Regression III: Advanced Methods
Syllabus - Summer 2020

Instructor:
Dave Armstrong
Canada Research Chair in Political Methodology
Department of Political Science
Department of Statistics and Actuarial Sciences (by courtesy)
University of Western Ontario

e: dave.armstrong@uwo.ca
Office Hours: TBD (and by appointment)
Course website: http://quantoid.net/teachicpsr/regression3

Teaching Assistants

Kathryn Overton	Chris Schwarz
Political Science	Political Science
University of New Mexico	NYU
koverton@unm.edu	cschwarz@nyu.edu

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 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.

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, we will spend review these ideas briefly in class where necessary. 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.

I recently experimented with allowing participants to use either Stata or R and found, from student evaluations, and from both R and Stata users that the support for both pieces of software was both distracting and unnecessary. Stata users found that it was reasonably easy to pick up R for the purposes of the course. Further, Some of the specific software used in the course does not exist (or exist in the same useful way) in Stata. I found that the implementation in Stata often required some programming (loops, macros, etc...) and that was a threshold many participants did not want to cross. Thus, R will be the only officially supported software. That said, I have (now quite old) Stata code for some of the examples in class and would be happy to share with interested participants. 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 taught in the first few weeks of the first session.

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, common LaTeXtypesetting 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" with R integration through "knitr". There are some nice literate programming tools (Sweave, knitr and StatWeave) that integrate LaTeX, R and Stata as well. 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 try it; everyone else is doing it. There is a LaTeXworkshop on the first Monday night of the Summer Program.

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:

Fox, John. (2015) Applied Regression Analysis and Generalized Linear Models. *3rd ed.* Thousand Oaks, CA: Sage Publications, Inc.

Fox, John and Sanford Weisberg. (2011) An R Companion to Applied Regression. *2nd ed.* Thousand Oaks, CA: Sage Publications, Inc.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. (2013) An Introduction to Statistical Learning with Application in R. New York: Springer. Available for free here.

The R Companion is a great book for those currently learning R. I would highly recommend getting the recently updated and expanded second edition. This is widely recognized as one of the best ways for Social Scientists to get into R.

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 3.5.0. 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. TeXWorks is a good, free editor for LaTeXthat works in most environments, including Windows and Mac(`http://www.tug.org/texworks/`). RStudio is a free IDE (Integrated Development Environment) for R that includes a nicely-featured text editor (`http://www.rstudio.org/`). If you're looking for a general purpose text editor, I often use Microsoft's VS Code (`https://code.visualstudio.com`), though there are other great options, too like Atom or SublimeText.

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, June 23)

   Readings:
   ∗ James et al. (2013) Chapter 2
   ∗ Fox (2015), Chapters 1 & 2
   ∗ Fox and Weisberg (2011), Chapters 1 & 2

2. Effective Model Presentation I (Wednesday, June 24)

   (a) Factors and contrasts; quasi-variances and graphical displays
   (b) Standardization and relative importance

   Readings:
   ∗ Armstrong II (2013)
   ∗ Armstrong II (2013)
   ∗ Silber, Rosenbaum and Ross (1995)

3. Effective Model Presentation II (Thursday, June 25)

   (a) Factors and contrasts; quasi-variances and graphical displays
   (b) Interactions and effect displays
   (c) Standardization and relative importance

   Readings:
   ∗ Berry, Golder and Milton (2012)
   − Kam and Franzese (2007)

4. Linearity Diagnostics - Ordinal Variables (Monday, June 29)

   (a) Linearity and ordinal variables

   (b) Alternating Least Squares Optimal Scaling (ALSOS)

   Readings:
   * Fox (2015) Chapters 4 & 12 (Sections 12.3-12.5)
   * Jacoby (1999)

5. Linearity Diagnostics - Continuous Variables (Tuesday, June 30)

   (a) Diagnosing linearity through residual plots

   (b) Fixing non-linearity with data transformations and polynomials

   Readings:
   * Fox (2015) Chapters 4 & 12 (Sections 12.3-12.5)
   * Fox and Weisberg (2011) Chapter 3
   – Box and Tidwell (1962)

6. Bootstrapping (Wednesday, July 1)

   (a) Bootstrapping and Jackknifing

   Readings:
   * James et al. (2013) Chapter 5
   * Fox (2015) Chapter 21
   – Davison and Hinkley (1997)
   – Stone (1974)
   – Efron and Tibshirani (1993)

7. Model Selection (Thursday, July 2)

   (a) Theoretical issues in model searching and post-data model construction

   (b) Cross-validation

   (c) Model selection criteria and multi-model inference.

   (d) Subset selection models

   Readings:
   ∗ Fox (2015) Chapter 22
   ∗ James et al. (2013) Chapter 6
   ∗ Burnham and Anderson (2004)
   – Leamer and Leonard (1983)
   – Leamer (1983)
   – Box (1976), Box and Hunter (1962)

8. Generalized Additive Models for Location, Scale and Shape (Monday, July 6)

   (a) GAMLSS framework.

   (b) Modeling higher moments.

   (c) Regression diagnostics in GAMLSS

   ∗ Stasinopoulos and Rigby (2007)
   – Stasinopoulos et al. (2017)
   – Rigby et al. (2020)

9. Comparing GAMLSS Diagnostics/Fixes to Conventional Diagnostics/Fixes (Tuesday, July 7)

   (a) Heteroskedasticity

   (b) Outliers and Influence

   ∗ Fox (2015) Chapters 12 & 13
   ∗ Fox and Weisberg (2011) Chapters 3 & 6
   – King and Roberts (2015)
   – Long and Ervin (2000)
   – Harvey (1976)
   – Cribari-Neto (2004), Cribari-Neto, Souza and Vasconcellos (2007), Cribari-Neto and da Silva (2011)

10. Regression Splines (Wednesday, July 8)

    (a) Regression Splines

    (b) Inference for regression smoothers

    Readings:
    ∗ Fox (2015) Chapters 17 & 18
    ∗ James et al. (2013) Chapter 7
    – Keele (2008) Chapters 2-6

11. Smoothing Splines (Thursday, July 9)

    (a) Smoothing Splines

    (b) Smoothers in Generalized Additive Models

    Readings:
    ∗ Fox (2015) Chapters 18
    ∗ James et al. (2013) Chapter 7
    – Harezlak, Ruppert and Wand (2018) Chapters 2 and 3
    – Ruppert, Wand and Carroll (2003) Chapters 3, 5, 6 & 8

12. Flexible Models: Tree-based Regression, Multivariate Adaptive Regression Splines (Friday, July 10)

    - Fundamentals of flexible models
    - Automatic variable selection
    - Inference and effects in statistical learning models
    - When (and when not) to use these kinds of models

    ∗ Montgomery and Olivella (2018)
    ∗ James et al. (2013) Chapter 7
    ∗ Berk (2016) Section 3.14

13. Regression Discontinuity Designs (Monday, July 13)

    Readings:
    ∗ Cattaneo, Idrobo and Titiunik (2017)
    ∗ Calonico, Cattaneo and Titiunik (2015)
    ∗ Keele (2015)
    ∗ Sekhon and Titiunik (2016)

14. Finite Mixture Models (Tuesday, July 14)

Readings:
∗ Imai and Tingley (2012)
∗ Grün and Leisch (2008)
∗ Grün and Leisch (2007)

15. Missing Data and Multiple Imputation (Wednesday, July 15)

(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:
∗ Fox (2015) Chapter 20
∗ van Buuren and Groothuis-Oudshoorn (2011) ∗ Honaker and King (2010)
∗ Cranmer and Gill (2013)
∗ Akande, Li and Reiter (forthcoming)
∗ Xia and Yang (2016)
∗ Resseguier, Giorgi and Paoletti (2011)
– Schafer (1997)
– Rubin (1987)

16. Mixtures and Missing Data in GAMLSS (Thursday, July 16)

(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:
∗ Fox (2015) Chapter 20
∗ van Buuren and Groothuis-Oudshoorn (2011) ∗ Honaker and King (2010)
∗ Cranmer and Gill (2013)
∗ Akande, Li and Reiter (forthcoming)
∗ Xia and Yang (2016)
∗ Resseguier, Giorgi and Paoletti (2011)
– Schafer (1997)
– Rubin (1987)

17. Wrap-up (Friday, July 17)

References

Akande, Olanrewaju, Fan Li and Jerome Reiter. forthcoming. "An Empirical Comparison of Multiple Imputation Methods for Categorical Data." *The American Statistician* .
URL: *https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1277158*

Armstrong II, David A. 2013. "factorplot: Improving Presentation of Simple Contrasts in GLMs." *The R Journal* 5(2):4–15.
URL: *http://journal.r-project.org/archive/2013-2/armstrong.pdf*

Berk, Richard A. 2016. *Statistical Learning from a Regression Perspective, 2nd ed.* Switzerland: Springer.

Berry, William, Matt Golder and Daniel Milton. 2012. "Improving Tests of Theories Positing Interaction." *Journal of Politics* 74(3):653–671.
URL: *https://files.nyu.edu/mrg217/public/jop2.pdf*

Box, George E. P. 1976. "Science and Statistics." *Journal of the American Statistical Association* 71(356):791–799.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1976.10480949*

Box, George E. P. and William G. Hunter. 1962. "A Useful Method for Model-Building." *Technometrics* 4(3):301–318.
URL: *https://www.tandfonline.com/doi/abs/10.1080/00401706.1962.10490015*

Box, George and P.W. Tidwell. 1962. "Transformation of the Independent Variables." *Technometrics* 4:531–550.
URL: *https://www.tandfonline.com/doi/abs/10.1080/00401706.1962.10490038*

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.
URL: *https://journals.sagepub.com/doi/10.1177/0049124104268644*

Calonico, Sebastian, Matias D. Cattaneo and Rocio Titiunik. 2015. "rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs." *The R Journal* 7(1):38–51.
URL: *https://journal.r-project.org/archive/2015/RJ-2015-004/RJ-2015-004.pdf*

Cattaneo, Matias D., Nicolas Idrobo and Rocio Titiunik. 2017. "A Practical Introduction to Regression Discontinuity Designs.". Working Manuscript.

URL: *http://www-personal.umich.edu/~titiunik/books/CattaneoIdroboTitiunik2017-Cambridge.pdf*

Cranmer, Skyler J. and Jeff Gill. 2013. "We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data." *British Journal of Political Science* 43:425–449.
URL: *https://doi.org/10.1017/s0007123412000312*

Cribari-Neto, Francisco. 2004. "Asymptotic Inference Under Heteroskedasticity of Unknown Form." *Computational Statistics and Data Analysis* 45:215–233.
URL: *https://doi.org/10.1016/s0167-9473(02)00366-3*

Cribari-Neto, Francisco, Tatiene C. Souza and Klaus L.P. Vasconcellos. 2007. "Inference Under Heteroskedasticity and Leveraged Data." *Communications in Statistics - Theory and Methods* 36(10):1877–1888.
URL: *https://doi.org/10.1080/03610920601126589*

Cribari-Neto, Francisco and Wilton Bernardino da Silva. 2011. "A New Heteroskedasticity-consistent Covariance Matrix Estimator for the Linear Regression Model." *Advances in Statistical Analysis* 95(1):129–146.
URL: *https://doi.org/10.1007/s10182-010-0141-2*

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.

Fox, John. 2015. *Applied Regression Analysis and Generalized Linear Models,* 3^{rd} *edition*. Thousand Oaks, CA: Sage, Inc.

Fox, John and Sanford Weisberg. 2011. *An R Companion to Applied Regression,* 2^{nd} *ed.* Thousand Oaks, CA: Sage.

Grün, Bettina and Friedrich Leisch. 2007. "Fitting Finite Mixtures of Generalized Linear Regressions in R." *Computational Statistics & Data Analysis* 51(11):5247–5252.

Grün, Bettina and Friedrich Leisch. 2008. "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*

Harezlak, Jaroslaw, David Ruppert and Matt P. Wand. 2018. *Semiparametric Regression with R*. New York, NY: Springer.

Harvey, Andrew C. 1976. "Estimating Regression Models with Multiplicative Heteroskedasticity." *Econometrica* 44(3):461–465.
URL: *https://doi.org/10.2307/1913974*

Honaker, James and Gary King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data." *American Journal of Political Science* 54(2):561–581.
URL: *https://doi.org/10.1111/j.1540-5907.2010.00447.x*

Imai, Kosuke and Dusting Tingley. 2012. "A Statistical Method for Empirical Testing of Competing Theories." *American Journal of Political Science* 56(1):218–236.
URL: *https://doi.org/10.1111/j.1540-5907.2011.00555.x*

Jacoby, William G. 1999. "Levels of Measurement and Political Research: An Optimistic View." *American Journal of Political Science* 43(1):271–301.
URL: *https://doi.org/10.2307/2991794*

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. New York, NY: Springer.
URL: *https://bit.ly/2QLmMuD*

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

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

Keele, Luke J. 2015. "Geographic Boundaries as Regression Discontinuities." *Political Analysis* 23:127–155.
URL: *https://doi.org/10.1093/pan/mpu014*

King, Gary and Margaret E. Roberts. 2015. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It." *Political Analysis* 23(2):159–179.
URL: *https://doi.org/10.1093/pan/mpu015*

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *The American Economic Review* 73(1):31–43.
URL: *https://www.jstor.org/stable/1803924*

Leamer, Edward E. and Herman Leonard. 1983. "Reporting the Fragility of Regression Estimates." *The Review of Economics and Statistics* 65(2):306–317.
URL: *https://doi.org/10.2307/1924497*

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.
URL: *https://doi.org/10.1080/00031305.2000.10474549*

Montgomery, Jacob and Santiago Olivella. 2018. "Tree-based Models for Political Science Data." *American Journal of Political Science* 62:739–744.
URL: *https://doi.org/10.1111/ajps.12361*

Resseguier, Noémie, Roch Giorgi and Xavier Paoletti. 2011. "Sensitivity Analysis When Data Are Missing Not-at-random." *Epidemiology* 22(2):282.
URL: *https://doi.org/10.1097/ede.0b013e318209dec7*

Rigby, Robert A., Mikis D. Stasinopoulos, Gillian Z. Heller and Fernanda De Bastiani. 2020. *Distributions for Modeling Location, Scale and Shape Using GAMLSS in R.* Boca Raton, FL: CRC Press.

Rubin, Donald. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: J. Wiley and Sons.

Ruppert, David, Matt P. Wand and Raymond J. Carroll. 2003. *Semiparametric Regression.* New York, NY: Cambridge University Press.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall/CRC.

Sekhon, Jasjeet S. and Rocio Titiunik. 2016. "Understanding Regression Discontinuity Designs as Observational Studies." *Observational Studies* 2:174–182.
URL: *https://doi.org/10.1108/s0731-905320170000038001*

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.
URL: *https://doi.org/10.1080/01621459.1995.10476483*

Stasinopoulos, D. and Robert Rigby. 2007. "Generalized Additive Models for Location Scale and Shape (GAMLSS) in R." *Journal of Statistical Software, Articles* 23(7):1–46.
URL: *https://www.jstatsoft.org/v023/i07*

Stasinopoulos, Mikis D., Robert A. Rigby, Gillian Z. Heller, Vlasios Voudouris and Fernanda De Bastiani. 2017. *Flexible Regression and Smoothing Using GAMLSS in R.* Boca Raton, FL: CRC Press.

Stone, Mervyn. 1974. "Cross-validation and Assessment of Statistical Predictions." *Journal of the Royal Statistical Society, Series B* 36(2):111–147.
URL: *https://doi.org/10.1111/j.2517-6161.1974.tb00994.x*

van Buuren, Stef and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software, Articles* 45(3):1–67.
URL: *https://www.jstatsoft.org/v045/i03*

Xia, Yan and Yanyun Yang. 2016. "Bias Introduced by Rounding in Multiple Imputation for Ordered Categorical Variables." *The American Statistician* 70(4):358–364.
URL: *https://doi.org/10.1080/00031305.2016.1200486*