PLSC 503: "Multivariate Analysis for Political Research"

Introduction
January 12, 2016

- Welcome!
- Course logistics.
  - Roughly 15 weeks of regression models, minus a few cancelled classes.
  - Course meets every Tuesday & Thursday, from 11:15-12:30, in Mateer 110.
  - Syllabus, notes, readings, homeworks, etc. on the github repo.
  - Books: Weisberg and Kennedy (buy them), plus other materials.
  - Mostly articles from political science – theory + recommended applications.
  - Grading: Ten homework assignments (@ 50 points), plus a final project (500 points); more on the latter, later.
  - Nick Dietrich is the teaching preceptor.
  - No office hours per se – e-mail me.
  - I don't care if you miss class.
  - I don't care what software you use (R, Stata are supported; R is very strongly preferred).
  - You better know some math, or at least not be afraid of it.
- About me.
  - What to call me.
  - Where I come from / background / history.
  - Personal stuff...
- About you all.
  - Field(s)?
  - Year / cohort?
  - Interests?
  - Previous training?
  - Personal stuff?

Notation

Arabic letters:

- Letters $a,\ b,\ c,\ d$ are usually constants
- $f,\ g,\ h$ are usually functions
- $i,\ j,\ k,\ t$ are usually indices/term identifiers
- $p,\ q$ are usually probabilities
- e and u are often used for error terms (e is also a specific value, $\approx 2.71828...$)
- $x,\ y,\ z$, are usually variables or scalars

Greek letters:

- $\alpha, \beta, \gamma, \delta, \zeta, \eta, \theta, \kappa, \lambda, \rho, \tau, \phi, \chi, \psi$, and ω are usually parameters/coefficients
- ϵ is often (usually) an error term
- μ almost always represents a mean
- σ almost always represents variation in some form
- π is a specific value ($\pi = 3.14159265...$), but can also be a probability

Typically we will use lower-case for scalars or vectors and upper-case for sets and matrices.

Logical and Relational Operators and Set Notation

Ways of writing simple and not-so-simple logical statements regarding groups of items or terms.

- {...} list the contents (elements) of a set
- (...) indicate open intervals; define a set which does not include the endpoints – e.g. if $U = (1, 3)$, then 3 is not an element of U
- [...] indicate closed intervals, defining a set which does contain the endpoints
- $\in$: "is an element of"; $\notin$: "is not an element of"
- $\ni$: "such that"
- $\exists$: "there exists" or "there is"; $\nexists$: "there does not exist"
- $\forall$: "for all"

- $\therefore$: "therefore"
- $\because$: "because"
- $\between$: "between"
- $\rightarrow$ or $\Rightarrow$: "implies" or "implies that" (also, "to" or "then")
- $\iff$: "if and only iff" (also, "iff")
- $\subseteq$: "is a subset of"; $\subset$: "is a proper subset of"
- $\cup$: "the union of"; $\cap$: "the intersection of"
- $\emptyset$: "the null/empty set" (a set with no elements)

Operations

Symbols indicating transformations and relationships between terms...

- $=$ (everybody knows this one): equals, or "consists of"
- $\neq$ (not equal to)
- $\approx$ (approximately equal to)
- $\equiv$ (equivalent to, same/defined as)
- $\sim$ (is distributed as)
- $\propto$ (is proportional to)
- $\doteq$ (approaches; is equal to in the limit)

Likewise, everybody is familiar with $+$, $-$, $\times$ (also * and $\cdot$), and $\div$.

You also need to know the *order of operations* (i.e., the order you do these calculations in...)

1. Parentheses
2. Exponents
3. Multiplication/division
4. Addition/subtraction

Please get the order of operations right!!! (E.g., $3X^2 \neq (3X)^2$, generally; $X = 0$ being the obvious exception).

Summation Notation

- $\sum(...)$ generally indicates summing inside the parentheses,
- $\sum_{i=1}^{N} X_i$ indicates summing the N values of X ...
- ...this can also be written $\sum_N X_i$.

Likewise, product notation indicates multiplication...

- $\prod_{i=1}^{T} X_i^2$ indicates successively multiplying the T values of X^2

Factorials

- $x! = x \times (x-1) \times (x-2) \times ... \times (2) \times (1)$
- So: $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

Exponents

- Indicate repeated self-multiplication (i.e., $X^2 = X \cdot X$).
- Negative exponents: $X^{-1} = \frac{1}{X}$, $X^{-2} = \frac{1}{X^2}$, etc.
- Fractional exponents: $X^{\frac{1}{2}} = \sqrt{X}$, $X^{\frac{1}{3}} = \sqrt[3]{X}$, etc.
- So: $X^{3/4} \equiv X^3 \times \sqrt[4]{X} = \sqrt[4]{X^3}$; $X^{-7/3} = \frac{1}{\sqrt[3]{X^7}}$, etc.

Linear Algebra

- Per normal conventions, I will generally denote matrices in **boldface**. So, $\mathbf{X}$ generally denotes something different from X.
- This will be clear in the slides. Since there's no good way of making such distinctions on a chalkboard, you'll just have to pay attention.

Regression: A Conceptual Overview

Since this is a course about regression, it behooves us to think a bit about what regression is, at a general level. To do that, today we'll spend a bit of time talking about what regression is *not* – that is, other things one might do with (multivariate) data, and the statistical / quantitative methods and approaches one uses to do them.

"Multivariate" Regression

It's easiest to think of *regression* as mapping a vector of responses (Y) to a matrix of predictors / covariates ($\mathbf{X}$). We can extend this to the case where Y is matrix-valued ($\mathbf{Y}$), where we get "true" multivariate regression. Multivariate linear regression might look like:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$$

An example is *vector autoregression* (VARs), a time-series approach where all elements of $\mathbf{Y}$ are also (generally) elements of $\mathbf{X}$.

Measurement

Exercises in data / dimensional reduction (going from multivariate to less-multivariate or even univariate).

An example are simple (additive, etc.) indices. For example, one could:

- Take each of five measures of "health" (IM, Fertility, LE, measles, and DPT percentages),
- "standardize" each (to put them all on the same "scale"), so that each is $\mu = 0$ and $\sigma = 1$, and then
- add (or subtract, as the case may be) them together.

Other, more complex methods include:

- Principal Components Analysis ($\mathbf{Y} = \mathbf{W}^{\mathrm{T}}\mathbf{X}$...)
- Factor Analysis (like PCA, but somewhat different...)
- Uni- and Multidimensional Scaling (e.g., Guttman & Mokken scaling, etc.).
- Structural Equation Modeling [used with continuous variables, where there is a strong *a priori* understanding that the variables measure the same underlying factor(s)]
- Item-Response (IRT) Models (a la the SATs... usually used with binary or ordinal-response data, rather than continuous indicators)

Classification

- Cluster Analysis (hierarchical or not; agglomerative or divisive, etc.).
- Classification and Regression “Trees” (akin to cluster analysis...) $\rightarrow$ random forests.
- Pattern Recognition (gene sequencing, etc.)
- Machine learning, support vector machines, etc.