The following document presents a discussion of the results of three OLS models that apply the same model specifications to three subtly different data sets. The first model, the "Baseline" model, predicts a country's standardized Gini coefficient based on the country's Polity score and Ethnolinguistic fractionalization. The model estimates the coefficient of ethnolinguistic fractionalization to be approximately 2, indicating that a one-unit increase in the ethnolingiusitic fractionalization index should correspond to an approximately two-unit increase in estimated standardized Gini coefficients.

The second, the "Missing Data" model, carries out the same regression, but employs list-wise deletion of observations that have missing values. The third model, the "Amelia Model," employs the Amelia protocol to impute the missing data. Figure 1 presents a coefficient plot for the estimated coefficients and 95^{th} percent confidence interval. Both the Missing Data and Amelia models find statically significant, although substantively minor, negative effects for a one-unit decrease for the Polity score on the standardized Gini coefficient. In these models, each one-unit decrease in Polity2 score corresponds to an approximately 0.1 and 0.15 decrease in predicted standardized Gini score.

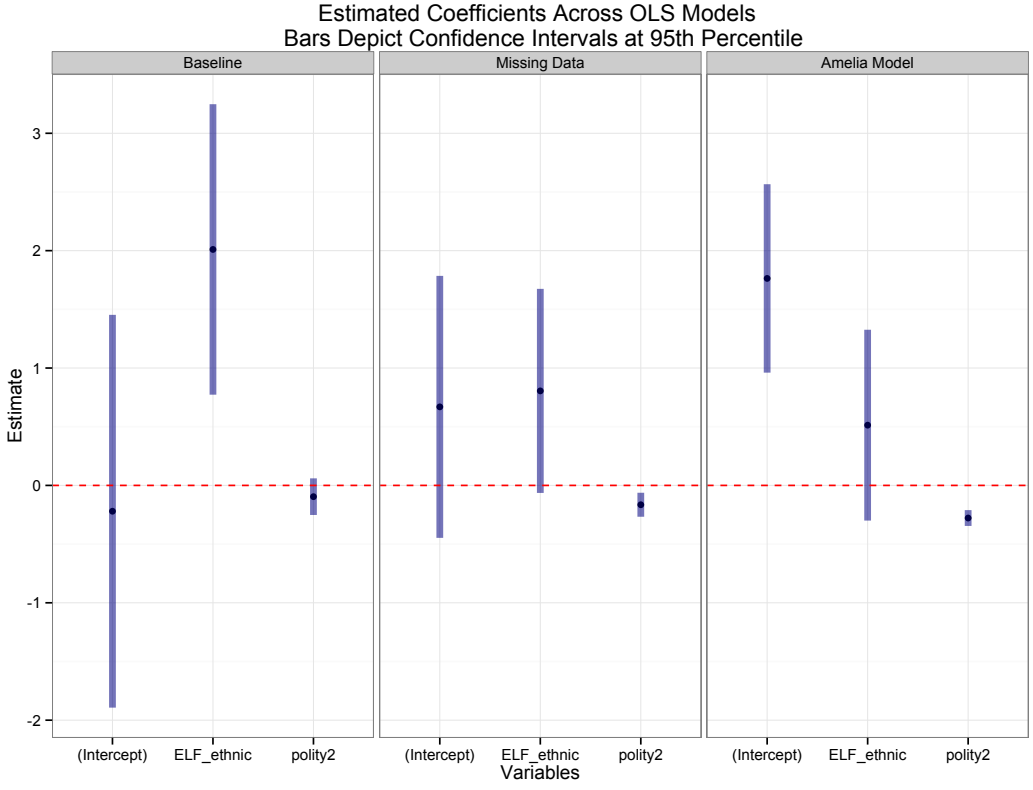

Figure 1: Coefficient Estimates, OLS Models

A results table for each of the three models is available in the Appendix.

One of the most immediately evident elements of the three models is that the Amelia model is closer to the model that uses listwise deletion. Neither the missing data model nor the Amelia model perform well on the ELF and Polity variables: both fail to find a statistically significant relationship for ELF and report a significant relationship for polity 2 where none should exist. Moreover, intercept changes dramatically between the full data model and the models using missing and imputed data. Taken together, these patterns are suggestive of a systematic difference between the full dataset and the data set with missing values. Without a systematic difference between the full and missing datasets, the results from the model that uses imputed data should be more similar to the full model. Interestingly, the confidence intervals become progressively smaller through the three models, giving the illusion of increasing precision.[1]

An investigation into the missing data revealed an underlying pattern: each of the six missing observations occur for the Polity variable when the Gini coefficient is 1 or greater. Figure 2 presents side-by-side visualizations of the full data set and the data set with missing data. This visualization emphasizes that the missing data is not randomly missing—a key assumption for data imputation to be effective. Imputation protocols, such as Amelia, create values for missing data by leveraging information from the entire data set. Essentially, imputation reinforces the underlying tendencies of the data. This is not a problem if data is truly randomly missing. However, imputation is vulnerable to data removal that changes the structure of the underlying data, which has the effect of distorting both the regression and the basis for imputation.

[1]The numeric magnitude of the respective confidence intervals can be seen in the Appendix.

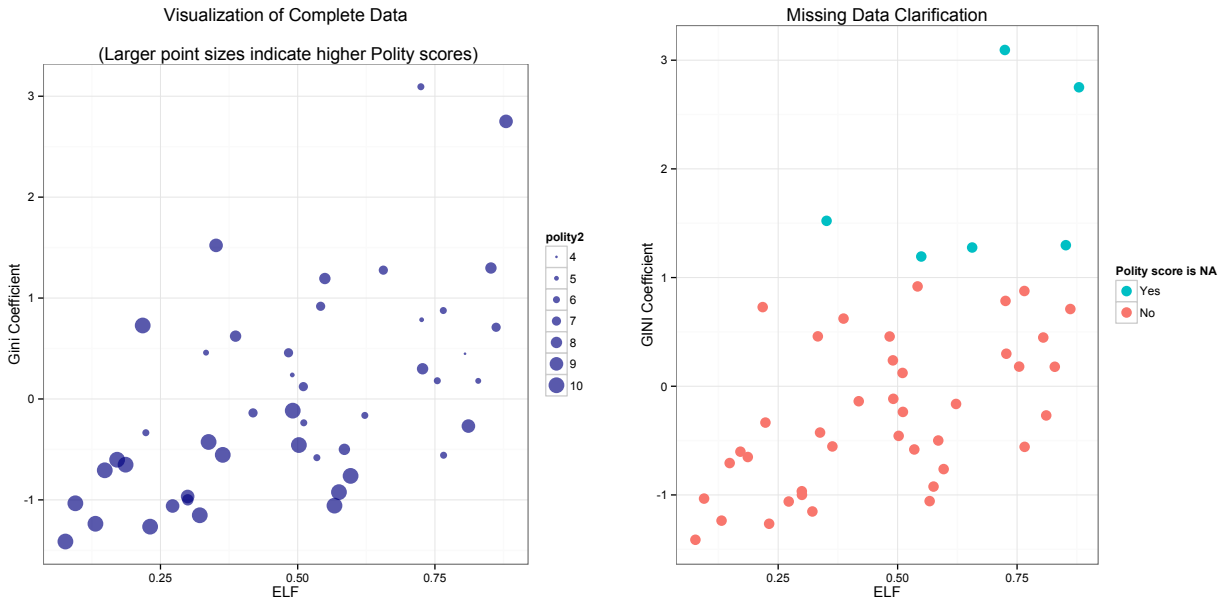

Figure 2: Full dataset (left) compared to data with missing values (right)

Numerical Appendix

The Numerical Appendix presents the estimated coefficients for the three OLS models. The final table quantifies the diminishing sizes of the confidence intervals for each estimated coefficient.

Full Model

	Estimate	Std.Error	T-Statistic	P-value	Lower95	Upper95	ModelID
(Intercept)	-0.22	0.85	-0.26	0.80	-1.89	1.45	Baseline
ELF_ethnic	2.01	0.63	3.18	0.00	0.77	3.25	Baseline
polity2	-0.10	0.08	-1.20	0.24	-0.25	0.06	Baseline

Listwise Deletion Model

	Estimate	Std.Error	T-Statistic	P-value	Lower95	Upper95	ModelID
(Intercept)	0.67	0.57	1.18	0.25	-0.45	1.79	Missing Data
ELF_ethnic	0.81	0.44	1.82	0.08	-0.06	1.67	Missing Data
polity2	-0.16	0.05	-3.16	0.00	-0.27	-0.06	Missing Data

Imputed Data Model

	Estimate	Std.Error	T-Statistic	P-value	Lower95	Upper95	ModelID
(Intercept)	1.63	0.70	2.34	0.02	0.27	3.00	Amelia Model
ELF_ethnic	0.79	0.59	1.34	0.19	-0.37	1.94	Amelia Model
polity2	-0.27	0.06	-4.32	0.00	-0.39	-0.14	Amelia Model

Coefficient estimation precision

Table 1: Size of confidence interval for variables on the three models

Model	Variable 1	Variable 2	Variable 3
Full Model	Intercept: 3.346	ELF ethnic: 2.475	polity2: 0.312
Listwise Deletion	Intercept: 2.232	ELF ethnic1 1.739	polity2 0.204
Amelia Imputation	Intercept: 1.605	ELF ethnic: 1.626	polity2: 0.135