Regression Analysis III: Advanced Methods

Dave Armstrong
University of Oxford
david.armstrong@politics.ox.ac.uk

Teaching Assistant:
Matthew Painter, Ohio State University
Painter.63@sociology.osu.edu

Course Website: http://users.ox.ac.uk/~polf0104/regressionIII.htm

1. Course Objectives

This course takes a modern, data-analytic approach to regression analysis, emphasizing graphical tools. The course will focus on the basic assumptions of linear regression, discussing various diagnostic tools for detecting violations of these assumptions and measures to adapt the linear model to accommodate data for which these assumptions are violated. The goal is to give a general overview of various modern extensions to the linear model. Effective ways of presenting the results of statistical models will also be discussed.

Topics to be covered are: assumptions of the linear model, graphical examination of data and transformations, regression diagnostics, linear model selection, robust and resistant regression, weighted least squares, generalized linear models, mixed-effects models, generalized least squares, bootstrapping and cross-validation, nonparametric regression (loess and smoothing splines), generalized additive models and vector generalized additive models.

2. Requirements

This course is part of a track of advanced courses that also includes *Maximum Likelihood* and *Bayesian Methods*. These courses are linked and integrated around the theme of providing advanced methodological training using the **R** statistical computing environment. Familiarity with **R** or **S**, or otherwise enrollment in the *Statistical Computing using R/S* course, is assumed. The course also assumes a good knowledge of regression analysis using matrix algebra and statistical inference for regression. If you do not meet these requirements, you should take one of the other regression courses instead (*i.e.,* Regression Analysis I or Regression Analysis II).

3. Course Texts

No one text tackles all of the subject matter of this course, but much of the material is covered in John Fox's applied regression texts:

- Fox, J. (forthcoming) *Applied Regression Analysis, Generalized Linear Models, and Related Methods, Second Edition.* Thousand Oaks: Sage.
- Fox, J. (2002) *An R and S-Plus Companion to Applied Regression*. Sage.

A draft of the first text is available for the cost of photocopying from the ICPSR summer program library in Helen Newbury. John Fox has allowed us to use the draft of this text under the condition that it is not distributed to students outside of this course. The latter book is excellent for those just beginning to use **R**. Other helpful books are listed in the reading list on the last page of this document. Readings specific to the daily lectures are given under the course schedule section of this document.

4. Computer Software

1. The R Base Package

All class demonstrations and assignments will be done using **R**, an implementation of the **S** language. Aside from its superb functionality and flexibility, **R** is also attractive because it is available free of charge. The base package and all of the add-on packages (called `libraries' in **S**) that we will use in the course will be installed in the computer lab. The ICPSR Summer Program will also provide a CD containing the latest version of **R**, a version of Latex, and a few other utilities that will be available for the cost of the CD. You can also download the base package and contributed packages from the **R** website <`http://www.r-project.org/`>. This site also includes links to several good manuals about **R** and information about add-on packages.

2. Add-on packages for R

Some of the packages that will be used in this course are:

- `boot`: Bootstrapping techniques associated with Davidson and Hinkley (1997)
- `car`: Regression diagnostics and other procedures associated with John Fox's (2002) *Companion to Applied Regression*
- `Design`: Many useful functions by Frank Harrell, including `robcov`for robust standard errors
- `effects`: A package by John Fox that creates effect displays for generalized linear models
- `foreign`: Imports SPSS, Stata and SAS data files
- `gam`: Functions for generalized additive models written by Trevor Hastie, following the philosophy in Hastie and Tibshirani (1990).
- `glmmML`: Generalized lineal models with random intercept
- `leaps`: For selecting model subsets
- `lqs`: Functions for robust regression
- `locfit`: Local regression techniques associated with Loader (1999)
- `MASS`: Modern statistical procedures from Venables and Ripley (2002)
- `mgcv`: For Generalized Additive Models
- `nlme` and `lme4`: Linear and nonlinear mixed-effects models
- `nls`: Nonlinear Least Squares
- `nnet`: Specifically for fitting neural networks, but also necessary for multinomial logit models
- `qvcalc`: A package by David Firth that calculates quasi-variances for dummy regressors in generalized linear models
- `relimp`: A package by David Firth that assesses the relative importance of the effects of

explanatory variables for generalized linear models

- `robustbase`: Various functions related to robust regression, including diagnostic plots using robust residuals, MM-estimation, and robust generalized linear models
- `sandwich`: Robust standard errors
- `scatterplot3d`: Creates 3-dimensional scatterplots
- `sm`: Smoothing procedures described in Bowman and Azzalini (1997)
- `splines` and `psplines`: Packages for smoothing splines
- `stepfun`: Returns and plots step functions
- `VGAM`: A package by Thomas Yee that fits Vector Generalized Additive Models

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

Packages → Install package(s) from CRAN...

We will also use two packages for **R** that are not yet available on the CRAN website, but are availaible freely on the authors' homepages:

1 `djmrgl`: A package by Duncan Murdoch that allows for dynamic 3D graphs <http://www.stats.uwo.ca/faculty/murdoch/software>

If you are using a Windows operating system you will need to download the `zip' versions of the files above. Once you have downloaded the files to a local directory, they can be easily installed using the menus within **R**:

Packages → Install package(s) from local zip files...

3. Other Related Software

You will find that a good text editor, such as WinEdt or Emacs (both of which can be adapted to **R**), is very useful for writing **R**-scripts. Also, equations and statistical output can be nicely presented using Latex, the code for which can be exported from **R** using the `xtable` package. My course notes will be created using PowerPoint and a free add-in for Latex called TeXPoint. For those who want to learn more about LaTeX, the Summer Program will be giving a set of lectures on the program that coincide with the third week of this course. Links to all the above programs, and some related manuals, can be found on my homepage: <http://socserv.mcmaster.ca/andersen/>.

5. Course Schedule

Each entry in the schedule below represents a single lecture. Suggested readings are starred (*); All other readings are supplemental. For some topics even the suggested readings are likely to be extensive, making it difficult for many of you to finish them before the lecture--I'll keep this in mind when preparing the lectures. Nonetheless, I suggest that you read them at some point, even if it is after the course has been completed.

Week 1

1. Tuesday, June 26: Preliminary Material
   (a) Goals of the course
   (b) Basics of least squares regression (properties of the estimator, assumptions, inference, regression in matrix form)

   Readings:
   - *Fox (forthcoming), Chapters 5, 6 and 9
   - *Fox (2002), Chapter 4

2. Wednesday, June 27: Examining and Transforming Data
   (a) Displays for univariate distributions (histograms, stem-and-leaf plots, density estimation, quantile comparison plots, boxplots)
   (b) Bivariate plots (parallel boxplots, scatterplots, bivariate density estimation, dynamic 3D histograms)
   (c) Multivariate plots (scatterplot matrices, dynamic 3D scatterplots, conditioning plots)
   (d) Power transformations for quantitative distributions
   (e) Logit and probit transformations for proportions

   Readings:
   - *Fox (forthcoming), Chapters 3 & 4
   - *Fox (2002), Chapter 3 (pp 85-106)
   - Jacoby (1997); Jacoby (1998); Cleveland (1993)

3. Thursday, June 28: Effective Presentation of Linear Models
   (a) Factors and contrasts; quasi-variances for dummy regressors
   (b) Fitted values, interactions, and effect displays
   (c) Standardization and relative importance

   Readings:
   - *Firth (2003)
   - *Fox (1987)
   - *Fox (forthcoming), Chapter 7
   - *Silber, Rosenbaum and Ross (1995)
   - Firth and Menezes (2004)

4. **Friday, June 29: Computer Lab (Perry Building)**
   Readings:
   - *Fox (forthcoming), Chapter 2
   - *Fox (2002), Chapters 1 & 2
   - Venables and Ripley (2002), Chapters 1-3

Week 2: Diagnostics

5. Monday, July 2: Diagnostics for the Linear Model I: Unusual Observations
   (a) Outliers, leverage and influential data
   (b) Hat values, studentized residuals, Cook's D
   Readings:
   - *Fox (forthcoming), Chapter 11
   - *Fox (2002), Chapter 6 (pp 191-201)
   - Cook and Weisberg (1999), Chapter 15
   - Jasso (1985); Jasso (1986); Kahn and Udry (1986)

6. Tuesday, July 3: Diagnostics for the Linear Model II: Nonlinearity, Nonnormality and Nonconstant Error Variance
   (a) Residual plots
   (b) Maximum likelihood methods for transformations
   (c) Box-Cox transformation of Y
   (d) Box-Tidwell Transformation of the Xs
   (e) Weighted least squares to adjust for nonconstant error variance
   (f) Robust standard errors
   (g) Polynomial regression
   Readings:
   - *Fox (forthcoming), Chapter 12 and 17
   - *Fox (2002), Chapter 3 (pp 106--117) & Chapter 6 (pp 201--216)
   - Cook and Weisberg (1999), Chapter 14

No class Wednesday, July 4th Holiday

7. Thursday, July 5: Diagnostics for the Linear Model III: Collinearity
   (a) Variance inflation factors
   (b) Principal components analysis
   (c) Collinearity and model selection
   (d) Ridge regression
   Readings:
   - *Fox (forthcoming), Chapter 13
   - *Fox (2002), Chapter 6 (pp 216--225)

8. Friday, July 6: Generalized Linear Models I
   (a) Limited dependent variables and problems with the OLS model
   (b) Binary logit and probit models

(c) Iteratively reweighted least squares
(d) Fitted probabilities and effect displays

Readings:

- *Fox (forthcoming), Chapter 15 (pp 341-362)
- *Fox (2002), Chapter 5 (pp 155-158)
- Long (1997), Chapters 3 & 4

Week 3

9. Monday, July 9: Generalized Linear Models II
   (a) Ordered probit and logit models
   (b) Multinomial logit models
   (c) Poisson models for count data
   (d) Diagnostics for GLMs

   Readings:
   - *Fox (forthcoming), Chapter 14 (pp 363-385) and Chapter 15
   - *Fox (2002), Chapter 5 (pp 167-188) & Chapter 6 (pp 225-233)
   - Long (1997), Chapters 5, 6 & 8
   - McCullagh and Nelder (1989)

10. Robust Regression (Tuesday, July 11)
   (a) Breakdown point, influence function, and various types of robust regression
   (b) M-estimation and extensions
   (c) Diagnostics for outliers revisited
   (d) Robust GLMs

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

11. Re-sampling Techniques for Regression (Wednesday, July 12)
   (a) Bootstrapping and Jackknifing
   (b) Cross-validation

   Readings:
   - *Fox (forthcoming), Chapter 21
   - Davison and Hinkley (1997)

12. Handling Dependent Data (Thursday, July 13)
   (a) Mixed-effects models for clustered and longitudinal data
   (b) Robust standard errors revisited

   Readings:
   - *Bryk and Raudenbush (1992), Chapters 1,2 & 4
   - *Pinheiro and Bates (2000), Chapters 1 & 2
   - Venables and Ripley (2002), Chapter 14

13. Friday, July 14: **Computer Lab (Perry Building)**

Week 4

14. Nonparametric Regression: I: Local Polynomial Regression (Monday, July 17)
    (a) Local regression
    (b) Span and bandwidth; polynomial degree; local weights
    (c) Robust local regression
    (d) Degrees of freedom

    Readings:
    - *Fox (forthcoming), Chapter 18
    - *Fox (2000a)
    - Loader (1999); Bowman and Azzalini (1997), Chapters 3 & 4

15. Nonparametric Regression II: Smoothing Splines (Tuesday, July 18)
    (a) Piecewise regression
    (b) Cubic smoothing splines
    (c) Thin plates smoothing splines
    (d) Degrees of freedom

    Readings:
    - *Marsh and Cormier (2002)
    - Schimek, ed. (2000), Chapters 1 (Eubank) & 2 (van der Linde)

17. Generalized Additive Models (Thursday, July 20)
    (a) Estimation and backfitting
    (b) Degrees of freedom
    (c) Cross-validation for smoothing parameters
    (d) Diagnostics
    (e) Generalized additive models for binary responses
    (f) Vector generalized additive models or ordered responses

    Readings:
    - *Hastie and Tibisharani (1990): Chapter 4-6
    - *Fox (2000b)
    - Schimek, ed. (2000), Chapter 10 (Schimek and Turlach)

18. Final Review

6. Reference List

Bowman, A.W. and A. Azzalini (1997) *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford, Oxford University Press.

Bates, D. M. and D. G. Watts (1988) *Nonlinear Regression Analysis and Its Applications*. Wiley.

Bryk, A.S. and S.W. Raudenbush (1992) *Hierarchical Linear Models*. Newbury Park: Sage.

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

Cleveland, W.S. (1993) *Visualizing Data*. Summit, NJ: Hobart Press.

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

Davison, A.C. and D.V. Hinkley (1997) *Bootstrap Methods and Their Application*. Cambridge University Press.

Firth, D. (2003) "Overcoming the reference category problem in the presentation of statistical models," *Sociological Methodology*, 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,' *Sociological Methodology*, 17:347-361.

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

Fox, J. (2000a) *Simple Nonparametric Regression*. (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-129). Thousand Oaks, CA: Sage.

Fox, J. (2000b) *Multiple and Generalized Nonparametric Regression*. (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-130). Thousand Oaks, CA: Sage.

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

Hastie, T.J. and R. Tibshirani (1990) *Generalized Additive Models*. London: Chapman Hall.

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.

Jacoby, W.G. (1997) *Statistical Graphics for Univariate and Bivariate Data*.(Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-117). Thousand Oaks, CA: Sage.

Jacoby, W.G. (1998) *Statistical Graphics for Visualizing Multivariate Data*. (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-120). Thousand Oaks, CA: Sage.

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.

Loader, C. (1999) *Local Regression and Likelihood*. New York: Springer.

Long, J.S. (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage.

Marsh, L.C. and D.R. Cormier (2002) *Spline Regression Models*. (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-137). Thousand Oaks, CA: Sage.

McCullagh, P. and J.A. Nelder (1989) Generalized Linear Models (2^{nd} Edition). New York: Chapman & Hall.

Pinheiro, J.C. and D.M. Bates (2000) *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.

Rousseeuw, P.J. and A.M. Leroy (1987)*Robust Regression and Outlier Detection*. New York: John Wiley & sons.

Schimek, M.G. (ed.) (2000) *Smoothing and Regression. Approaches, Computation, and Application* New York: John Wiley & Sons.

Silber, J.H., P.R. Rosenbaum and 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.

Yohai, V.J. (1987) 'High breakdown point and high efficiency robust estimates for regression,' *The Annals of Statistics*, 15:642-656.