

Midterm Assignment

Soomin Oh

Discussion on the Three Models

Figure 1 illustrates the coefficient plots of the three models I have obtained using (a) the original data with no missing data, (b) missing data whose rows were listwise deleted, and (c) imputed data using Amelia. To reiterate, the dependent variable is inequality and the independent variables are polity and ethnolinguistic fractionalization (ELF).

According to the original model, ELF is a strong predictor of inequality (more ethnolinguistic fractionalization is associated with higher levels of inequality) whereas polity is not (statistically insignificant; as a sidenote, the coefficient is negative indicating that higher polity levels is associated with lower levels of inequality). The coefficient estimate for both of the independent variables across the models are different, and the difference for ELF is more pronounced. Polity, for all three models, falls within 0 and -0.3 with a narrow 90 and 95 confidence interval. One thing to note is that in the original model, coefficient for polity includes 0, so there is a chance that the result has been obtained through random chance, but for the other two models, polity falls outside 0, making them significant coefficients. Substantively, this indicates that when the data goes through listwise deletion and/or imputation, a variable that is not statistically significant could become statistically significant, posing a threat to valid interpretation.

As for ELF, it is a significant variable in the original model, though it has a large confidence interval and thus lacks accuracy in prediction. For the other two models using listwise deletion and imputation, the variable is not statistically significant, i.e. the confidence interval contains 0. This shows that listwise deletion and imputation can lead the coefficient to become insignificant although the original data says it is significant.

This exercise was fruitful in enhancing one's ability to know the mechanics of OLS and to see what listwise deletion and imputation do to the model, but there are a few things that make the interpretation difficult. First, the data with NAs lack country names, which make it hard to see which countries were dropped. For example, it could be the case that a certain group of countries with lower ELF were dropped hence making ELF insignificant when deleted listwise. Second, the lack of control variables or a theoretical background to the empirical test makes it hard to make substantive comments other than stating the visible results. One last comment is that neither listwise deletion nor imputation come close to bringing the empirical results close to the original model.

Figure 1: Coefficient Plots of the Models

