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.

For the first time in the course’s history I will support both R and Stata. This departure comes as Stata’s capabilities have evolved to a point where most (and I stress, most) of what we want to do can be done easily in Stata. There will certainly be things that we want to do that are impossible (or very difficult), but enough of what we care about is available that it seems worth considering as an option. A couple of things to note:

1. 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. If you choose to use Stata, you should be comfortable with the idea of using the underlying programming language to write do files. While many of the things we want to do in Stata are already written, some things will still require a bit of extra work on our part and this work is often most efficiently conducted by writing little programs to accomplish the tasks. If you want to use this course as an opportunity to strengthen your R skills, but have little familiarity with that software, you should take the R workshop that I teach in the first two weeks of the first session.

If you’re one of those “glutton-for-punishment” types, you may also find it useful to learn L\TeX. L\TeX 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. L\TeX has its own bibliographic software (Bib\TeX) 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 and StatWeave) that integrate L\TeX, R and Stata 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 (and covered well) in:


The R 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:


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 3.0.1. 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.

For those using Stata, version 12 made significant improvements over version 11, most notably for our purposes is the debut of marginsplot, which we will use extensively.
4.1 Related Software

A good text editor is invaluable when using R and \LaTeX. \LaTeX{}Works is a good, free editor for \LaTeX{} that works in most environments, including Windows and Mac(\url{http://www.tug.org/texworks/}). RStudio is a free, recently-released IDE (Integrated Development Environment) for R that includes a nicely-featured text editor (\url{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 (\url{http://www.winedt.com/}), for Windows and TextMate (\url{http://macromates.com/}) are my favorites, but there are many other options as well.

5 Course Schedule

Each entry represents a single topic. 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, 25)

   (a) Goals for the course

   (b) Getting started with R.

Readings:

* Fox (2008), Chapters 1 & 2
* Fox and Weisberg (2011), Chapters 1 & 2
  – Venables and Ripley (2002), Chapters 1-3
2. OLS I: The Basics of Least Squares Regression (Wednesday, June 26)

(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 and Weisberg (2011), Chapter 4
* Gill (1999)
* Clarke (2005)
  – Gelman and Stern (2006)
  – Lewis-Beck and Skalaban (1990), Achen (1990), King (1990)

3. Graphics (Thursday, June 27)

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

Readings:

* Fox and Weisberg (2011) Chapter 7
* Kastellec and Leoni (2007)
  – Venables and Ripley (2002) Chapter 4
4. OLS II: Effective Presentation (Friday, June 28 and Monday July 1)
   (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)
* Berry, Golder and Milton (2012)
* Brambor, Clark and Golder (2006)
* Silber, Rosenbaum and Ross (1995)
  – Braumoeller (2004)
  – Firth and Menzes (2004)
  – Kam and Franzese (2007)

5. Diagnostics I: Linearity (Tuesday, July 2)
   (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 and Weisberg (2011) Chapter 3
* Jacoby (1999)
  – Cook and Weisberg (1999) Chapter 16
  – Box and Tidwell (1962)
  – Breiman and Friedman (1985a,b), Pregibon and Vardi (1985), Buja and Kass (1985),
    Fowlkes and Kettering (1985)

6. Non-Linearity, Smoothing and Splines (Wednesday, July 3)
   (a) Nonparametric Smoothing - Lowess
   (b) Inference for regression smoothers
   (c) Regression Splines
Readings:

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

7. Generalized Additive Models (Friday, July 5)

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

Readings:

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

8. Lab I: (Monday, July 8)

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

9. Re-sampling Techniques and Regression (Tuesday, July 9)

(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)

10. Diagnostics II: Outliers and Influential Data (Wednesday, July 10)

(a) Outliers, leverage and influential data
(b) Hat values, standardized residuals, Cook’s D
Readings:

* Fox (2008) Chapter 11
* Fox and Weisberg (2011) Chapter 6 (pp 101-201)
* Cook and Weisberg (1999) Chapter 15

11. Fixing Outliers and Influential Data: Robust Regression (Thursday, July 11)
   
   (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 (Friday, July 12)

   (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 and Weisberg (2011) Chapters 3 & 6
* Long and Ervin (2000)
* Harvey (1976)
13. Lab II (Monday, July 15)
   (a) Outliers and Robust Regression
   (b) Heteroskedasticity
   (c) Model Selection

14. Model Selection (Tuesday, July 16)
   (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)
   * Burnham and Anderson (2004)

15. Finite Mixture Models (Wednesday, July 17)

   Readings:
   * Imai and Tingley (Forthcoming)
   * Grün and Leisch (2008)
   * Grün and Leisch (2007)

16. Missing Data and Multiple Imputation (Thursday, July 18)

   (a) What's 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)
   * Resseguier, Giorgi and Paoletti (2011)
     – Allison (2001)
– Schafer (1997)
– Rubin (1987)

17. Lab III (Friday, July 19)

(a) Mixture Models
(b) Missing Data and Multiple Imputation
References


URL: [http://www.quantoid.net/factorplot_armstrong.pdf](http://www.quantoid.net/factorplot_armstrong.pdf)


URL: [https://files.nyu.edu/mrg217/public/jop2.pdf](https://files.nyu.edu/mrg217/public/jop2.pdf)


URL: [http://www.jstatsoft.org/v28/i04](http://www.jstatsoft.org/v28/i04)


