Beliefs about Data Quality and Implications for Inference

Max Gallop
Duke University

Simon Weschle
Duke University

May 19, 2012

1 Setup

Consider an observation y_i, which we model as $f(y_i) = \beta_1 x_i + \gamma^T Z_i + \varepsilon_i$ where $f(\cdot)$ is some link function, β_1 the coefficient associated with variable x_i that we care about, Z_i a vector of covariates we don't care about and γ the associated vector of coefficients and ε_i is a draw from some distribution with mean 0. We suspect that x_i does not correctly capture the variable it purports to measure and that there are other variables that systematically affect how well x_i measures the true data. We could also suspect that y_i is not measured correctly, but this is not discussed here for now. Possible examples for such data are:

- Historical records for the number of fatalities of wars. This series becomes potentially more accurate over time as better logistics and data processing capability makes it easier to keep track of the number of fatalities.
- A cross-national survey asking respondents e.g. whether they have bribed officials, have been the victim of police brutality, or whether they have received clientelistic benefits in exchange for vote support. One suspicion is that respondents in countries with a more repressive regime are less likely to answer "yes" to each of those questions.
- The W-ICEWS data reports the monthly number of events for certain events of interest for almost all countries in the world. These events are gleaned from news stories. Since correspondents that are based in some country often also cover a number of other countries in the region, one can expect that the further countries are away from the nearest correspondent the more events will go unreported.

The goal of the approach is to quantify *beliefs* about how a variable systematically influences the data quality of another variable and see how this affects inference, e.g. whether results change compared to using the assumption that all variables are measured without systematic error.

2 Math

2.1 Bernoulli

Suppose we think the x_i are independent Bernoulli trials. But we think the variable is not perfectly measured. Denote the true data by $t_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. We believe that whenever $t_i = 0$ then $x_i = 0$, but when $t_i = 1$, $x_i = 1$ with probability π_i and $x_i = 0$ with probability $1 - \pi_i$. So whenever the true outcome is no success the data will record this, but when the true outcome is a success it is only reported with some case-specific probability. We cannot observe the detection probability π_i or the true data t_i, only x_i is known. But we suspect that the detection probability is systematically influenced by some other variable w_i, here modeled through a logit link

$$log\left(\frac{\pi_i}{1 - \pi_i}\right) = \alpha_0 + \alpha_1 w_i \tag{1}$$

where w_i is an observable variable, e.g. in the examples above time, regime repressiveness and distance to the nearest news correspondent. α_0 and α_1 are the two quantities which we have beliefs about and that need to be quantified for the analysis. This is done by specifying two prior distributions (the one on α_0 is not strictly necessary) that represent our beliefs. Given draws from this prior distribution, we can create a distribution of t_i, which we can feed into the main regression and record the resulting changes in β_1 and γ.