Beliefs about Data Quality and Implications for Inference

Max Gallop
Duke University

Simon Weschle
Duke University

May 24, 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 Normal

Suppose we think that the x_i are independent draws from a normal distribution. But we think the variable is not perfectly measured. Denote the true data by $t_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. We believe that there is some variable m_i that quantifies how much x_i differs from t_i. In particular, we think that

$$m_i = \alpha_0 + \alpha_1 w_i \tag{1}$$

where w_i is an observable variable, e.g. in the examples above time, regime repressiveness and distance to the nearest news correspondent. This leads to a model of the observed data:

$$x_i = t_i + m_i \tag{2}$$

Our approach is as follows: α_0 and α_1 are two unknown quantities, but we have beliefs about them. They are quantified by specifying two prior distributions denoted by $p(\tilde{\alpha}_0)$ and $p(\tilde{\alpha}_1)$. Denote the draws from this distribution by $\tilde{\alpha}_0^{(j)}$ and $\tilde{\alpha}_1^{(j)}$. This gives us a distribution of our beliefs on m_i:

$$\tilde{m}_i^{(j)} = \tilde{\alpha}_0^{(j)} + \tilde{\alpha}_1^{(j)} \tag{3}$$

We can therefore construct estimates of the true data from the observed data and our beliefs about m_i:

$$\hat{t}_i^{(j)} = x_i - \tilde{m}_i^{(j)} \tag{4}$$

So instead of estimating $f(y_i) = \beta_1 x_i + \gamma^T Z_i + \varepsilon_i$, we estimate $f(y_i) = \beta_1 \hat{t}_i + \gamma^T Z_i + \varepsilon_i$ instead. If our beliefs about α_0 and α_1 are correct, the β_1 resulting from the "belief model" will approach the true β_1 from estimating $f(y_i) = \beta_1 t_i + \gamma^T Z_i + \varepsilon_i$.