Logistic Regression in Rare Events Data

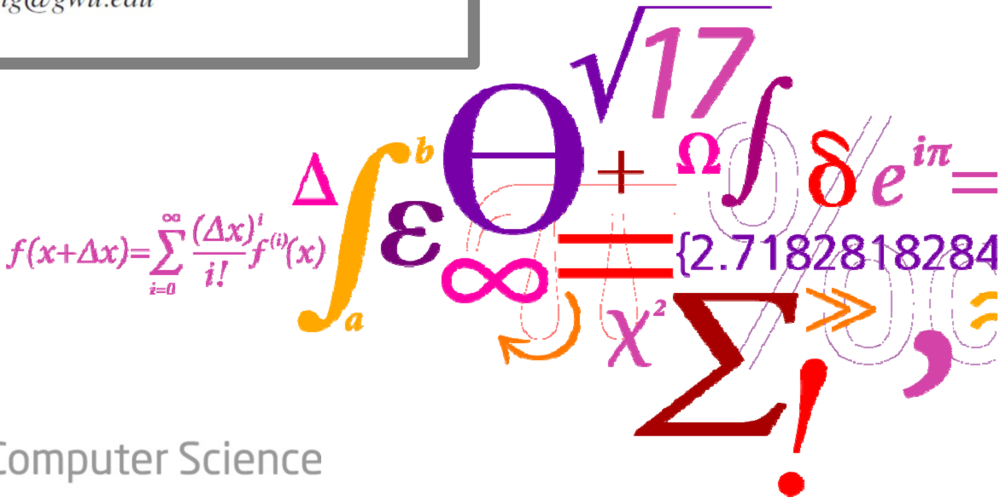
Presented by Morten Mørup

Political analysis, 2001

Logistic Regression in Rare Events Data

Gary King
Center for Basic Research in the Social Sciences, 34 Kirkland Street, Harvard University, Cambridge, MA 02138
e-mail: King@Harvard.Edu
http://GKing.Harvard.Edu

Langche Zeng
Department of Political Science, George Washington University, Funger Hall, 2201 G Street NW, Washington, DC 20052
e-mail: lzeng@gwu.edu




Political analysis, 2001

Logistic Regression in Rare Events Data

Gary King
Center for Basic Research in the Social Sciences, 34 Kirkland Street, Harvard University, Cambridge, MA 02138
e-mail: King@Harvard.Edu
http://GKing.Harvard.Edu

Langche Zeng
Department of Political Science, George Washington University, Funger Hall, 2201 G Street NW, Washington, DC 20052
e-mail: lzeng@gwu.edu

Abstract

We study rare events data, binary dependent variables with dozens to thousands of times fewer ones (events, such as wars, vetoes, cases of political activism, or epidemiological infections) than zeros ("nonevents"). In many literatures, these variables have proven difficult to explain and predict, a problem that seems to have at least two sources. First, **popular statistical procedures, such as logistic regression, can sharply underestimate the probability of rare events.** We recommend corrections that outperform existing methods and change the estimates of absolute and relative risks by as much as some estimated effects reported in the literature. Second, **commonly used data collection strategies are grossly inefficient for rare events data**. The fear of collecting data with too few events has led to data collections with huge numbers of observations but relatively few, and poorly measured, explanatory variables, such as in international conflict data with more than a quarter-million dyads, only a few of which are at war. As it turns out, more efficient sampling designs exist for making valid inferences, such as sampling all available events (e.g., wars) and a tiny fraction of nonevents (peace). This enables scholars to save as much as 99% of their (nonfixed) data collection costs or to collect much more meaningful explanatory variables. We provide methods that link these two results, enabling both types of corrections to work simultaneously, and software that implements the methods developed.

Logistic regression

$$Y_i \sim \text{Bernoulli}(Y_i \mid \pi_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$\pi_i = \frac{1}{1+e^{-\mathbf{x}_i \beta}}$$

$$L(\beta \mid \mathbf{y}) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$\ln L(\beta \mid \mathbf{y}) = \sum_{\{Y_i=1\}} \ln(\pi_i) + \sum_{\{Y_i=0\}} \ln(1-\pi_i)$$

$$= -\sum_{i=1}^{n} \ln\left(1 + e^{(1-2Y_i)\mathbf{x}_i \beta}\right)$$

$$\hat{\beta} = \underset{\beta}{\text{argmax}} \ln L(\beta \mid \mathbf{y})$$

Variance of the ML estimate

$$V(\hat{\beta}) = \left[\sum_{i=1}^{n} \pi_i (1 - \pi_i) \mathbf{x}_i' \mathbf{x}_i \right]^{-1}$$

Rare events more informative?

In general yes as π_i closer to 0.5

Data collection strategies

Case-control design

Correcting prior

$$\hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$$

τ: Fraction of ones in the population

$\bar{y}$: Fraction of ones in the sample

Weighting loss function

$$\ln L_w(\beta \mid \mathbf{y}) = w_1 \sum_{\{Y_i=1\}} \ln(\pi_i) + w_0 \sum_{\{Y_i=0\}} \ln(1-\pi_i)$$
$$= -\sum_{i=1}^{n} w_i \ln\left(1 + e^{(1-2y_i)\mathbf{x}_i\beta}\right)$$

$$w_1 = \tau/\bar{y}$$
$$w_0 = (1-\tau)/(1-\bar{y})$$
$$w_i = w_1 Y_i + w_0(1-Y_i)$$

Issue of bias in ML estimate facing rare events

STATISTICS IN MEDICINE, VOL. 2, 71–78 (1983)

BIAS CORRECTION IN MAXIMUM LIKELIHOOD LOGISTIC REGRESSION

ROBERT L. SCHAEFER
Department of Mathematics and Statistics, Miami University, Oxford, Ohio, U.S.A.

$$\text{bias}(\hat{\beta}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\xi$$

$$\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'.$$

$$\xi_i = 0.5Q_{ii}[(1+w_1)\hat{\pi}_i - w_1]$$

$$\mathbf{W} = \text{diag}\{\hat{\pi}_i(1-\hat{\pi}_i)w_i\}$$

$$\tilde{\beta} = \hat{\beta} - \text{bias}(\hat{\beta}) \qquad V(\tilde{\beta}) = (n/(n+k))^2 V(\hat{\beta})$$

Bayesian averaging

$$\Pr(Y_i = 1) = \int \Pr(Y_i = 1 \mid \boldsymbol{\beta}^*) P(\boldsymbol{\beta}^*) d\boldsymbol{\beta}^*$$

$$P(\beta^*) \sim N(\tilde{\beta}, var(\tilde{\beta}))$$

$$\Pr(Y_i = 1) \approx \tilde{\pi}_i + C_i$$
$$C_i = (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1 - \tilde{\pi}_i)\mathbf{x}_0 V(\tilde{\beta})\mathbf{x}_0'$$

Issue of tails facing rare events