Regression III: Advanced Methods
Syllabus - Summer 2009

Instructor: Dave Armstrong
Until August: Department of Politics & IR and Nuffield College (Oxford)
From September: Department of Political Science (UW-Milwaukee)

e: davearmstrong.ps@gmail.com
Course website: http://www.quantoid.net/ICPSR.php

Teaching Assistant: Matthew Painter
Department of Sociology
Ohio State University

e: painter.63@sociology.osu.edu

Overview and Course Objectives

The Regression III course takes a considerably different form than the first two regression courses at the Summer Program. This course will hopefully prepare you for the things you will encounter as you (attempt to) publish quantitative work with linear models. Initial linear model classes focus on the assumptions and theoretical considerations of linear models and generally walk you through estimation and interpretation. Good courses also deal with diagnostics, though these often get less time than they should. Further, it is not always obvious what violations of these assumptions will lead to in practical terms. This course will provide you with a systematic approach to assessing, fixing and presenting your linear model results. Though we focus almost exclusively on the linear model (we will allude to nonlinear models occasionally), the logic we follow will be helpful in dealing with nonlinear models as well.

This is a class that deals exclusively with observational data - those not collected in experimentally controlled environments. As such, we will spend little time on ANOVA and no time at all talking about concerns that are specific to the analysis of experimental data.

Requirements

This course is a practical, data-analytic extension of what you learned in your department's linear models class or the Regression II class at the ICPSR Summer Program. As such, I assume you are familiar with the types of things taught in these courses - Gauss-Markov assumptions, properties of OLS estimators, and statistical inference for linear

model coefficients. While I assume this knowledge exists, I will spend time reviewing these results briefly in class. If you are not sure where you belong in the series of linear models courses at the Summer Program, please see me or the Summer Program director and we will make sure you end up the most appropriate class.

Participants being formally assessed in this course (i.e., those taking the course for University of Michigan credit or those wishing to receive a letter indicating successful completion of the course) will be required to attend all lectures and complete all homework assignments. Further, these individuals should notify me if they are being formally assessed, especially those taking the class for UM credit.

For reasons that will likely become apparent rather quickly, computation will be done exclusively in the **R** statistical computing environment. This is not optional and implies two things:

- We will spend time going over R-code to generate the results we talk about in class. The class will be theoretically driven, but will have a considerable amount of associated computer work. If you are looking for a purely theoretical class, this is very likely not the one for you.
- It is essential for you to know or be learning the **R** language. Many of the results we generate in class are either impossible or (more likely) considerably more difficult to generate in many other general purpose statistical software packages (e.g., Stata and SPSS). This year, the Summer Program has split it's education in **R** into two separate courses. There is an introductory course for those with little to no background in **R** that will be taught by me from the 23rd to the 30th of June. For those already familiar with **R**, John Fox is teaching a workshop called "Programming in **R**," which will be held from the 23rd of June to the 3rd of July. John is probably one of the smartest **R** people you will ever meet. Even if you know **R** already, I am sure that you would learn many new and useful tools for programming in **R** and I would encourage you in the strongest possible way to take his course.

Obviously, those not being formally assessed in the course can attend without making a commitment to learn **R**. I have no problem with this, but please be aware that this will have a considerable (negative) impact on your ability to employ some of these techniques when you go back to your home institution. We will have lab classes that will allow you to get some hands-on practice with the results we generate in class.

If you're one of those "glutton-for-punishment" types, you may also find it useful to learn LaTeX. LaTeXis a system for typesetting documents. People find it most useful for typesetting documents that are heavy on mathematical notation, but this is just the tip of the iceberg. LaTeXhas its own bibliographic software (BibTeX) and will automatically build (and re-build) tables of contents, lists of figures and lists of tables. It also automatically numbers (and re-numbers when necessary) tables, figures and equations, changing

appropriately formed references to those objects when table, figure or equation numbers change. Best of all, the most LaTeXtypesetting engines (I use MikTeX) are free as are a number of powerful text editors that serve as front-ends to this software. Everything I present in class is written in LaTeX; specifically, the slides are all made with a package called "Beamer". Further, there are those who see LaTeXas a sort of secret handshake for nerds. So, if you want to be one of the "cool" kids, then you should definitely give it a shot.

Course Text(s)

No one text effectively presents all of the material that will be covered in this course. That said, much of the material is covered in:

- Fox, John. (2008) Applied Regression Analysis and Generalized Linear Model 2nd ed. Thousand Oaks, CA: Sage Publications, Inc
- Fox, John. (2002) An **R** and S-PLUS Companion to Applied Regression. Thousand Oaks, CA: Sage Publications, Inc.

The **R** and S-PLUS Companion is a great book for those currently learning **R**. This is widely recognized as one of the best ways for Social Scientists to get into **R**. The Applied Regression book is a great general purpose regression book. Much of what we talk about will be covered in other regression books. If you've got a particular favorite, then it might be worth supplementing your reading from your chosen regression book with pieces from the Fox book that are not covered by your favorite. Some books that I think are pretty good (depending on your orientation toward visualization, etc...) are:

- Gujarati, Damodar N. (2002) Basic Econometrics. 4th ed. New York: McGraw Hill/Irwin.
- Wooldridge, Jeffrey M. (2005) Introductory Econometrics. 3rd ed. Mason, OH: Southwestern.
- Cook, R. Dennis and Sanford Weisberg. (1999) Applied Regression Including Computing and Graphics. New York: John Wiley & Sons, Inc.

We will also use a number of other books and articles to deal with more specialized issues. These are listed below (along with the appropriate chapters/pages) for the classes in which we use them.

Software

One of **R**'s main virtues from the grad-student point of view is that the base package and all of the add-ons (called libraries in **R**) are free. You can download the basepackage of **R** from the Comprehensive **R** Archive Network (CRAN) `http://www.cran.r-project.org`.

As of this writing, the most recent version is 2.9.0. **R** is updated a couple of times per year so you'll have to look back here periodically for updates. We will be using a number of add-on libraries, including:

boot Bootstrapping techniques associated with Davison and Hinkley (1997).

car Regression diagnostics and other procedures associated with Fox (2002).

effects Creates effect displays for GLMs according to Fox (2003) and Fox and Andersen (2006).

foreign Imports data from SPSS, Stata and Sas

gam Functions for generalized additive models written by Trevor Hastie in the spirit of Hastie and Tibshirani (1990)

leaps Methods for subset selection following Miller (2002)

MASS Functions supporting Venables and Ripley (2002)

mgcv Methods for generalized additive models following Wood (2006).

nlme and lme4 Methods for linear and non-linear mixed effects models following Pinheiro and Bates (2000)

qvcalc package to calculate and present quasi-variances following Firth (2003) and Firth and Menzes (2004)

Rcmdr John Fox's GUI front end for **R**.

relimp Package written by David Firth to implement procedures in Silber, Rosenbaum and Ross (1995).

robustbase Package to implement robust regression, as in Yohai (1987).

rpanel User-friendly front-end to TCL/TK widgets.

RWinEdt Package that creates links between **R** and WinEdt.

scatterplot3d Tools for creating 3-dimensional scatterplots.

splines Package for creating smoothing splines.

Many of the above packages are included with the base distribution of **R**. However, some must be downloaded separately from CRAN. The easiest way to do this is first install the base package and then use the menus within **R** to automatically download and install required packages:

Packages → Install package(s) from CRAN

Related Software

A good text editor is invaluable when using **R** and LaTeX. I use WinEDT for both of these tasks. This comes free for 30 days and then with increasing frequency provides quite annoying reminders to buy the software. In fact, if you let it go long enough, every time you hit a key, the message will pop up. For students, the license is 30*and*40 for educational purchasers. This can be downloaded from `http://www.winedt.com`.

Emacs is another useful editor that can be used for both LaTeXand **R**. You can get information on how to install the **R** implementation of this (Emacs Speaks Statistics), from `http://ess.r-project.org/`. If you follow the directions, you should be able to get this up and running in relatively little time. There are other editors and you're certainly welcomed to use any other one you want, but these are the ones I know.

Course Schedule

Each entry represents a single class meeting. Readings are designated either as suggested (*) or supplemental (-). For most of you, this is not the only class you are taking and as the weeks fly by, your time will undoubtedly be too limited to read everything indicated in the syllabus. However, this should serve as a nice reference to which you can return if the intricacies of a particular topic have faded from your memory.

1. Preliminary Material (Tuesday, June 23)

   - Introduction and Goals for the course
   - Getting started with R.

   Readings:
   *Fox (2008) Chapters 1&2
   *Fox (2002) Chapters 1&2
   -Venables and Ripley (2002) Chaters 1-3

2. OLS 1: The Basics of Least Squares Regression (Wednesday, June 24)

   - Least-squares fit
   - Properties of the least-squares estimator
   - Statistical inference
   - Regression in matrix form

   Readings:
   *Clarke (2005)
   *Gill (1999)
   *Fox (2008) Chapters 5, 6 & 9
   *Fox (2002) Chapter 4
   -Abbott (1998, 1988), Achen (1990), Lewis-Beck and Skalaban (1990)

3. Graphics (Thursday, June 25)

   - Traditional and Lattice Graphics in R.
   - Types of R graphs
   - Graphical elements
   - Building R graphs

   Readings:
   *Fox (2002) Chapter 7
   *Jacoby (1997, 1998, 2006)
   *Kastellec and Leoni (2007)
   *Murrell (2006) Chapters 1-4
   -Sarkar (2008), Venables and Ripley (2002) Chapter 4

4. OLS II: Effective Presentation (Friday, June 26)

   - Factors and contrsts; quasi-variances and graphical displays for dummy regressors
   - Fitted values, interactions and effect displays
   - Standardization and relative importance

   Readings:
   *Brambor, Clark and Golder (2006)
   *Braumoeller (2004)
   *Firth (2003)
   *Silber, Rosenbaum and Ross (1995)
   *Firth and Menzes (2004), Kam and Franzese (2007)

5. Diagnostics I: Linearity (Monday, June 29)

   - Diagnosing linearity through residual plots
   - Fixing non-linearity with data transformations
   - Linearity and ordinal variables

   Readings:
   *Fox (2008) Chapters 4 & 12 (sec 12.3-12.5)
   *Fox (2002) Chapter 3
   *Jacoby (1999)
   -Cook and Weisberg (1999) Chapter 16, Box and Tidwell (1962)

6. Non-Linearity, Smoothing and Splines (Tuesday, June 30)

    - Nonparametric Smoothing - Lowess
    - Inference for regression smoothers
    - Regression Splines

    Readings:
    *Fox (2008) Chapters 17 & 18
    *Keele (2008)
    -Fox (2000*b*,*a*)

7. Generalized Additive Models (Wednesday, July 1)

    - Estimation and Backfitting
    - Degrees of freedom
    - Cross-validation for smoothing parameters
    - Diagnostics

    Readings:
    *Fox (2000*a*)
    *Keele (2008) Chapters 4-6
    -Hastie and Tibshirani (1990), Wood (2006)

8. Lab I: (Thursday, July 2)

    - Non-linearity transformations
    - Smoothers and splines
    - Generalized additive models

9. Diagnostics II: Outliers and Influential Data (Friday, July 3)

    - Outliers, leverage and influential data
    - Hat values, standardized residuals, Cook's D

    Readings: *Fox (2008) Chapter 11
    *Fox (2002) Chapter 6
    *Cook and Weisberg (1999) Chapter 15
    -Jasso (1985, 1996), Kahn and Udry (1986)

10. Fixing Outliers and Influential Data: Robust Regression (Monday, July 6)

    - Breakdown point, influence function and various types of robust regression
    - M-estimation (and extension) and interative reweighted least squares
    - Diagnostics for outliers revisited
    - Robust GLMs

    Readings:
    *Andersen (2008)
    *Fox (2008) Chapter 19
    -Cantoni and Ronchetti (2001), Rousseeuw and Leroy (1987)

11. Diagnostics III: Non-constant error variance and collinearity (Tuesday, July 7)

    - Residual plots
    - ML transformations of Y
    - Weighted least squares
    - Heteroskedastic linear regression
    - Robust standard errors

    Readings: *Fox (2008) Chapters 12 & 13
    *Fox (2002) Chapters 3 & 6
    *Harvey (1976)
    *Long and Ervin (2000)
    -Cook and Weisberg (1999) Chapter 14

12. Model Selection (Wednesday, July 8)

    - Theoretical issues in model searching and post-data model construction
    - Model selection criteria and multi-model inference.
    - Subset selection models

    Readings:
    *Burnham and Anderson (2004)
    *Fox (2008) Chapter 22
    *Leamer (1974, 1983), Leamer and Leonard (1983)
    *Mallows (1974)
    *Miller (1984)

13. Bootstrapping and other Resampling Schemes for Regression (Thursday, July 9)

    - Bootstrapping and Jack-knifing
    - Corss-validation

    Readings:
    *Fox (2008) Chapter 21
    -Davison and Hinkley (1997)
    -Efron and Tibshirani (1993)

14. Lab II (Friday, July 10)

    - Outliers and Robust Regression
    - Heteroskedasticity
    - Model Selection
    - Resampling

15. Handling Dependent Data (Monday, July 13)

    - Mixed-effects models for clustered and longitudinal data
    - Robust standard errors revisited

    Readings:
    *Snijders and Bosker (1999)Chapters 1-5
    *Steenbergen and Jones (2002)
    -Raudenbush and Bryk (2002)Chapters 1,2 & 4
    -Pinheiro and Bates (2000)

16. Multilevel Change Models [Matthew Painter] (Tuesday, July 14)

    Readings:
    *Bliese and Ployhart (2002)
    *Raudenbush and Bryk (2002) Chapter 6
    *Singer and Willett (2003) Part 1 (especially Chapters 1-5)
    -Barnes et al. (2006), Downey, vonHipple and Beckett (2004), Curran (2003), Kim and Sakamoto (2008)

17. Fixed- and Random-effects Models for TSCS Data (Wednesday, July 15)

    - Fixed Effects assumptions
    - Random effects assumptions
    - Choosing between fixed- and random-effects

    Readings:
    *Wooldridge (2002) Chapters 10&11
    -Hsiao (1986)

18. Finite Mixture Models for Unobserved Heterogeneity (Thursday, July 16)

    - Estimating finite mixtures
    - Dealing with problems of label switching
    - Diagnosing and interpreting models

    Readings:
    *Grün and Leisch (2006*b*)
    *Grün and Leisch (2006*a*)
    *Leisch (2004)

19. Lab III (Friday, July 18)

    - Models for Dependent data
    - Models for TSCS data
    - Mixture Models

References

Abbott, Andrew. 1988. "Transcending General Linear Reality." *Sociological Theory* 6(2):169–186.

Abbott, Andrew. 1998. "The Causal Devolution." *Sociological Methods and Research* 27:148–181.

Achen, Christopher H. 1990. "What Does "Explained Variance" Explain?: Reply." *Political Analysis* 2(1):173–184.

Andersen, Robert. 2008. *Modern Methods for Robust Regression*. Thousand Oaks, CA: Sage.

Barnes, Grace M, Joseph H. Hoffman, John W Welte, Michael P Farrell and Barbara A Dintcheff. 2006. "Effects of Parental Monitoring and Peer Deviance on Substance Use and Delinquency." *Jounral of Marriage and Family* 68:1084–1104.

Bliese, Paul D. and Robert E. Ployhart. 2002. "Growth Modeling Using Random Coefficient Models: Model Building, Testing and Illustrations." *Organizational Research Methods* 5(4):362–387.

Box, George and P.W. Tidwell. 1962. "Transformation of the Independent Variables." *Technometrics* 4:531–550.

Brambor, Thomas, William Clark and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14(1):63–82.

Braumoeller, Bear F. 2004. "Hypothesis Testing and Multiplicative Interaction Terms." *International Organization* 58(4):807–820.

Burnham, Kenneth P. and David R. Anderson. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods and Research* 33(2):261–304.

Cantoni, Gustavo E. and Elvezio Ronchetti. 2001. "Robust Inference for Generalized Linear Models." *Journal of the American Statistical Association* 96:1022–1030.

Clarke, Kevin. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22(4):341–352.

Cook, R. Dennis and Sanford Weisberg. 1999. *Applied Regression Including Computing and Graphics*. New York: Wiley & Sons, Inc.

Curran, Patrick J. 2003. "Have Multilevel Models Been Structural Equation Models All Along?" *Multivariate Behavioral Research* 38(4):529–569.

Davison, Anthony C. and D.V. Hinkley. 1997. *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.

Downey, Douglas B., Paul T. vonHipple and Broh A. Beckett. 2004. "Are Schools the Great Equalizer? Cognitive Inequality During the Summer Months and the School Year." *American Sociological Review* 69(5):613–635.

Efron, Bradley and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Firth, David. 2003. "Overcoming the Reference Category Problem in the Presentation of Statistical Models." *Sociological Methodology* 33:1–18.

Firth, David and Renee X. De Menzes. 2004. "Quasi-Variances." *Biometrika* 91(1):65–80.

Fox, John. 2000*a*. *Multiple and Generalized Nonparametric Regression*. Thousand Oaks: Sage.

Fox, John. 2000*b*. *Nonparametric Simple Regression*. Thousand Oaks: Sage.

Fox, John. 2002. *An R and S-Plus Companion to Applied Regression*. Thousand Oaks: Sage Publications.

Fox, John. 2003. "Effect Displays in R for Generalised Linear Models." *Journal of Statistical Software* 8(15):1–27.

Fox, John. 2008. *Applied Regression Analysis and Generalized Linear Models,* 2^{nd} *edition*. Thousand Oaks, CA: Sage, Inc.

Fox, John and Robert Andersen. 2006. "Effect Displays for Multinomial and Proportional-Odds Logit Models." *Sociological Methodology* 36:225–255.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3):647–674.

Grün, Bettina and Friedrich Leisch. 2006*a*. Finite Mixture Model Diagnostics Using the Parametric Bootstrap. In *Proceedings of the Junior Scientist Conference 2006*, ed. Wilfried Elmenreich and Hans Kaiser. Vienna University of Technology Vienna, Austria: pp. 301–302.

Grün, Bettina and Friedrich Leisch. 2006*b*. Fitting Finite Mixtures of Linear Regression Models with Varying & Fixed Effects in R. In *Compstat 2006—Proceedings in Computational Statistics*, ed. Alfredo Rizzi and Maurizio Vichi. Physica Verlag, Heidelberg, Germany pp. 853–860.

Harvey, Andrew C. 1976. "Estimating Regression Models with Multiplicative Heteroskedasticity." *Econometrica* 44(3):461–465.

Hastie, Trevor J. and Robert J. Tibshirani. 1990. *Generalized Additive Models*. Boca Raton, FL: Chapman & Hall/CRC.

Hsiao, Cheng. 1986. *Analysis of Panel Data*. Cambridge.

Jacoby, William G. 1997. *Statistical Graphics for Univariate and Bivariate Data*. Thousand Oaks, CA: Sage.

Jacoby, William G. 1998. *Statistical Graphics for Visualizing Multivariate Data*. Thousand Oaks, CA: Sage.

Jacoby, William G. 1999. "Levels of Measurement and Political Research: An Optimistic View." *American Journal of Political Science* 43(1):271–301.

Jacoby, William G. 2006. "The Dot Plot: A Graphical Display for Labeled QUantitative Values." *The Political Methodologist* 14(1):6–14.

Jasso, Guillermina. 1985. "Marital Coital Frequency and the Passage of Time: Estimating the Separate Effects of Spouses' Ages and Marital Duration, Birth and Marriage Cohorts, and Period Influences." *American Sociological Review* 50(2):224–241.

Jasso, Guillermina. 1996. "Is It Outlier Deletion or is it Sample Truncation? Notes on Science and Sexuality." *American Sociological Review* 51(5):738–742.

Kahn, Joan R. and J. Richard Udry. 1986. "Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions." *American Sociological Review* 51(5):734–737.

Kam, Cindy and Robert J. Franzese. 2007. *Modeling and Interpreting Interactive Hypotheses in Regression Analyses*. Ann Arbor: University of Michigan Press.

Kastellec, Jonathan P and Eduardo L. Leoni. 2007. "Using Graphs Instead of Tables in Political Science." *Perspectives on Politics* 5(4):755–771.

Keele, Luke J. 2008. *Semiparametric Regression for the Social Sciences*. New York: Wiley & Sons, Inc.

Kim, ChangHwan and Arthur Sakamoto. 2008. "The Rise of Intra-Operational Wage Inequality in the United States, 1983-2002." *American Sociological Review* 73(1):129–157.

Leamer, Edward E. 1974. "False Models and Post-Data Model Construction." *Journal of the American Statistical Association* 69(345):122–131.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *The American Economic Review* 73(1):31–43.

Leamer, Edward E. and Herman Leonard. 1983. "Reporting the Fragility of Regression Estimates." *The Review of Economics and Statistics* 65(2):306–317.

Leisch, Friedrich. 2004. "FlexMix: A general framework for finite mixture models and latent class regression in R." *Journal of Statistical Software* 11(8):1–18.
URL: *http://www.jstatsoft.org/v11/i08/*

Lewis-Beck, Michael S. and Andrew Skalaban. 1990. "The R-squared: Some Straight Talk." *Political Analysis* 2(1):153–171.

Long, J. Scott and Laurie H. Ervin. 2000. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model." *The American Statistician* 54(3):217–224.

Mallows, Colin L. 1974. "Some Comments on C_p." *Technometrics* 15:661–675.

Miller, Alan. 2002. *Subset Selection in Regression,* 2^{nd} *edition*. Boca Raton, FL: Chapman & Hall/CRC.

Miller, Alan J. 1984. "Selection of Subsets of Regression Variables." *Journal of the Royal Statistical Society, Series A* 147(3):389–425.

Murrell, Paul. 2006. *R Graphics*. Boca Raton, FL: Chapman & Hall/CRC.

Pinheiro, Jose C. and Douglas M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.

Raudenbush, Stephen and Anthony Bryk. 2002. *Hierarchical Linear Models*. Thousand Oaks, CA: Sage.

Rousseeuw, Peter J and Annick M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: Wiley & Sons, Inc.

Sarkar, Deepayan. 2008. *Lattice : multivariate data visualization with R*. Use R New York; London: Springer Science+Business Media.

Silber, Jeffrey H., Paul R. Rosenbaum and Richard N. Ross. 1995. "Comparing the Contributions of Groups of Predictors: Which Outcomes Vary with Hospital Rather Than Patient Characteristics." *Journal of the American Statistical Association* 90(429):7–18.

Singer, Judith D. and John B Willett. 2003. *Applied Longitudinal Data Analysis*. New York: Oxford University Press.

Snijders, Tom A. B. and Roel Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage.

Steenbergen, Marco and Bradford S. Jones. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46(1):218–237.

Venables, William N. and Brian D. Ripley. 2002. *Modern Applied Statistics with S,* 4^{th} *edition*. 3 ed. New York: Springer.

Wood, Simon. 2006. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.

Wooldridge, Jeffrey. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Yohai, Victor J. 1987. "High Breakdown Point and High Efficiency Robust Estimates for Regression." *Annals of Statistics* 15:642–656.