Accounting for Coding Uncertainty in Political Science Data

Anton Strezhnev, Dana Higgins, Connor Huff | Harvard University | astrezhnev@fas.harvard.edu, danahiggins@fas.harvard.edu, cdezzanihuff@fas.harvard.edu

Motivation

How can social scientists assess how uncertainty in variable coding decisions affects their results?

- Social scientists frequently make use of datasets where one or more variables are generated through human coding decisions.
- However, the process of coding these cases is often extremely difficult (trust us, we have done it).
- Events of interest to political science researchers are often shrouded in secrecy, there is ambiguity about the exact chain of events, and there are incentives for political actors to misrepresent political events.
- These difficulties make the process of determining the correct code a particular variable should receive for a given event a challenging endeavor.
- Uncertainty in the coding of independent variables has the potential to bias regression results.
- **We propose a tool that allows researchers to focus their data validation efforts and assess the impact of coding uncertainty on their results.**

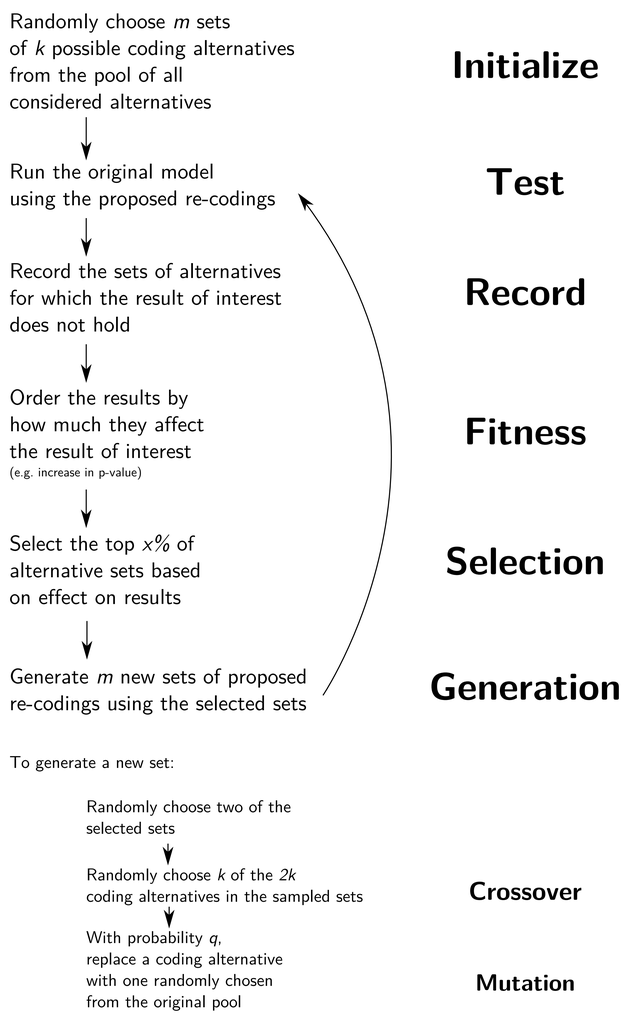
Finding Sensitive Codings

- Researchers have a limited amount of time to spend re-checking data coding decisions. How can this effort be focused on the "optimal" set of cases to validate?
- Solution: "Robustness checks" - Re-run the model many times on datasets that use alternative coding decisions and investigate those that "break" the original result.
- This is easy if we only consider one observation at a time - we run n regressions. Once we consider more than one alternative at a time, an exhaustive search becomes computationally intensive
- For k alternative codings of a binary variable per test, we need to run $\binom{n}{k}$ regressions.
- For a medium-sized political science dataset of about 700 observations and $k = 3$, an exhaustive search would take $\binom{700}{3} = 56,921,900$ regressions.
- Assuming an optimistic $.025$ seconds per iteration in `R`, **this would take around 395 hours or over 2 weeks**.
- If our coded variable is not dichotomous, this task becomes even more time-consuming!

A Genetic Search Algorithm

- Our goal is to find at least one set of alternative codings that eliminates the substantive finding of our original regression.
- The search space can be substantially limited by pruning alternatives that are unlikely to negatively impact the results.
- We propose a genetic search algorithm to implement a more efficient search across the space of possible sensitive coding decisions.
- The intuition is that we focus our regressions on variations on those sets of alternatives that have the largest influence on the results.

Figure 1: A Genetic Search Algorithm for Finding Sensitive Coding Sets



Results

- We apply our method to Hyde and Marinov (2014) "Information and self-enforcing democracy: The role of international election observation."
- The main empirical finding is that protests after elections in weakly institutionalized democracies are more likely when international election monitoring organizations issue negative reports about the fairness of the election than when they do not.
- Both independent and dependent variables are coded from analyses of election events (the NELDA database - Hyde and Marinov (2012)). We evaluate the robustness of coding decisions for the main "negative report" variable which is

  *...coded from the official reports and press releases from international observers, and equal to 1 if observers seriously questioned the winner of the election or the legitimacy of the process (165 election events). Most observer reports include some criticism, and **only those statements that are quite critical** are considered a NEGATIVE REPORT (Hyde and Marinov, 2014 p. 343).*

- Using the provided replication data, we estimate the full logistic regression model in Hyde and Marinov (with all controls) and obtain a positive coefficient estimate for the "negative report" variable on post-election protest.

Table 1: Replicated coefficient estimate from Hyde and Marinov (2014)

Variable	Logit Coefficient	Standard Error	p-value
Negative Report	0.828	0.378	0.0284

- We find that if we consider only one alternate coding at a time, the result remains - which is promising.
- When we look at two or three alternative codings simultaneously, we find a number of cases that might break the results (assuming a $.05$ threshold for rejection of the null). Two examples are

Table 2: Robustness Check - Two re-coding proposals

Case	Original Coding	Proposed Alternative
Cote d'Ivoire - 2000 Executive Election	1	0
Ethiopia - 2005 Legislative Election	1	0

Variable	New Coefficient Estimate	Standard Error	p-value
Negative Report	0.678	0.378	0.072

Table 3: Robustness Check - Three re-coding proposals

Case	Original Coding	Proposed Alternative
Georgia - 2003 Parliamentary Election	1	0
Niger - 1996 Executive Election	1	0
Ethiopia - 2005 Legislative Election	1	0

Variable	New Coefficient Estimate	Standard Error	p-value
Negative Report	0.604	0.381	0.113

- Luckily for the authors, the original codings appear very reasonable. In all of these cases, a review of the historical record shows very strong criticisms from at least one election observer.