PLSC 503: "Multivariate Analysis for Political Research"

Regression: A Conceptual Overview
January 17, 2017

Some Housekeeping...

- The github repo is now up and running; go there for all your PLSC 503 needs.
- The books are not in the bookstore; you can do better (price-wise) buying them on-line anyway.
- There is a set of "Introduction to R" slides on the github repo; take a look at them if you want, as they might teach you something (or maybe not).

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, we'll also 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.

Regression Revealed

The basic idea of regression is one of *conditional distributions*. More specifically, regression usually (but not always) involves a *single* response / dependent variable Y, and a set of *multiple* independent variables / covariates $\mathbf{X}$. What we are usually interested in is the distribution of Y given $\mathbf{X}$:

$$\Pr(Y|\mathbf{X}) = f(\mathbf{X}) \tag{1}$$

In other words, we want to know something about the conditional density / "shape" of Y over a range of values of our covariates $\mathbf{X}$, where we'll typically denote $\mathbf{X}$ as a $N \times k$ matrix (and Y as a $N \times 1$ vector). The reasons we might want to know this are several:

- We might be interested in knowing / learning about *associations* between Y and the elements of $\mathbf{X}$,
- we might want to know the *causal effect* of (one or more elements of) $\mathbf{X}$ on Y, and/or
- we might want to *predict* what Y is likely to be at some particular configuration of values for $\mathbf{X}$.

These three things – association, causation, and prediction – are what we do 90 percent or so of the time we undertake quantitative analyses.

All that seems easy enough, in principle. In practice, things get a bit more complicated. In particular, the term in equation (1) implies two things:

1. The distribution of Y is *conditional on all variables in* $\mathbf{X}$, and
2. The conditional distribution is, in fact, conditional on the *joint distribution* of the elements of $\mathbf{X}$. That complicates matters, because the variables in $\mathbf{X}$ are very likely to be conditionally non-independent of each other. So, in conditioning on $\mathbf{X}$, we cannot simply do so "one at a time;" rather, we must consider the full *joint* distribution of $\mathbf{X}$ to get at (1).

As an intuitive matter, consider the example below...

An Example: Infant Mortality in 2000

The slides have a series of figures that plot infant mortality rates (that is, infant deaths per 1000 live births, denoted *IM*) against some factors that might be thought to be correlated with them. Consider each of the following variables:

1. *Life Expectancy (*LE*)* (in years):
   - What is the distribution of infant mortality at $LE = 45$? At $LE = 55$? At $LE = 75$?
   - What (in general) can one say about how $\mathrm{E}(IM)$ and $\mathrm{Var}(IM)$ change over different values of LE?
2. Figure 2 shows the "residuals" – the difference between the infant mortality we actually observe for that country and what we would expect if the country were "on the line."
3. Figure 3 shows infant mortality plotted against *Fertility* (measured as births per woman):
   - One might ask the same questions as above...
4. Figure 4 shows infant mortality plotted against *Wealth* (measured as GDP per capita):
   - Ditto:
     - Monotonic
     - Curvilinear / diminishing returns in IM to wealth
     - High variance at low levels of wealth, decreasing as wealth increases
   - The second wealth figure (Figure 5) – with axes plotted on the log scale – shows how the log-log relationship between the two is linear...

5. Figure 6 shows infant mortality as a function of *Democracy* (measured as POLITY IV -10 to 10 score):
   - Non-monotonic, curvilinear (inverted-U-shaped)
   - Relatively constant variance
6. The second *Democracy* plot (Figure 7) conditions on both *Democracy* and (in a simple dichotomous way) *wealth*...
   - Rich countries have low infant mortality, low *variance* in infant mortality, and *IM* is unresponsive to *democracy*.
   - Poor countries have (on average) higher infant mortality, more highly variable infant mortality, and IM is responsive (in that curvilinear way discussed above) to *democracy*.

All of this underscores exactly how complex regression modeling is; in the last figure above, for example, we are conditioning simultaneously (and in a very simple way) on only two variables, and yet there's a lot of nuance presented. Trying to do the same for three, four, five, etc. would be exceedingly difficult (though not impossible...).

What Else Is There?

Measurement

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

An example are simple (additive, etc.) indices, of the sort described in the slides. Here' I've just:

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

The result is the simple additive index in the last row/column of Figure 9 in the slides. Of course, there are a host of other, more complex approaches to statistical measurement as well. These include:

- Principal Components Analysis ($\mathbf{Y}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{W}$...)
- 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...) → random forests.
- Pattern Recognition (gene sequencing, etc.)
- Machine learning, support vector machines, etc.

So What Good Is Regression?

If we consider the three (general) things that one wants to do with data – describe, explain, and predict – then regression-like models tend to fall most clearly in the "explanation" category. They tend to be:

- multivariate, but not *too* multivariate
- theory-driven, not atheoretical
- of greatest interest when *marginal* quantities / associations are of interest.

This little table outlines the three "things" we do with data, and suggests why regression tends to fall in the middle column.

	Description	**Explanation**	**Prediction**
Task	Summarize data	Correlation/causation	Forecast OOS / future data
Emphasis	Data	Theory / Hypotheses	Outcomes
Focus	Univariate	Multivariate	Multivariate
Typical Application	Summarize / "reduce" data	Discuss marginal associations between predictors and an outcome of interest	Optimize out-of-sample predictive power / minimize prediction error