Regression III

Introduction

Dave Armstrong

Instructional Staff

Instructor: Dave Armstrong

E-mail: dave.armstrong@uwo.ca
Course Website: http://quantoid.net/teachicpsr/regression3/
Office Hours: 1:30-2:30 PM M-TH (or by Appointment)

Teaching Assistants:

Chris Schwarz (NYU, Political Science)
E-mail: cschwarz@nyu.edu
Office Hours: TBD

Kathryn Overton (U of New Mexico, Political Science)
E-mail: koverton@unm.edu
Office Hours: TBD

Course Materials

The course material will be posted in two places:

- My website will serve as the location of record for the course material and will stay active long after the course has ended.
- UMich Canvas
    - I will post links to my website for all of the material on `quantoid.net`
    - ICPSR is recording the course for later viewing and these recordings will only appear on Canvas.

What you need (R)

- R: I am using R v 4.1.0. If you're using an earlier version, please upgrade if you can.
- **Optional:** You should have some sort of IDE for R (RStudio, sublime, atom, vs code). I use Rstudio - it's not best on every dimension, but its combination of features make it a great tool for R and related technologies.
- **Optional:** If you are using a machine that prevents installing software, you could use RStudio Cloud which is a web-based RStudio distribution.
    - If this describes you, reach out to me and I can give you access to my RStudio.cloud instance.

Organization of Lectures

Each day, we will do the following (approximately):

- 40 minutes of lecture
- 5 minute break
- 40 minutes of lecture
- 5 minute break
- 25-30 minutes applied work

Homework

4-5 homework assignments

Classroom Management

Obviously, we are using Zoom as the platform for the course. Here are a few tips that will hopefully keep us all rowing in the same direction.

- The "reactions" button gives several options that allow you to respond to prompts non-verbally. Please use these to raise your hand, respond to yes-no questions or respond to questions about the pace of the course.

Getting Help

In Class

- You can use the slack group to ask questions that I can answer to the whole class.
- Or you can chat with the TAs directly by sending them a direct message in slack.
    - If you have a more complicated issue that requires a video chat in the moment, the TAs may have you join a different google meeting, so you should also be logged into a google account (either your UM account or a different one).

Outside of Class

- We will each have drop-in office hours M-F. We will try to cover a wide range of times.
- We will all also be available by appointment outside of class time.

Note Slides

Throughout the presentation there are slides (html) that have notes boxes in them.

- You can type in the text boxes to make some notes for yourself.
- If you click the "s" key, you will be allowed to draw on the slides with your mouse, trackpad or screen (if you have a touch device).
- You can then print the slides from the browser to PDF after your are done giving you pdf slides with your notes embedded.
    - This works best from Chrome.

I will put a notes slide after every slide from here on out.

What are we doing in the course?

- Broad view of regression (tracing the dependence of y on X).
  - Model Selection
  - Diagnostics
  - Testing
  - Presentation
- Think a lot about "Robustness" (again in broad terms)

Prerequisites:

- Regression (in matrix form),
- Understanding of Statistical Inference,
- MLE (would be nice, but not a pre-requisite *per se*)

Notes

Type notes here...

Course Books

Fox, John. (2016) ***Applied Regression Analysis and Generalized Linear Models***, 3^{rd} ed. Thousand Oaks, CA: Sage Publications, Inc.

Fox, John and Sanford Weisberg. (2018) ***An R Companion to Applied Regression***, 3^{rd} ed. Thousand Oaks, CA: Sage Publications, Inc.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. (2013) ***An Introduction to Statistical Learning with Applications in R***. New York: Springer pdf link

A more detailed list is at the back of the course syllabus

Notes

Type notes here...

The model

As a motivating example, let's say that we estimate:

$$y = b_0 + b_1 x_1 + b_2 x_2 + e$$

We identify $H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$.

- Presumably this means that we have a theory that suggests linearity (a particular functional form) of the relationship between x_1 and y.
- Normally, we would do a significance test on b_1 and that would tell us whether the estimated relationship is significantly different from zero.
- Assuming we reject H_0, do we interpret this as evidence that our theory is right?

Notes

Type notes here...

We might not be right...

There are a couple of potential impediments to rejecting H_0 meaning we're right.

- Functional form and the nature of models (Clarke and Primo, 2012)
  - Logical fallacy of affirming the consequent.
- Models involved:
  - Theory $\rightarrow$ Empirical Model
  - Concepts $\rightarrow$ Measures
  - Empirical Model $\rightarrow$ Measures.
- Better than nothing doesn't mean best.

Notes

Type notes here...

What does it mean to be right?

- If our hypotheses are a good description of the world, the functional form should be right.
  - Our original H_A becomes the new H_0 tested against H_{flex}, one where we remove functional form restrictions.
- If our hypothesis is about additivity, then there shouldn't be interesting interactions with other variables.
- If our hypothesis is right, then it should work for all data points.

Notes

Type notes here...

Understanding the Bias-Variance Tradeoff

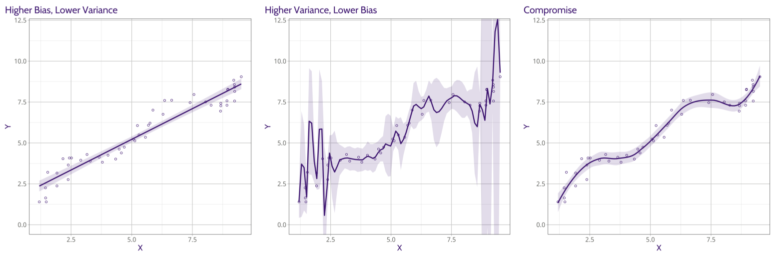
- Bias: difference between true dependence of y on x and the estimated dependence of y on x. Often we describe this as the difference between estimating a parametric model and interpolating the points, as closely as possible.

- Variance: the sampling variability of the regression line around the points.

Notes

Type notes here...

Understanding the Bias-Variance Tradeoff 2



Notes

Type notes here...

Bias-Variance Tradeoff

There is (nearly) always a bias-variance tradeoff to be made.

- Can characterize the bias-variance tradeoff with the Mean Squared Error (MSE).
  - $MSE = Bias^2 + Variance$
  - Lower MSE models have a better bias-variance tradeoff.

Notes

Type notes here...

Evaluating "rightness" of B-V Tradeoff

Method	Linearity	Simple Interactions	Complex Interactions
Splines	✓		
Penalized Splines	✓		
MARS	✓	✓	
Polywog	✓	✓	
CART	✓		✓
Random Forest	✓		✓

Notes

Type notes here...

Model Testing and Selection

- Theory testing - selecting between two known models (generally operationalizing H_0 and H_A.
  - Evaluating strength of evidence for a set of known models.
- Feature selection - finding the most important variables.
  - All subsets regression.
  - Ridge Regression/LASSO/Elastic-Net
  - MARS
  - Decoupling Shrinkage and Selection (DSS).

Notes

Type notes here...

Other neat applications of regression

- Regression Discontinuity Designs
- Finite Mixtures
- Missing data/Multiple imputation

Notes

Type notes here...

More conventional diagnostics (with a couple of tweaks)

- Outliers
  - Robust Regresion as diagnostic
- Heteroskedasticity
  - Robust standard errors (there are lots of them)
  - Trouble with robust standard errors
  - Bootstrapping for appropriate inference.

Notes

Type notes here...

Importance of Gauss-Markov Assumptions

Now we know that the OLS estimator $\mathbf{b}$ is linear, unbiased, and efficient. What assumptions did we have to make to get there?

- Linearity
  - $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, or equivalently $E(\varepsilon) = 0$
  - No perfect collinearity (or $\mathbf{X}$ of full-rank).
- Unbiasedness
  - ε independent from $\mathbf{X}$

Notes

Type notes here...

Importance of Gauss-Markov Assumptions II

- Efficiency
  - Homoskedasticity: $V(\varepsilon|\mathbf{X}) = \sigma^2$, or equivalently $V(\varepsilon|\mathbf{X}) = \sigma^2\mathbf{I}_n$
- Approximately correct type I error rate:
  - Assume a functional form of the error distribution: $\varepsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$

Notes

Type notes here...

F-test (just a reminder)

- Assume we have an OLS model with k explanatory variables that produces residual sum of squares RSS for the *full* model.
- Now, place q linear restrictions on the model coefficients (e.g., set some of them to zero) and generate a new residual sum of squares RSS_0 for the *restricted* model.

$$F_0 = \frac{\frac{RSS_0 - RSS}{q}}{\frac{RSS}{n-k-1}}$$

The statistic F_0 is distributed F with q and $n - k - 1$ degrees of freedom.

Notes

Type notes here...

Tomorrow

- Effective Presentation of Linear Model Results.

Notes

Type notes here...