University of Illinois
Fall 2012

Department of Economics
Roger Koenker

Economics 508

Lecture 24
Ecological Regression

The phrase "ecological regression" refers to the common desire to use standard regression methods to infer individual behavior from spatial aggregated data. The most typical situation would seem to be the attempt to infer ethnic voting behavior from precinct level data in political science. Thus, for example we might like to know what proportion of Hispanics voted for Obama in the 2012 election.

1. Goodman's Regression

Given data on the proportion of Hispanics, x_i and the proportion of the vote going to Obama, y_i, in a large number of precincts, we are tempted to estimate the ecological regression,

$$y_i = a + bx_i + u_i.$$

If it were reasonable to assume that the coefficients a and b were constant over precincts and u_i was well-behaved, then we could interpret $\hat{a}$ as an estimate of the proportion voting for Obama among non-Hispanics, and $a + b$ as the proportion of Hispanics voting for Obama, corresponding to the extreme cases of $x_i = 0$ and $x_i = 1$, respectively. This is sometimes described as Goodman's (1953) regression. Note that a potentially embarrassing drawback of this approach is the possibility that we could end up with estimates of the two parameters a and b that fall outside the interval $[0, 1]$. Various refinements are possible, most obviously a weighting by the size of the regions, but there is nothing to ensure that such refinements are going to help make the estimates more accurate.

2. Method of Bounds

An alternative is the "method of bounds" introduced by Duncan and Davis (1953). Freedman (1999) illustrates this approach with an example using CPI data from Washington state. Suppose we know that 0.079 of the population is foriegn born and 0.344 of the population have "high income." We are interested in the proportion, p of the foreign born who have "high income." We know that

$$0.344 = 0.079p + (1 - 0.079)q$$

where q denotes the proportion of the native born with high income. This reveals the essential problem: we have only one equation to determine two unknowns. But all is not lost, suppose we solve for q in terms of p and

Method	p	q
Truth	0.35	0.28
Nbd	0.34	0.36
Goodman	0.29	0.85
King	0.30	0.72

then observe that p must be between zero and one. This implies that $q \in [0.288, 0.374]$. Try it! Manski (2007) elaborates this idea in many other contexts. Manski argues that many problems in econometrics have this form, that models are inherently underidentified, but some bounds can be placed on parameter estimates based on careful analysis of the probability structure of the problem. Another example of this sort of analysis is the case of regression data in which we observe intervals $y_i \in [\underline{y_i}, \overline{y_i}]$ and we would like to make inferences about the standard regression model. The challenge in all such models is to carefully specify the probability structure of the model, and when the identified set in non-unique to find a practical way to make inferences about these sets.

3. Random Coefficients

Now, suppose that we have many observations on (x_i, y_i) as above and we would like to consider the random coefficient model,

$$y_i = p_i x_i + q_i(1 - x_i).$$

This is obviously a generalization of the Goodman model. King (1997) in an influential (and controversial) book on the subject assumes that (p_i, q_i) are drawn iid-ly from a bivariate normal distribution truncated to respect the requirement that they should lie in $[0, 1]^2$. This model has a reasonably tractable likelihood and can be therefore estimated by maximum likelihood. Freedman compares three methods of estimating (p_i, q_i): the Goodman regression, the King regression, and a simple model that he calls the neighborhood model that assumes that outcomes are determined by geography not demography. In his formulation, the neighborhood model assumes that $p_i = q_i = y_i$ in each region. This is obviously quite extreme, but in Freedman's example, where we know the correct answer thanks to the crosstabs provided by the CPI, the neighborhood approach is better than the others in estimating the mean of the p's and q's. I reproduce his table here.

There is an interesting connection of the King method to medical imaging called tomography. In tomography a 3d image is reconstructed from many 2d slices. In King's tomography plot, each point (x_i, y_i) appears as a line in the parameter space of (p, q)'s. Pairs of these lines have intersections that can be taken as meta-observations to which we try to fit the normal model. Of course, there doesn't seem to be any compelling reason to think that the normal model is appropriate. So it would seem prudent to explore other less parametric approaches. One such approach is the Kiefer-Wolfowitz (1956)

nonparametric MLE, which would replace the normal model with a discrete mixing distribution with a relatively small number of mass points. If you were very lucky these mass points might yield an interpretable clustering of the regions.

References

Freedman, D.A. (1999) Ecological Inference and the Ecological Fallacy, in *International Encyclopedia of the Social & Behavioral Sciences*, available as `http://statistics.berkeley.edu/tech-reports/549.pdf`

Goodman, L. 1953 Ecological regression and the behavior of individuals. *American Sociological Review* 18: 66364

Duncan, O. D, Davis B 1953 An alternative to ecological correlation. *American Sociological Review* 18: 66566

King, G. 1997 *A Solution to the Ecological Inference Problem.* Princeton University Press

Manski, C. (2007) *Identification for Prediction and Decision*, Harvard U. Press