Bounds for Logistic Regression Coefficients with Nonignorable Missing Outcomes

Brenton Kenkel
Department of Political Science, University of Rochester

Abstract

I develop a new method to estimate logistic regression coefficients when there is nonignorable missingness or measurement error in the outcome variable. The estimator finds the set of all coefficient vectors that could be obtained under any assumption about the missing outcomes.

Methodological Approach: Partial Identification

There is a binary outcome Y generated by a logistic model,

$$\Pr(Y_i = 1) = \frac{1}{1 + \exp(-x_i'\beta)} = \Lambda(x_i'\beta),$$

and the goal is to estimate β. The indicator variable U_i represents whether Y_i is unobserved. The missingness-generating process is an unknown function φ, where $\Pr(U_i = 1) = \varphi(x_i, Y_i)$.

If $U_i = 0$ for all observations, we can estimate β via maximum likelihood, solving for

$$\frac{\partial \ell}{\partial \beta_j} = X_j' \left(Y - \Lambda(\mathbf{X}\beta)\right) = 0, \qquad j = 1, \ldots, p.$$

Otherwise, we must either assume missingness at random or specify the exact missingness-generating process to obtain point estimates for the coefficients.

Instead, I use a partial-identification approach to obtain bounds on the coefficients. For each observation, define the bounds Y_i^0 and Y_i^1 as

$$(Y_i^0, Y_i^1) = \begin{cases} (0, 1) & \text{if } U_i = 1 \\ (Y_i, Y_i) & \text{if } U_i = 0 \end{cases}$$

Combining these with the maximum likelihood score function (similar to Manski and Tamer 2002), we can estimate the "identified set" as the set of all β such that

$$\begin{aligned} Z_j' \left(Y^0 - \Lambda(\mathbf{X}\beta)\right) &\leq 0 \\ Z_j' \left(\Lambda(\mathbf{X}\beta) - Y^1\right) &\leq 0, \end{aligned} \qquad j = 1, \ldots, p,$$

where Z_j is a positive-valued affine transformation of X_j.

Computation

With more than two covariates, a grid search is infeasible. I use the following sampling procedure:

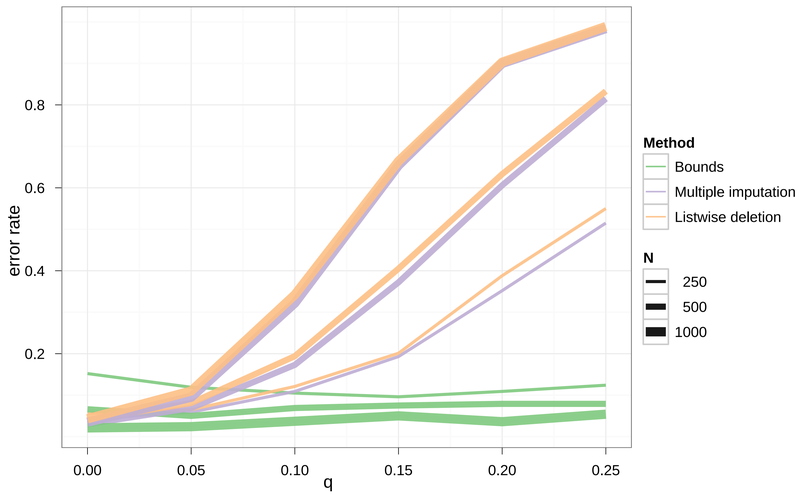
1. Randomly "fill in" the missing outcome values and run logistic regression to compute initial estimate $\hat{\beta}$ and variance matrix $\hat{\mathbf{\Sigma}}$
2. Sample from $N(\hat{\beta}, c\hat{\mathbf{\Sigma}})$ and check whether each draw meets the condition, until enough members of the identified set have been obtained.

Monte Carlo Simulation

In each of 1,000 replications (per combination of N and q),

1. Data and missingness for N entries generated by
$$\Pr(Y_i = 1) = \Lambda(-1 + X_{2i})$$
$$\Pr(U_i = 1) = 0.1 + q \cdot I(Y_i = X_{1i}),$$
where $X_{1i} \sim \text{Bernoulli}(0.5)$ and $X_{2i} \sim N(1, 1)$.
2. Record if estimated bounds on X_1 coefficient contain 0 (true value).
3. Record if coefficient on X_1 is statistically significant under listwise deletion and multiple imputation.

Results



Areas of Use

The method can be used for both nonignorable missingness and measurement error (when the potentially mismeasured outcomes are known). Some use cases:

- Nonignorable nonresponse in survey data
- To avoid arbitrary decisions in data collection when some outcomes do not meet prespecified coding rules
- To establish robustness of a result under different measures of the same outcome variable, or when scholars disagree about how to treat a set of cases

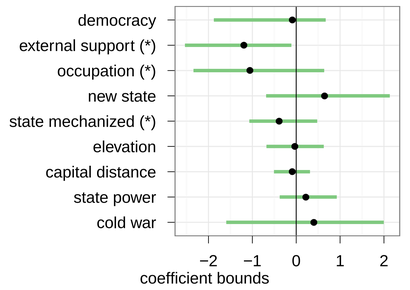
In the last two cases, applied researchers commonly drop the troublesome observations from the dataset—which causes bias for the same reasons as in standard missing-data situations.

Application 1: Lyall (2010) on Counterinsurgencies

Y: successful counterinsurgency ($N = 286$, with 153 successes)

U: "difficult to code" draw (39 instances)

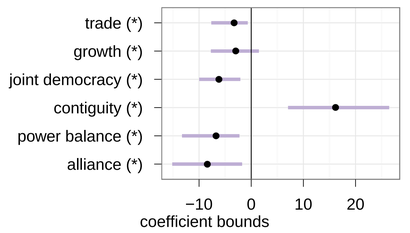
Lines: Estimated bounds
Points: Original coefficients
Starred: Statistically significant in original analysis



Application 2: Oneal and Russett (1997) on Liberal Peace

Y: militarized international dispute onset ($N = 20{,}990$, with 405 disputes)

U: ongoing dispute (542 instances)



Interpreting the Results

The bounds contain the set of coefficients that could be obtained as a *point estimate* under some assumption about the missing values—no population inference is implied.

If the bounds contain 0, the sample definitely does not provide evidence in favor of a directional hypothesis. However, failure to contain 0 only establishes that the sign of the sample estimate is robust to missingness/measurement error.

References

Jason Lyall. 2010. "Do Democracies Make Inferior Counterinsurgents? Reassessing Democracy's Impact on War Outcomes and Duration." *International Organization*, 64(1): 167–192.

Charles F. Manski and Elie Tamer. 2002. "Inference on Regressions with Interval Data on a Regressor or Outcome." *Econometrica* 70(2): 519–546.

John R. Oneal and Bruce M. Russett. 1997. "The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950–1985." *International Studies Quarterly* 41(2): 267–293.