Beliefs about Data Quality and Implications for Inference

Max Gallop
Duke University

Simon Weschle
Duke University

May 31, 2012

1 Idea

Suppose we are interested in a variable $\mathbf{t}$ with elements t_i, where $i = 1, \ldots, n$. The variable is not perfectly measured and we observe $\mathbf{x}$ instead. We suspect that there is a variable $\mathbf{w}$ that systematically influences how close the elements x_i are to t_i. Possible examples 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. So here $\mathbf{t}$ is the true number of fatalities, $\mathbf{x}$ the number of observed fatalities and $\mathbf{w}$ would be time.
- 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 variable $\mathbf{w}$ systematically influences how the true values $\mathbf{t}$ are reported in $\mathbf{x}$. Given these subjective beliefs, the approach can then perform a sensitivity analysis and give information on how robust the results are to suspected systematic measurement error.

2 Math

2.1 Continuous Variable

Suppose $\mathbf{t}$ (and its realization $\mathbf{x}$) is a continuous variable with support on the real line. We model the relationship between true and observed values as follows:

$$t_i = x_i \cdot m_i \tag{1}$$

where m_i is our *belief* about how well x_i measures t_i. If $m_i = 1$ we believe that x_i measures t_i correctly, if $m_i < 1$ we believe that x_i is larger than t_i and if $m_i > 1$ we believe that x_i is smaller than t_i. The value of m_i is modeled as a polynomial function of w_i:

$$m_i = \alpha_0 + \sum_{j=1}^{k} \alpha_j w_i^j \tag{2}$$

where k is the order of the polynomial.

2.1.1 Example

In this example we estimate a regression $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{4i}$, but the measurement $\mathbf{x_1}$ of $\mathbf{t_1}$ is systematically biased. Suppose the true mechanism is $t_{1i} = x_{1i} \cdot m_i$ and $m_i = 1.2 + 0.05 w_i$. Further assume our beliefs about $\mathbf{m}$ are correct but we express some uncertainty. In particular, our beliefs are $\alpha_0 \sim N(1.2, 0.05)$ and $\alpha_0 \sim N(0.05, 0.01)$. Figure 1 gives the the relation between $\mathbf{w}$ on the horizontal axis and $\mathbf{m}$ on the vertical axis. The dots are the mean beliefs and the dashed lines present the area where we are 95% confident the true value lies within. Here the belief is that $\mathbf{t_1}$ is correctly measured at low values of $\mathbf{w}$ but $\mathbf{t_1}$ is much larger than $\mathbf{x_1}$ at high values of $\mathbf{w}$ (up to almost 50%).

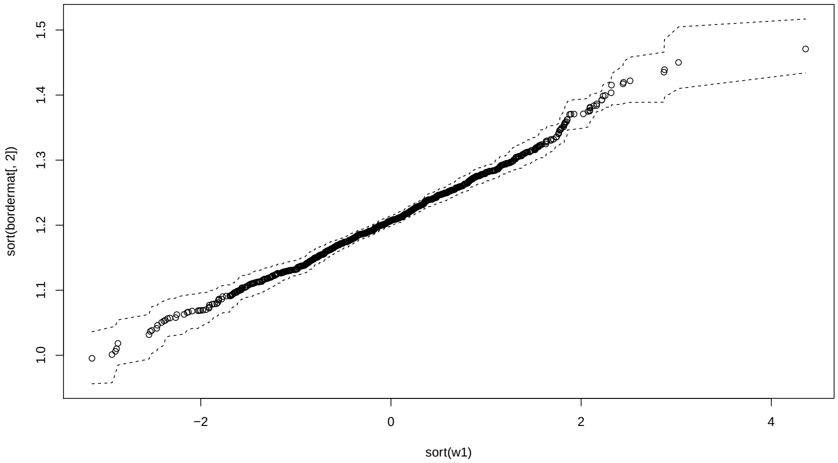

Figure 1: Description.

Figure 2 gives the posterior distribution of β_1, β_2 and β_3 for three different regressions. The black curves are the posteriors if we know the true values $\mathbf{t_1}$. The blue lines are what we get if we use $\mathbf{x_1}$ instead. Clearly the estimated coefficients are biased. Especially the estimated effect of the variable that is measured incorrectly on $\mathbf{y}$ is too large. The red lines give the posterior distributions after we corrected using our correct beliefs (based on 100 draws from our distribution of beliefs about α_0 and α_1). The posterior is relatively close to the correct posterior.

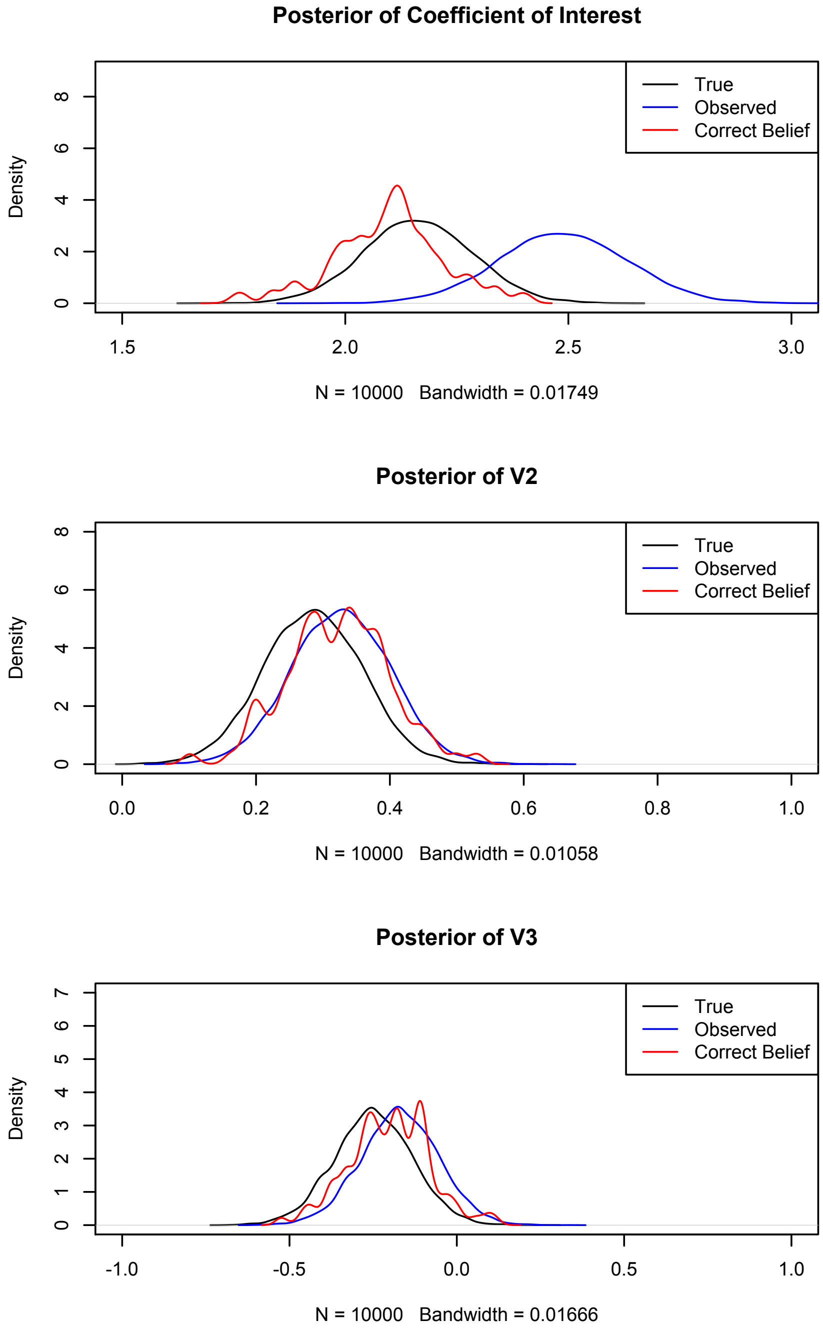

Figure 2: Description.

2.2 Categorical Variable

Suppose $\mathbf{t}$ (and its realization $\mathbf{x}$) is a categorical variable with s categories.

2.3 Binary Variable

Suppose $\mathbf{t}$ (and its realization $\mathbf{x}$) is a binary variable.