Spring 2015 POLSCI.733 MLE Midterm

Dag Tanneberg

Missing data poses severe problems to regression analysis. At the very least it diminishes estimation efficiency, because less observations are available for analysis. More importantly, however, already a small amount of missing data may bias regression results severely if missingness is not completely unrelated to the observed data and the parameters to be estimated (Missing Completely At Random, MCAR). 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 states of missingness. Solid lines denote the partial effect without any missing data, shaded lines represent estimated slopes from 100 Amelia imputations. Patterned lines denote 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.

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 political regime type because the partial effect for Polity 2 is -0.096 and not statistically significant. Although higher levels of democracy tend to concur with lower levels of inequality this observation should not be generalized from the sample. The pattern of missing data systematically biases these results and the bias depends on the strategy employed to deal with missingness.

List wise deletion attenuates the effect of ethnic fractionalization and renders it statistically insignificant (Est.: .80, Std. Error: .44). In contrast, the coefficient on Polity 2 increases in absolute size and becomes statistically significant (Est.: $-.16$, Std. Error: 0.05). Thus, list wise deletion reverts the 'true' process in the data. Multiple imputation ranks as a best practice alternative of list wise deletion. However, using the obligatory settings it actually makes things worse because it confirms the results from list wise deletion and increases the bias (Ethnic fractionalization: Est.: 0.33, Std. Error: 0.45; Polity 2: Est. -0.29, Std. Error: 0.04). This perverse result neither depends on the random seed nor on the number of imputations as can be seen from the shaded lines. Rather, while list wise deletion ignores any causal process that might introduce missingness, multiple imputation fails to tap into it in this exercise.

In conclusion, given non-MCAR data list wise deletion will yield misleading results, but multiple imputation may offer little improvement if it is not tailored to fit the process that induces missingness.

Figure 1: Demonstrating the effect of missing data