Data Recipe #34:
Nonparametric Plot with Two Levels of Aggregation

Jesse M. Shapiro

The Problem. A federal job training program grants additional funds to counties whose poverty rate exceeds a cutoff. You want to study the program's effect on unemployment. You have data on poverty rate at the county level, but you can only measure unemployment at the state level. Is there any way to make the standard "RD" plot to visualize the relationship between unemployment and poverty rate to see if there is a discontinuity where the program kicks in?

The Solution. Under the right assumptions, you can make this plot by exploiting the fact that averages of linear functions are still linear.

The Math. Let x_{cs} denote the poverty rate in county c in state s and let y_{cs} denote the county's unemployment rate. If we measured y_{cs} we could plot its mean as a function of x_{cs}, say by constructing groups of counties with similar values of x_{cs}. Let $z(x_{cs})$ denote a vector of indicators for whether values of x_{cs} fall in pre-specified bins, defined, say, by percentiles or integer values of x_{cs}. Computing the value of y_{cs} for each bin is equivalent to estimating the parameters β in the regression

$$\mathrm{E}(y_{cs}|x_{cs}) = z(x_{cs})'\beta.$$

Unfortunately we don't measure y_{cs}, but we do measure $y_s = \sum w_{cs} y_{cs}$ where w_{cs} are known weights (say, the size of the labor force in each county). Let x_s and w_s be vectors of poverty rates and weights for the counties in state s. Notice that if

$$\mathrm{E}(y_{cs}|x_s, w_s) = z(x_{cs})'\beta \tag{1}$$

then it follows also that

$$\mathrm{E}(y_s|x_s, w_s) = \mathrm{E}\left(\sum w_{cs} y_{cs}|x_s, w_s\right) = \left(\sum w_{cs} z(x_{cs})\right)'\beta.$$

We can therefore make the desired plot by regressing state-level unemployment y_s on a vector $(\sum w_{cs} z(x_{cs}))$ which records the share of each state that falls into each poverty-rate bin.

The Warning. The assumption in (1) is not free. It can be wrong for reasons that are well-known in the literature on ecological regression and the ecological fallacy. In words, it says that a county's unemployment rate is not related to the poverty rate of other counties in the state, once we know the county's own poverty rate. This is almost certainly not literally true, and whether it provides a good enough approximation to yield a useful plot will depend on your setting.

The Recipe. This is easy to do:

```
use county_poverty.dta, clear
egen bin = cut(poverty), group(100)
tab bin, gen(zbin)
collapse (mean) zbin* [w=laborforce], by(state)
merge state 1:1 using unemployment.dta, keepusing(unemployment)
regress unemployment zbin*, noconstant
[plot coefficients on zbin*]
```

Hope it comes out right!