.2; Goldberger 28.2

1. OLS Estimation (*Greene 8.4.1-2*)
2. OLS Estimation with Consistent Estimation of Asymptotic Covariance Matrices (*Greene 8.4.3-4*)
3. Testing for Heteroskedasticity (*Greene 8.5*)
4. GLS (=WLS) and FGLS (FWLS) (*Greene 8.6-7*)
5. Applications, Conclusions (*Greene 8.8-9*)

**C. Correlated Disturbances, Lagged Dependent-Variables, & Time-Series Models (3/27)**

*Greene 19-20, 22*; King 7; Beck, *Political Analysis*; Gujarati 12, 17, 21-22; Kennedy 8.4, 16-17; Kmenta 8.3; Goldberger 28.3

1. OLS Estimation, and Consistent Estimation of Asymptotic Covariance Matrices (*Greene 19.1-5*)
2. FGLS Estimation of ARIMA(p,d,q) Models (*Greene 19.8-9*)
3. (Time-)Lagged Dependent-Variable Models (*Greene 20.1-20.4.3, 20.5-20.5.2*; King 7; Beck, *Political Analysis*)
4. Forecasting (*Greene 19.12, 20.4.4*)
5. Testing for (Temporal) Autocorrelation (*Greene 19.7, 20.5.3*)
6. (Non-)Stationarity and Cointegration (*Greene 22-22.3*)
7. ARIMA, Panel/TSCS, VAR, ARCH, & Other Acronymious Extensions (*Greene 19.6, 19.10-11, 19.13-14, 20.6-7, 21, 22.4-5*)

**V. Advanced Topics**

**A. Qualitative and Limited Dependent-Variable Models Endogeneity, Simultaneous Equations, Strategies for Causal Analysis (3/31, 4/2, 4/3)**

1. Binary Choice: Logit and Probit (*King 5.1-3, 6*; Aldrich & Nelson; Menard; Gujarati 16; Greene 23.1-9; Kennedy 15.1; Kmenta 11.5)
2. Multiple Choice: Multinomial Logit and/or Probit (*King 5.4-5*; Greene 23.10-12; Kennedy 15.2-3; Kmenta 11.6)
3. Durations and Counts (*King 5.7-9*; Greene 25; Kennedy 15.4)
4. Truncation, Censoring, and Selection (*Greene 24*; King 9)

**B. FURTHER TOPICS (4/7, 4/9, 4/10)**

1. Endogeneity, Simultaneous Equations, & Strategies for Causal Analysis
  - a. Systems of Equations, Endogeneity, and Simultaneous Equations Models (*Greene 10, 13.1-3*)
  - b. Instrumental-Variables (Method-of-Moments) Strategies (*Greene 12, 13.4-10*; Bartels, *AJPS*; Gujarati 18-20; Kennedy 10; Kmenta 13; Goldberger 30-4; King 8.2)
  - c. Vector Autoregression and Vector Error-Correction (Reprise) (*Greene 20.6, 22.3*)
  - d. Matching, Propensity-Score Matching, and “Causal Analysis” (*Greene 24.5.4-6*)
  - e. Regression-Discontinuity Design (*Franzese & Hays, unpublished IPES conference-paper*)
2. Models for Multilevel, Panel, and Time-Series-Cross-Section Data
  - a. Models for Multilevel/Hierarchical Data (*Franzese Political Analysis*)
  - b. Models for Panel Data (*Greene 9*; Gujarati 15.12; Kmenta 12.2-3)
  - c. Models for Time-Series-Cross-Section Data (*Beck & Katz, Political Analysis*; Franzese & Hays *Political Analysis*)