

Regression III: Advanced Methods Syllabus - Summer 2011

Instructor: Dave Armstrong
Department of Political Science
University of Wisconsin - Milwaukee
e: davearmstrong.ps@gmail.com
Course website: <http://www.quantoid.net/ICPSR.php>

Teaching Assistant: Kelly Gleason
Department of Political Science
University of Wisconsin - Milwaukee
e: kgleason@uwm.edu
w: <http://www.uwm.edu/~kgleason>

1 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 when 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.

2 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 ideas 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.

This course will employ the R statistical computing environment exclusively. This is not optional. This means two things for the course:

1. We will spend time going over R-code used 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.
2. 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). If you are participating in both sessions of the program, you could take the R workshop taught in the first session (by John Fox, one of the smartest R people you will probably ever meet) or the one taught in the second session (by me).

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 . \LaTeX is 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. \LaTeX has 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, common \LaTeX typesetting engines are free (see <http://www.latex-project.org/ftp.html> for links to the software appropriate for your OS). Everything I present in class is written in \LaTeX ; specifically, the slides are all made with a package called "Beamer". There are some nice literate programming tools (Sweave) that integrate \LaTeX and R as well. Further, there are those who see \LaTeX as a sort of secret handshake for nerds. So, if you want to be one of the "cool" kids, then you should definitely try it; everyone else is doing it.

3 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. (2011) An R and S-PLUS Companion to Applied Regression. *2nd ed.* Thousand Oaks, CA: Sage Publications, Inc.

The R and S-PLUS Companion is a great book for those currently learning R. I would highly recommend getting the recently update and expanded second edition. 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: 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.

4 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 packages in R) are free. You can download the base package of R from the Comprehensive R Archive Network (CRAN) website <http://www.cran.r-project.org>. As of this writing, the most recent version is 2.13.0. 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 user-contributed packages that we will discuss as they become relevant.

4.1 Related Software

A good text editor is invaluable when using R and \LaTeX . \TeX Works is a good, free editor for \LaTeX that works in most environments, including Windows and Mac (<http://www.tug.org/texworks/>). RStudio is a free, recently-released IDE (Integrated Development Environment) for R that includes a nicely-featured text editor (<http://www.rstudio.org/>). There are a couple of pay options that are good general-purpose text editors for Mac and Windows that integrate nicely with \LaTeX and R (as well as a bunch of other languages) - WinEDT (<http://www.winedt.com/>), for Windows and TextMate (<http://macromates.com/>) are my favorites, but there are many other options as well.

5 Course Schedule

Each entry represents a single schedule. 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, July 19)

- (a) Goals for the course
- (b) Getting started with R.

Readings:

- * Fox (2008), Chapters 1 & 2
- * Fox (2002), Chapters 1 & 2
- Venables and Ripley (2002), Chapters 1-3

2. OLS I: The Basics of Least Squares Regression (Wednesday, July 20)

- (a) Least-squares fit
- (b) Properties of the least-squares estimator
- (c) Statistical inference
- (d) Regression in matrix form

Readings:

- * Fox (2008), Chapters 5, 6 & 9
- * Fox (2002), Chapter 4
- * Gill (1999)
 - Gelman and Stern (2006)
- * Clarke (2005)
 - Lewis-Beck and Skalaban (1990), Achen (1990), King (1990)

3. Graphics (Thursday, July 21)

- (a) Traditional, Lattice and Grid graphics
- (b) Types of R graphs
- (c) Graphical elements
- (d) Building R graphs

Readings:

- * Fox (2002) Chapter 7
- * Murrell (2006) Chapters 1-4
- * Jacoby (1997, 1998, 2006)
- * Kastelec and Leoni (2007)
 - Venables and Ripley (2002) Chapter 4

4. OLS II: Effective Presentation (Friday, July 22)

- (a) Factors and contrasts; quasi-variances and graphical displays
- (b) Fitted values, interactions and effect displays
- (c) Standardization and relative importance

Readings:

- * Armstrong (2011)
- * Firth (2003)
- * Brambor, Clark and Golder (2006)
- * Braumoeller (2004)
- Silber, Rosenbaum and Ross (1995) – Kam and Franzese (2007)
- Firth and Menzes (2004)

5. Re-sampling Techniques and Regression (Monday, July 25)

- (a) Bootstrapping and Jackknifing
- (b) Cross-validation

Readings:

- * Fox (2008) Chapter 21
- * Stone (1974)
- Efron and Tibshirani (1993)
- Davison and Hinkley (1997)
- Ronchetti, Field and Blanchard (1997)

6. Diagnostics I: Linearity (Tuesday, July 26)

- (a) Diagnosing linearity through residual plots
- (b) Fixing non-linearity with data transformations
- (c) Linearity and ordinal variables

Readings:

- * Fox (2008) Chapters 4 & 12 (Sections 12.3-12.5)
- * Fox (2002) Chapter 3
- * Jacoby (1999)
- Cook and Weisberg (1999) Chapter 16
- Box and Tidwell (1962)
- Breiman and Friedman (1985 a,b), Pregibon and Vardi (1985), Buja and Kass (1985), Fowlkes and Kettering (1985)

7. Non-Linearity, Smoothing and Splines (Wednesday, July 27)

- (a) Nonparametric Smoothing - Lowess
- (b) Inference for regression smoothers
- (c) Regression Splines

Readings:

- * Fox (2008) Chapters 17 & 18
- Fox (2000*b,a*)
- * Keele (2008) Chapters 2 & 3

8. Generalized Additive Models (Thursday, July 28)

- (a) Estimation and Backfitting
- (b) Degrees of freedom
- (c) Cross-validation for smoothing parameters
- (d) Diagnostics

Readings:

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

9. Lab I: (Friday, July 29)

- (a) Non-linearity transformations
- (b) Smoothers and splines
- (c) Generalized additive models

10. Diagnostics II: Outliers and Influential Data (Monday, August 1)

- (a) Outliers, leverage and influential data
- (b) Hat values, standardized residuals, Cook's D

Readings:

- * Fox (2008) Chapter 11
- * Fox (2002) Chapter 6 (pp 101-201)
- * Cook and Weisberg (1999) Chapter 15
- Jasso (1985, 1996), Kahn and Udry (1986)

11. Fixing Outliers and Influential Data: Robust Regression (Tuesday, August 2)

- (a) Breakdown point, influence function and various types of robust regression
- (b) M-estimation (and extension) and iterative re-weighted least squares
- (c) Diagnostics for outliers revisited
- (d) Robust GLMs

Readings:

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

12. Diagnostics III: Non-constant error variance and collinearity (Wednesday, August 3)

- (a) Residual plots
- (b) ML transformations of Y
- (c) Weighted least squares
- (d) Heteroskedastic linear regression
- (e) Robust standard errors

Readings:

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

13. Model Selection (Thursday, August 4)

- (a) Theoretical issues in model searching and post-data model construction
- (b) Model selection criteria and multi-model inference.
- (c) Subset selection models

Readings:

- * Fox (2008) Chapter 22
- * Leamer (1983)
- * Leamer and Leonard (1983)
- * Box (1976), Box and Hunter (1962)
- Freedman (1991*b,a*), Berk (1991), Blalock (1991), Mason (1991)
- Miller (2002), Breiman (1992), Breiman and Spector (1992)
- * Burnham and Anderson (2004)

14. Lab II (Friday, August 5)

- (a) Outliers and Robust Regression
- (b) Heteroskedasticity
- (c) Model Selection

15. Missing Data and Multiple Imputation (Monday, August 8)

- (a) Whats the problem with missing data?
- (b) When can we fix it?
- (c) How do we impute the data and use those imputations?

Readings:

- * Mcknight et al. (2007)
- * Allison (2001)

16. Random Forests Regression (Tuesday, August 9)

- (a) Classification by Random Forests.
- (b) Regression by Random Forests.

Readings:

- * Liaw and Wiener (2002)
- Breiman (2001)
- Breiman (1996)

17. Finite Mixture Models (Wednesday, August 10)

Readings:

- * Imai and Tingley (2011)
- * Grün and Leisch (2008a)
- * Grün and Leisch (2008b)
- * Grün and Leisch (2006)

18. Quick Intro to Spatial Data Analysis (Thursday, August 11)

- (a) What are the main problems with regression of spatially organized data
- (b) Assessing the extent to which geography causes problems
- (c) Estimating regression models for spatially organized data

Readings

- * Bivand, Pebesma and Gomez-Rubio (2008) Chapters 1, 3, 8-10

19. Lab III (Friday, August 12)

- (a) Random Forests
- (b) Mixture Models
- (c) Missing Data and Multiple Imputation
- (d) Spatial Regression Models

References

- Achen, Christopher H. 1990. "What Does "Explained Variance" Explain?: Reply." *Political Analysis* 2(1):173–184.
- Allison, Paul D. 2001. *Missing Data*. Thousand Oaks, CA: Sage.
- Andersen, Robert. 2008. *Modern Methods for Robust Regression*. Thousand Oaks, CA: Sage.
- Armstrong, David A. 2011. "factorplot: Improving Presentation of Simple Contrasts in GLMs." Working paper, available: <http://www.quantoid.net/factorplot.pdf>.
- Berk, Richard A. 1991. "Toward a Methodology for Mere Mortals." *Sociological Methodology* 21:315–324.
- Bivand, Roger S., Edzer J. Pebesma and Virgilio Gomez-Rubio. 2008. *Applied spatial data analysis with R*. Use R New York; London: Springer.
- Blalock, Hubert M. 1991. "Are There Really Any Constructive Alternatives to Causal Modeling?" *Sociological Methodology* 21:325–335.
- Box, George E. P. 1976. "Science and Statistics." *Journal of the American Statistical Association* 71(356):791–799.
- Box, George E. P. and William G. Hunter. 1962. "A Useful Method for Model-Building." *Technometrics* 4(3):301–318.
- 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.
- Breiman, Leo. 1992. "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error." *Journal of the American Statistical Association* 87(419):738–754.
- Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24(2):123–140.
- Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.
- Breiman, Leo and Jerome H. Friedman. 1985a. "Estimating Optimal Transformations for Multiple Regression and Correlation." *Journal of the American Statistical Association* 80(391):580–598.

- Breiman, Leo and Jerome H. Friedman. 1985*b*. “Estimating Optimal Transformations for Multiple Regression and Correlation: Rejoinder.” *Journal of the American Statistical Association* 80(391):614–619.
- Breiman, Leo and Philip Spector. 1992. “Submodel Selection and Evaluation in Regression. The X-Random Case.” *International Statistical Review* 60(3):291–319.
- Buja, Andreas and Robert E. Kass. 1985. “Estimating Optimal Transformations for Multiple Regression and Correlation: Comment.” *Journal of the American Statistical Association* 80(391):602–607.
- 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.
- Davison, Anthony C. and D.V. Hinkley. 1997. *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.
- 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.
- Fowlkes, E.B. and J.R. Kettering. 1985. “Estimating Optimal Transformations for Multiple Regression and Correlation: Comment.” *Journal of the American Statistical Association* 80(391):607–613.
- 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. 2008. *Applied Regression Analysis and Generalized Linear Models, 2nd edition*. Thousand Oaks, CA: Sage, Inc.

- Freedman, David A. 1991a. "A Rejoinder to Berk, Blalock and Mason." *Sociological Methodology* 21:353–358.
- Freedman, David A. 1991b. "Statistical Models and Shoe Leather." *Sociological Methodology* 21:291–313.
- Gelman, Andrew and Hal Stern. 2006. "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant." *The American Statistician* 60(4):328–331.
- Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3):647–674.
- Grün, Bettina and Friedrich Leisch. 2006. 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. 2008a. Finite Mixtures of Generalized Linear Regression Models. In *Recent Advances in Linear Models and Related Areas*, ed. Shalabh and Christian Heumann. Physica Verlag.
- Grün, Bettina and Friedrich Leisch. 2008b. "FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters." *Journal of Statistical Software* 28(4):1–35.
URL: <http://www.jstatsoft.org/v28/i04>
- Harvey, Andrew C. 1976. "Estimating Regression Models with Multiplicative Heteroskedasticity." *Econometrica* 44(3):461–465.
- Hastie, Trevor and Robert Tibshirani. 1990. *Generalized Additive Models*. New York: Chapman and Hall.
- Imai, Kosuke and Dustin Tingley. 2011. "A Statistical Method for Empirical Testing of Competing Theories." Working paper, available: <http://imai.princeton.edu/research/index.html>.
- Jacoby, William. 1997. *Statistical Graphics for Univariate and Bivariate Data*. Thousand Oaks, CA: Sage Publications.
- 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.
- King, Gary. 1990. "Stochastic Variation: A Comment on Lewis-Beck and Skalaban's 'The R-Squared'." *Political Analysis* 2(1):185–200.
- 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.
- Lewis-Beck, Michael S. and Andrew Skalaban. 1990. "The R-squared: Some Straight Talk." *Political Analysis* 2(1):153–171.
- Liaw, Andy and Matthew Wiener. 2002. "Classification and Regression by randomForest." *RNews* 2/3:18–22.
- 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.
- Mason, William M. 1991. "Freedman is Right as Far as He Goes, but There is More, and It's Worse. Statisticians Could Help." *Sociological Methodology* 21:337–357.
- McKnight, Patrick E., Katherine M. McKnight, Souraya Sidani and Aurelio Jose Figueredo. 2007. *Missing data a gentle introduction*. New York: Guilford Press.
- Miller, Alan. 2002. *Subset Selection in Regression, 2nd edition*. Boca Raton, FL: Chapman & Hall/CRC.
- Murrell, Paul. 2006. *R Graphics*. Boca Raton, FL: Chapman & Hall/CRC.

- Pregibon, Daryl and Yehuda Vardi. 1985. “Estimating Optimal Transformations for Multiple Regression and Correlation: Comment.” *Journal of the American Statistical Association* 80(391):598–601.
- Ronchetti, Elvezio, Christopher Field and Wade Blanchard. 1997. “Robust Linear Model Selection by Cross-validation.” *Journal of the American Statistical Association* 92(439):1017–1023.
- Rousseeuw, Peter J and Annick M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: Wiley & Sons, Inc.
- 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.
- Stone, Mervyn. 1974. “Cross-validation and Assessment of Statistical Predictions.” *Journal of the Royal Statistical Society, Series B* 36(2):111–147.
- Venables, W.N. and B.D. Ripley. 2002. *Modern Applied Statistics with S-PLUS*. 4 ed. New York: Springer.
- Wood, Simon. 2006. *Generalized Additive Models: An Introduction with R*. London: Chapman and Hall/CRC.