Spring 2015 POLSCI.733 MLE Midterm

Dag Tanneberg

Missing data generates severe problems for regression analysis. At the very least it diminishes efficiency because less data is available for parameter estimation. More importantly, already small amounts of missing data may bias regression results if missingness is not completely at random (MCAR), i.e. unrelated to the observed data and the parameters to be estimated. Using the example regression of standardized net income inequality on ethnic fractionalization and regime type (Polity 2) this effect becomes readily apparent from Figure 1. Each panel compares the estimated slope for either predictor under different regimes of missingness.[1] Solid lines capture the partial effect without missing data, shaded lines represent estimated slopes from 100 Amelia imputations. Dashed lines report the partial effect of either predictor under list wise deletion respectively averaged over all Amelia imputations. Finally, the obligatory seed 6886 is given by a dotted line. Figure 1 demonstrates the grave consequences of misguided missing data management.

Under full information the estimated coefficient for ethnic fractionalization is 2.01 and statistically significant at the conventional .95 level. Hence, judging from the sample higher ethnic heterogeneity tends to be associated with higher inequality. The reverse holds for Polity 2 because the partial effect of -0.096 is not statistically significant. Although higher levels of democracy tend to concur with lower levels of inequality this observation might not travel beyond the sample. The pattern of missing data systematically biases these results. Thereby the amount, but not direction, of bias depends on the management of missing data.

List wise deletion attenuates the effect of ethnic fractionalization and renders it statistically insignificant (Est.: .80, Std. Er.: .44). In contrast, the coefficient on Polity 2 increases in absolute size and becomes statistically significant (Est.: $-.16$, Std. Er.: 0.05). Thus, list wise deletion misrepresents the 'true' data generating process dramatically. Multiple imputation ranks as a best practice alternative to list wise deletion. However, using the obligatory settings for this examination things turn from bad to worse. Using seed 6886 Amelia confirms the results from list wise deletion and it increases the bias (Ethnic fractionalization: Est.: 0.33, Std. Er.: 0.45; Polity 2: Est. -0.29, Std. Er.: 0.04). As can be seen from the shaded lines this perverse result depends neither on the random seed nor on the number of imputations. Rather, while list wise deletion generally abstracts from any causal process introducing missingness, multiple imputation fails to tap into it in this exercise.

[1] In each panel the alternative predictor has been held constant at the mean of the complete data.

Figure 1: The impact of missing data on regression results