The problem of the rareness of the outcome variable

Wonjae

What is the problem?

One of the inherent problems in conflict data is the relative rareness of events (Gates 2001; King & Zeng 2001). Rare events indicate "binary dependent variables characterized as by dozens to thousands of times fewer ones (events such as wars or coups) than zeroes (nonevents)" (King & Zeng 2001, 693).

Why is it problematic?

The basic problem is having a number of units in a panel that have no events. This means that "the country-specific indicator variables corresponding to the all-zero countries perfectly predict the zeroes in the outcome variable (Gates 2001; King & Zeng 2001; King 2001). Perfect prediction.

The Sources of the Problem

According to King & Zeng (2001), the source of the rare events problems can be summarized in two ways. The first one is "researchers' reliance on logit coefficients, which are biased in small samples." The method of computing probabilities of events in logit analysis can be inappropriate in finite samples of rare-events data, leading to errors in the same direction as biases in the coefficients. Thus, it causes underestimation of event probabilities. Even in some cases, "the error can be as large as the reported estimated effects."

A second source of problems in analyzing rare events lies in "how data are collected." Confronting the tradeoff between gathering more observations and including better or additional variables, researchers tend to choose very large data sets with few explanatory variables.

How to correct it?

The simplest way of correcting the problem is decreasing the rareness of the event. By lowering the threshold of what constitutes events, expanding the data selection period, or other ways, we may reduce the need to correct for rareness.

Meanwhile, King & Zeng suggest a way of correcting the rare events problem. That is, by selecting data on the dependent variables, they insist that we can increase the efficiency of subsequent data collections by changing the optimal tradeoff in favor of fewer observations and more sophisticated measures that more closely reflect the desired concepts. They have devised a way of correcting rare events problem in "Relogit program."

Conclusion

Collier & Hoeffler (2001) report that accounting for rareness makes no substantial difference to their results. Hopefully, rareness might not a serious problem in many cases (Gates 2001).

References

Collier, P and A Hoeffler (1998) 'On Economic Causes of Civil War', *Oxford Economic Papers*, 50, 563-573.

Green, DP, SY Kim, and DH Yoon (2001) 'Dirty Pool', *International Organization* 55, 441-468.

King, G (2001) 'Proper Nouns and Methodological Propriety: Pooling Dyads in International Relations Data', *International Organization* 55, 497-507.

King, G and L Zeng (2001) 'Explaining Rare Events in International Relations', *International Organization* 55, 3, Summer, 693 – 715.

King, G and L Zeng (2000) 'Logistic Regression in Rare Events Data', *Ipolitical Analysis* 9, 2.

Oneal, JR and B Russett (2001) 'Clear and Clean: The Fixed Effects of the Liberal Peace', *International Organization* 55, 469-485.

Gates, Scott. 2001. “Empirically assessing the causes of civil war.” Working paper.