

**ADVANCED MULTIVARIATE STATISTICS
(A.K.A.: SLOUCHING TOWARD CAUSATION)
NYU, DEPARTMENT OF SOCIOLOGY
WEDNESDAYS 10:00 – 12:00**

DALTON CONLEY, INSTRUCTOR
OFFICE: 6 WASHINGTON SQUARE NORTH, ROOM 20
OFFICE HOURS: BY APPOINTMENT ONLY

This course will present the graduate student who has already completed one semester of statistics with a deeper engagement into the question of how we make causal claims within the context of analyzing observational data. Given this intent, the focus will be on the operationalization of research questions, the management of data and the interplay between methodology and theory. If you never calculate a mean or standard deviation ever again in your life, this course should still be of use to you since we will be thinking about questions of research design and causal inference in a way that would inform even comparative historical or ethnographic research. If the student walks away with just one concept deeply embedded in her/his thinking, I will be happy. That concept is the following: that association does not equal causation.

In short, it is a course about how to think like a social scientist. Statistics just happen to be the best vehicle to get at the relevant cognitive structures. From a pragmatic point of view, by the end of this course, the student should be able to understand *when* to use a given technique and *how* to use it in the SPSS, STATA or R statistical package.

To these ends, about five weeks into the term, students will have to read the assigned journal article(s) and prepare themselves to discuss them in class as if they were the editorial board for *ASR* or another social science journal. Many of the assigned articles throughout the semester have been written by myself and coauthors. This is meant to provide you with a guide to my particular thinking about the issues of statistical inference in the social sciences. That is, you can see how my methodology has evolved over the last few years as I (and sociology more generally) have grappled with the issues of unobserved heterogeneity (selection bias or spuriousness) and endogeneity (reverse causation). Don't be afraid to rip my articles to shreds—that's what they are there for (and I certainly will tear them apart myself). One of the things we need to learn is to take a licking and keep on ticking—this is how “normal” science works. As I run through a lot of course material in the first half of the class, I may not be able to address the weekly

reading examples so thoroughly (or at all), but you are still expected to come to class having read them. I hope to get through all the “equations” in the first half of the semester (though the material may spill over if we fall behind) in order to leave the second half of the semester as a more workshop setting: discussing and debating the reading assignments, coming up with the project proposals, and presenting the final assignments.

In addition to four paper critiques, you will have two data assignments; the first is given on the very first day of class (you have two weeks to complete it). The second one is given on week six (and you have four weeks to complete it). Lastly, your final assignment is to write a three- to four-page proposal for a project that addresses any research question you want in a quantitative framework, specifies the data that you will use (and includes an extract of those data under separate cover), and makes a decent attempt to “slouch toward causation” using one or more of the techniques that we will have discussed throughout the semester. Drafts of these proposals will be presented orally in the final weeks of the class for feedback and revision.

The class will meet once a week; however, I expect, when scheduled, students to attend the quantitative seminar series running on Wednesdays, 12:00 – 1:30 in the CASSR seminar room. This exposure to current quantitative research will aid your understanding of methods. Students are also going to be expected to familiarize themselves with the Panel Study of Income Dynamics (PSID), the General Social Survey (GSS) or another dataset that receives the approval of the instructor. Given the longitudinal design of the PSID and its snowball sampling technique (following households and split-offs from those households each year since 1968), it will allow us to address many of the pertinent research issues of the course while giving the students hands-on experience in coding and managing data sets themselves. The GSS, by contrast, provides us with a repeated cross section to address other methodological issues (such as the joint estimation of age and cohort effects, testing for regression breaks, and time series analysis).

TEXTS

Each week, students are required to read and be prepared to respond to the assigned empirical research papers; these will be available on JSTOR or through download on my homepage: homepages.nyu.edu/~dc66. There is also a methodological text which is required reading, *Introductory Econometrics: A Modern Approach*, by Jeffrey Wooldridge (Southwestern College Publishers). In addition, there are two texts that I recommend that everyone acquire at some

point:

Recommended for your Bookshelf:

William H. Greene, *Econometric Analysis*. (Prentice Hall, 5th Edition)

Edward Tufte, *The Visual Display of Quantitative Information*. (Graphics Press, 2nd Edition)

Useful Websites:

Mathworld. “The World’s Most Extensive Mathematics Resource”

<http://mathworld.wolfram.com/>

UCLA Academic Technology Services Statistical Computing Resources

<http://www.ats.ucla.edu/stat/>

:

Stata Archives and listserve

<http://www.stata.com/statalist/>

ASSIGNMENTS AND GRADING

A one page methodological critique of the assigned substantive articles for a given week is due on six occasions during the semester (author’s choice). The final exam will ask about statistical techniques and will also include an article critique. Finally, this is expected to be an interactive, research-generating course, so class participation is key to its success. Grades will be calculated according to the following method:

Weekly article critiques	20 percent
Data Tasks	15 percent
Final Assignment	20 percent
Final Exam	20 percent
Class Participation	25 percent

COURSE OUTLINE AND READINGS

Week One: Introduction to Data and Software + Regression Review and Violations

- I. SPSS & STATA resources
- II. GSS and PSID
- III. Mediation versus Moderation
 - I. Subgroup Regressions versus Interaction Terms
 - II. Interpreting Interaction Terms
- IV. Coefficient Comparison across/within Models
- V. Hierarchical Models: Chow Test
- VI. Multiple Hypothesis Testing: Bonferroni and False Discovery Rate
- VII. Assumption Violations and Diagnostics (BLUE: Best Linear Unbiased Estimator)
 - A. Multicollinearity: VIFs
 - B. Heteroscedasticity: White Test, Partial Plots
 - C. Non-normality: Functional Forms

Text: Chapters 1, 2, 3, 4, 7 and 8 (5 would be good too if time)

In detail:

- Assumptions (BLUE): Wooldridge (version 2E): Chapters 2 and 3 especially MLR.1 through MLR.4 for unbiasedness plus MLR.5 for efficiency and the Gauss-Markov Theorem on page 103 for BLUE (*best linear* is MLR.5 and *unbiased estimate* is MLR.1-MLR.4). In sum, these make up the Gauss-Markov.
- Multicollinearity: Chapter 3 especially pages 95-101.
- Heteroskedasticity: All of Chapter 8. This reviews the White and Bruesch-Pagan test as well.
- Non-normality functional forms: Chapter 2, section 4; Chapter 6, section 6.2
- Mediation versus Moderation and interaction terms: Chapter 7, section 7.4
- Coefficient comparison and chow test, hierarchical models: Chapter 7, section 7.4; Chapter 4 (entire) and especially sections 4.2 and 4.5

Readings: Conley, D. "Capital for College: Parental Assets and Postsecondary Schooling." *Sociology of Education*. 2001.

Conley, D., R. Glauber, S. Olasky. 2004. "Sibling Similarity and Difference in Socioeconomic Status." Institute for Research on Poverty Working Paper, University of Wisconsin at Madison.

DATA ASSIGNMENT #1: Data detective: How did the percentage of marriages with a taller woman than man change over a recent 17 year period?

Week Two: Specification Error

January 27

- I. Unobserved Variable Bias, Selection and Endogeneity
- II. Ramsey Reset Test
- III. Random Effects and Fixed Effects Models:
 - a. Breusch-Pagan
 - b. Hausman Test

Text:

- Review chapters 4, 6, 8; read Chapters 9, 13, 14. Chapter 9 is about specification errors in general and Chapter 14 deals explicitly with fixed and random effects models. Chapter 9 presents Ramsey RESET test for functional form misspecification and unobserved variable bias. Breusch-Pagan is presented in chapter 8.

Week Three: Time Series Analysis

February 3

- I. Seasons and Trends
- II. Pooled Analysis
- III. Autocorrelation
- IV. Unit Roots
- V. Lagged Dependent Variables

Readings: Conley, D. and K. Springer. 2001. "Welfare State and Infant Mortality." *American Journal of Sociology*. 107:768-807.

Text:

- Chapters 10, 11, 12, and 18, and also see Chapter 9 Section 9.2 for introduction to lagged dependent variables.

Week Four: Introduction to Fixed Effects

February 10

Readings: Conley, D. and N.G. Bennett. 2000. "Is Biology Destiny? Birth Weight and Life Chances." *American Sociological Review*. 65:458-467.

Guo, G., L. VanWey. 1999. "Sibship Size and Intellectual Development: Is the Relationship Causal?" *American Sociological Review*. 64:169-187.

Figlio, D. "Names, Expectations and the Black-White Test Score Gap." Mimeo, Department of Economics, University of Florida.

Text:

- Review chapters 13 and 14

**Week Five: Difference-in-Difference-in-Difference: Extensions of Fixed Effects
(including applications to time series)**

February 17

Readings: J. Gruber. “The Incidence of Mandated Maternity Benefits.” *The American Economic Review*. 84:622-641.

O. Ashenfelter and C. Rouse. 1997. “Income, Schooling and Ability: Evidence from a New Sample of Identical Twins.” NBER Working Paper w6106. published in *Quarterly Journal of Economics*. 113:253-284.

J. Bound and G. Solon, “Double Trouble: On the Value of Twins-Based Estimation of the Returns to Schooling.” NBER Working Paper w6721; published as *EEDR* 1999. 18:169-182.

Conley, D., K. Strully and N.G. Bennett. 2003. “A Pound of Flesh or Just Proxy? Using Twin Differences to Estimate the Effect of Birth Weight on Life Chances.” NBER Working Paper w9901.

Text:

- Chapter 13 especially section 13.2

DATA ASSIGNMENT #2: What is the likelihood that an African American living in a low wealth family (bottom quartile) in 1984 will remain in that bottom quartile as an adult?

**Week Six: Instrumental Variable Approaches
February 24**

- I. Finding an Instrument
- II. Two-stage least squares

Readings: Conley, D., R. Glauber. “Parental Educational Investment and Children's Academic Risk: Estimates of the Impact of Sibship Size and Birth Order from Exogenous Variation in Fertility.” Working Paper, CASSR.

Angrist, J. 1990. “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records.” *The American*

Economic Review. 80:313-336.

Figlio, D. 2003. "Boys Named Sue: Disruptive Children and their Peers."
Mimeo, Department of Economics, University of Florida.

Text:

- Chapters 15 and 16

Week Seven: IVs Continued: The Holy Grail Beyond Reach
March 3

- I. Weak Instruments
- II. Violations of the Excludability Assumption

Readings: Angrist, J. and A. Krueger. 1994. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*. NBER Working Paper No. w3572

Bound, J. and D.A. Jaeger. 1994. "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data", *Journal of Labor Economics*, Vol. 12. NBER Working Paper No. w5835

M. Rosenzweig and K. Wolpin. 2000. "Natural 'Natural Experiments' in Economics" *Journal of Economic Literature*. 38:827-874.

Lleras-Muney A. "The Relationship between Education and Adult Mortality in the United States." NBER Working Paper No. w8986.

Text:

- Review Chapter 15 – both topics are covered in this chapter.

Week Eight: Regression Discontinuity
March 10

- I. Cutoff Criterion
- II. Visual Examination
- III. Pre- Post- Analysis

Readings: W. van der Klaauw. 2002. "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach." *International Economic Review*.

Week Nine: A Taste of Semi- and Non-parametric Estimation
March 24

- I. Bootstrapping
- II. Jackknife
- III. Extreme Bounds
- IV. Monte Carlo Methods
- V. Regression Breaks: the BIC test
- VI. Spline Estimates

Readings: Appendix of Conley, D. and K. Springer. 2001. "Welfare State and Infant Mortality." *American Journal of Sociology*. 107:768-807.

Conley, D., K. Strully and N. Bennett. 2004. "Twin Differences in Birth Weight: Estimating the Effects of Genotype and Prenatal Environment on Neonatal and Post-neonatal Mortality." *Journal of Economics and Human Biology*.

Week 10: Experimental Design
March 31

- I. Power Analysis
- II. Field Experiment Considerations

Viewings: <http://www.streamingmeeting.com/webmeeting/matrixvideo/nber/index.html>
(Parts I & II only)

YOUR TURN:

STUDENT PAPER CRITIQUES / WORKSHOPPING / PROJECT PRESENTATIONS

April 7, 14, 21, 28