

# 2005 ICPSR SUMMER PROGRAM

## REGRESSION ANALYSIS III: ADVANCED METHODS

William G. Jacoby  
Department of Political Science  
Michigan State University

June 27 – July 22, 2005

This course will take a modern, data-analytic approach to the multiple regression model. Our coverage of the material will emphasize the ways that graphical tools can augment traditional methods for describing how the conditional distribution of a dependent variable changes along with the values of one or more independent variables. The course will examine the basic nature and assumptions of the linear regression model, diagnostic tools for detecting violations of the regression assumptions, and strategies for dealing with situations in which the basic assumptions are violated. The overall goal is to provide: (1) New insights about regression analysis; (2) a general overview of various modern extensions to the traditional linear model; and (3) innovative, effective methods for presenting the results from statistical investigations of empirical data.

Specific topics to be covered include: data visualization and transformation; assumptions of the linear model; regression diagnostics and model assessment; robust and resistant regression; weighted least squares; generalized linear models, resampling methods; nonlinear regression; nonparametric regression; generalized additive models; and graphical regression.

### Prerequisites and Requirements

This workshop is one element in a track of advanced courses which also includes the workshops on *Maximum Likelihood Estimation for Generalized Linear Models*, *Bayesian Methods for the Social and Behavioral Sciences*, and *Advanced Topics in Maximum Likelihood Estimation*. These courses are integrated around the theme of cutting-edge methodological training for relatively advanced graduate students. All of these courses assume that you are already familiar with the **R** (or **S-Plus**) statistical computing environment or that you are enrolled in the course on *Statistical Computing Using R/S*.

The *Regression III* workshop also assumes a good “working knowledge” of regression analysis, including statistical inference. It would be best if you already have some experience with matrix algebra. However, concurrent enrollment in *Mathematics for Social Scientists II* should provide sufficient exposure to the matrix concepts and techniques that will be required for this workshop.

Regular class attendance is assumed, but not required. I assume that you will be keeping up with the reading assignments provided below. At the same time, however, I understand the time pressures that are inherent in a work-intensive environment like the ICPSR Summer Program. Therefore, I recognize that you will frequently be unable to complete all of the assigned readings prior to the relevant class meetings. Nevertheless, I do suggest that you read the assignments at some point, even if you have to do so after the end of the workshop.

As you probably already know, data analysis and statistical methods are not “spectator sports.” So, there will be regular homework assignments (approximately two per week). All of these exercises will involve computer-assisted data analysis and they should be completed using R or S-Plus. You should also write up your answers to the assignments using the L<sup>A</sup>T<sub>E</sub>X document preparation system.

## Course Textbooks

No single text presents all of the subject matter from this workshop. However, most of the material is covered in the following three books:

Cook, R.D. and S. Weisberg. (1999) *Applied Regression Including Computing and Graphics*. New York: John Wiley and Sons.

Fox, J. (1997) *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.

Fox, J. (2002) *An R and S-Plus Companion to Applied Regression*. Thousand Oaks, CA: Sage.

It is perhaps not necessary that you purchase all of these texts, but I highly recommend that you do so- you will find them all to be important and useful reference sources. Additional readings specific to particular lectures are given in the class schedule; complete citations are provided in the list of references at the end of this syllabus.

## Course Software

A prominent feature of modern statistical methodology is its computationally intensive nature. At the same time, it would be impossible to construct the graphical tools that we will use in this workshop using pencil and paper. For these reasons, statistical software will be particularly important for achieving the course objectives.

Most class examples and demonstrations will rely on the **R** computing environment (or its commercial alternative, **S-Plus**). Along with its superb functionality and flexibility, **R** is also attractive because it is available free of charge. There are two ways for you to obtain this software: First, you can download it from the R website:

`http://www.r-project.org`

This site also provides a great deal of useful information about R and many links to additional material (e.g., manuals, FAQ's, newsletter, etc.). Second, the ICPSR Summer Program will provide a CD-ROM containing the latest version of R for a minimal charge (to cover processing costs). This CD may also include a version of the L<sup>A</sup>T<sub>E</sub>X document preparation system and a few other utility programs.

Two important advantages of **R** are its extensible nature and active user community. Many researchers have written functions and procedures (called “packages”) that extend **R**'s capabilities far beyond the features that are available in its basic implementation. We will use a number of these packages throughout the workshop and they will be introduced as the need arises. All of the packages can be downloaded from the web and most of them are available at the *Comprehensive R Archive Network* (CRAN) site, or one of its mirror sites around the world.

Although most of our computing will be carried out in **R**, we will also use another program called **Arc**. This small, but powerful, software package can be downloaded from the following web site:

`http://www.stat.umn.edu/arc/`

**Arc** is a supplement to the Cook and Weisberg (1999) text that is mainly intended to operationalize the dynamic graphical methods for fitting and assessing regression models presented in that book.

## Course Web Site

The website for the *Regression III* workshop is located at:

`http://polisci.msu.edu/jacoby/icpsr/regress3/`

The contents of this website will evolve and expand as the workshop proceeds through the subject matter. It will contain the syllabus, handouts, datasets, assignments, computing materials, and links to other relevant sites on the Worldwide Web. Be sure to check this site frequently to obtain the latest information and materials.

# Class Schedule

Each entry in the schedule below represents a single lecture. The required readings for each topic are marked by an asterisk (\*). All other citations are supplemental materials (marked with a dash, '-').

## 1. Preliminary Material

- A. Goals of the course
- B. Getting started with **R**

### *Readings:*

- \* Fox (1997), Chapters 1 and 2
- \* Fox (2002), Chapters 1 and 2
- \* Cook and Weisberg (1999), Chapters 1 and 2
- Venables and Ripley (2002), Chapters 1-3

## 2. Examining Data

- A. Graphical displays
- B. Transformations

### *Readings:*

- \* Fox (1997), Chapters 3 and 4
- \* Fox (2002), Chapter 3
- \* Cook and Weisberg (1999), Chapters 3, 5, 8
- Jacoby (1997)
- Jacoby (1998)
- Cleveland (1993)

## 3. General Linear Models I: The Basics of Least Squares Regression

- A. Graphical fitting
- B. Least-squares fitting
- C. Properties of the least-squares estimator
- D. Statistical inference
- E. The regression model in matrix form

### *Readings:*

- \* Fox (1997), Chapters 5, 6, 9, Appendixes C and D
- \* Fox (2002), Chapter 4
- Cook and Weisberg (1999), Chapters 6-7

#### **4. General Linear Models II: The Geometry of the Regression Model**

- A. Vector geometry
- B. Vector representation of the regression model
- C. The data ellipsoid and model fit

*Readings:*

- \* Fox (1997), Chapter 10
- \* Monette, G. (1990)
- Cook and Weisberg (1999), Chapter 4

#### **5. General Linear Models III: Effective Presentation**

- A. Dummy regressors
- B. Fitted values, interactions, and effect displays
- C. Standardization and relative importance

*Readings:*

- \* Firth (2003)
- \* Fox (1987)
- \* Fox (1997), Chapter 7
- \* Silber, Rosenbaum, Ross (1995)
- \* Firth and Menezes (2004)

#### **6. Regression with Categorical Dependent Variables I**

- A. Limited dependent variables and problems with the OLS model
- B. Binary logit and probit models
- C. Fitted probabilities and effect displays

*Readings:*

- \* Fox (1997), Chapter 15 (pp 438-465, 487-489)
- \* Fox (2002), Chapter 5 (pp 155-158)
- Cook and Weisberg (1999), Chapter 21
- McCullagh and Nelder (1989)
- Long (1997), Chapters 3 and 4

#### **7. Regression with Categorical Dependent Variables II**

- A. Ordered probit and logit models
- B. Multinomial logit models
- C. Generalized linear models

- D. Poisson models for count data
- E. Diagnostics for generalized linear models

*Readings:*

- \* Fox (1997), Chapter 15 (pages 466-489)
- \* Fox (2002), Chapter 5 (pages 167-188), Chapter 6 (pages 225-233)
- Cook and Weisberg (1999), Chapters 22-23
- Long (1997), Chapters 5, 6, 8

## **8. Regression Diagnostics I: Unusual Observations**

- A. Outliers, leverage, and influence
- B. Hat values and studentized residuals
- C. Case-deletion statistics

*Readings:*

- \* Fox (1997), Chapter 11
- \* Fox (2002), Chapter 6 (pages 191-201)
- Cook and Weisberg (1999), Chapter 15

## **9. Regression Diagnostics II: Nonlinearity, Nonnormality, and Heteroskedasticity**

- A. Residual plots
- B. Visual and maximum likelihood methods for determining transformations
- C. Weighted least squares to adjust for nonconstant error variance
- D. Robust standard errors

*Readings:*

- \* Fox (1997), Chapter 12
- \* Fox (2002), Chapter 3 (pages 106-117), Chapter 6 (pages 201-216)
- \* Cook and Weisberg (1999), Chapters 13-14

## **10. Diagnostics for the Linear Model III: Collinear regressors**

- A. Variance inflation
- B. Principal components analysis
- C. Collinearity and model selection
- D. Ridge regression

*Readings:*

- \* Fox (1997), Chapter 13
- \* Fox (2002), Chapter 6 (pp 216-225)

## 11. Robust Regression

- A. M-Estimation and iteratively reweighted least squares
- B. Bounded influence regression

### *Readings:*

- \* Fox (1997), Chapter 14 (pages 405-435)
- \* Rousseeuw and Leroy (1987)

## 12. Resampling Techniques for Regression

- A. Bootstrapping and jackknifing
- B. Cross-validation

### *Readings:*

- \* Fox (1997), Chapter 16
- Davison and Hinkley (1997)

## 13. Nonlinear Regression

- A. Transformable nonlinearity
- B. Polynomial regression
- C. Orthogonal polynomials
- D. Nonlinear least squares

### *Readings:*

- \* Fox (1997), Chapter 14 (pages 388-404)
- \* Cook and Weisberg (1999), Chapters 7 (pages 147-153) and 16
- Bates and Watts (1988)

## 14. Nonparametric Regression I: Local Polynomial Regression

- A. Local regression (Loess)
- B. Loess fitting parameters
- C. Robust local regression
- D. Degrees of freedom

### *Readings:*

- \* Fox (1997), Chapter 14 (pages 417-433)
- \* Jacoby (2000)
- Fox (2000a)
- Loader (1999)
- Bowman and Azzalini (1997), Chapters 3 and 4

## 15. Nonparametric Regression II: Smoothing Splines

- A. Piecewise regression
- B. Cubic smoothing splines
- C. Thin plates smoothing splines
- D. Degrees of freedom

### *Readings:*

- \* Marsh and Cormier (2002)
- Schimek (2000), Chapters 1 (Eubank) and 2 (van der Linde)

## 16. Additive Regression Models

- A. Estimation and backfitting
- B. Degrees of freedom
- C. Cross-validation for smoothing parameters
- D. Diagnostics

### *Readings:*

- \* Fox (1997), Chapter 14 (pages 425-435)
- \* Fox (2000b)
- \* Hastie and Tibshirani (1990), Chapters 4 and 5

## 17. Generalized Additive Models

- A. Generalized additive models for binary dependent variables
- B. Vector generalized additive models for ordered dependent variables

### *Readings:*

- \* Hastie and Tibshirani (1990): Chapter 6
- \* Schimek (2000), Chapter 10 (Schimek and Turlach)
- Beck and Jackman (1998)

## 18. Graphical Regression

- A. Model checking plots
- B. Visualizing regression with more than two predictors
- C. Sequentially combining predictors

### *Readings:*

- \* Cook and Weisberg (1999), Chapters 17-20
- Cook and Weisberg (1994), Chapters 6-8, 11.

## References

- Beck, N. and S. Jackman. (1998) "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42: 596-627.
- Bowman, A.W. and A. Azzalini. (1997) *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford, UK: Oxford University Press.
- Bates, D. M. and D. G. Watts. (1988) *Nonlinear Regression Analysis and Its Applications*. New York: John Wiley and Sons.
- Cleveland, W.S. (1993) *Visualizing Data*. Summit, NJ: Hobart Press.
- Cook, R.D. and S. Weisberg (1994) *An Introduction to Regression Graphics*. New York: John Wiley and Sons.
- Cook, D.R. and S. Weisberg. (1999) *Applied Regression Including Computing and Graphics*. New York: John Wiley and Sons.
- Davison, A.C. and D.V. Hinkley. (1997) *Bootstrap Methods and Their Application*. Cambridge, UK: Cambridge University Press.
- Firth, D. (2003) "Overcoming the Reference Category Problem in the Presentation of Statistical Models." *Sociological Methods and Research* 33: 1-18.
- Firth, D. and R. X de. Menezes. (2004) "Quasi Variances." *Biometrika*, 91: 65-80.
- Fox, J. (1987). "Effect displays for generalized linear models." In C. C. Clogg (Editor), *Sociological Methodology 1987*. Washington DC: American Sociological Association.
- Fox, J. (1997) *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.
- Fox, J. (2000a) *Simple Nonparametric Regression*. Thousand Oaks, CA: Sage.
- Fox, J. (2000b) *Multiple and Generalized Nonparametric Regression*. Thousand Oaks, CA: Sage.
- Fox, J. (2002) *An R and S-Plus Companion to Applied Regression*. Thousand Oaks, CA: Sage.
- Hastie, T.J. and R. Tibshirani (1990) *Generalized Additive Models*. London: Chapman and Hall.
- Jacoby, W.G. (1997) *Statistical Graphics for Univariate and Bivariate Data*. Thousand Oaks, CA: Sage.
- Jacoby, W.G. (1998) *Statistical Graphics for Visualizing Multivariate Data*. Thousand Oaks, CA: Sage.

- Jacoby, W.G. (2000) "Loess: A Nonparametric, Graphical Tool for Depicting Relationships Between Variables." *Electoral Studies* 19: 577-613.
- Jasso, G. (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: 224-241.
- Jasso, G. (1986) "Is It Outlier Deletion or Is It Sample Truncation? Notes on Science and Sexuality." *American Sociological Review* 51: 738-742.
- Kahn, J.R. and J.R. Udry. (1986) "Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions." *American Sociological Review* 51: 734-737.
- Loader, C. (1999) *Local Regression and Likelihood*. New York: Springer.
- Long, J.S. (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.
- Marsh, L.C. and D.R. Cormier (2002) *Spline Regression Models*. Thousand Oaks, CA: Sage.
- McCullagh, P. and J.A. Nelder. (1989) *Generalized Linear Models (Second Edition)*. New York: Chapman and Hall.
- Monette, G. (1990) "Geometry of Multiple Regression and 3-D Graphics." In J. Fox and J.S. Long (Editors) *Modern Methods of Data Analysis*. Newbury Park, CA: Sage.
- Rousseeuw, P.J. and A.M. Leroy. (1987) *Robust Regression and Outlier Detection*. New York: John Wiley and Sons.
- Schimek, M.G. (Editor). (2000) *Smoothing and Regression. Approaches, Computation, and Application*. New York: John Wiley and Sons.
- Silber, J.H.; P.R. Rosenbaum; R.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
- Venables, W.N. and B. Ripley (2002) *Modern Applied Statistics with S (Fourth Edition)*. New York: Springer-Verlag.