题目：On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data
作者：Francisco Louzada，Paulo H. Ferreira-Silva，Carlos A.R. Diniz

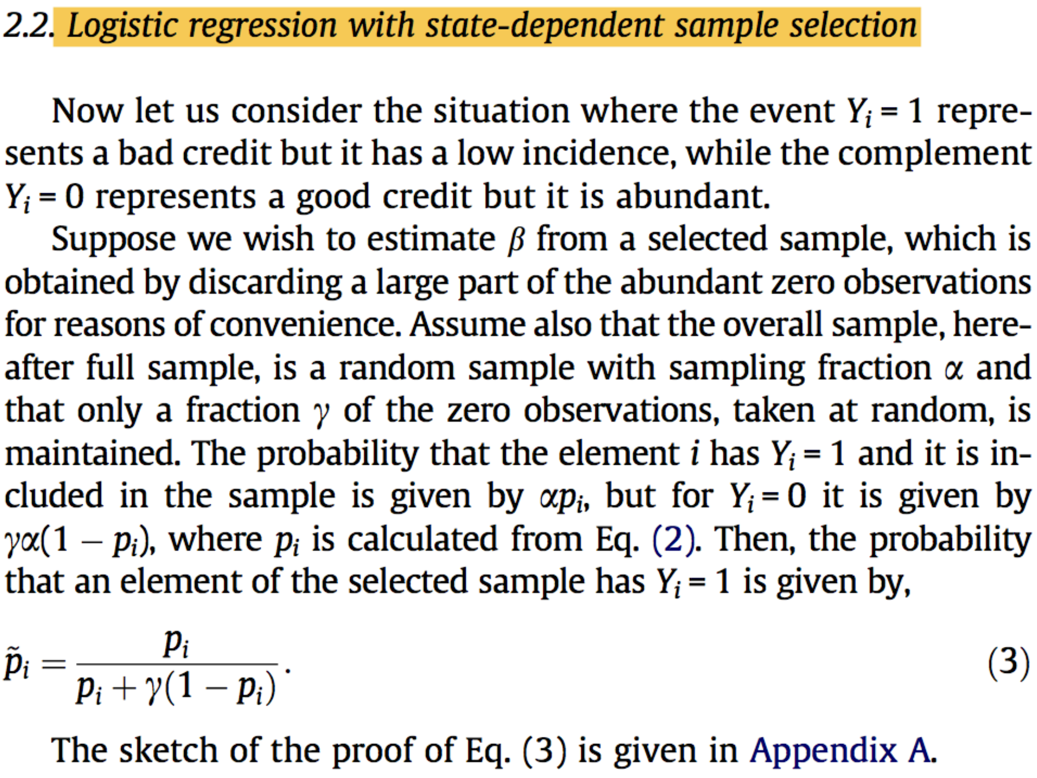
领域：信用评分

核心：两种回归模型(朴素逻辑回归模型和状态依赖于样本选择的逻辑回归模型)应用于模拟数据的性能

实现方法：

一、状态依赖于样本选择的逻辑回归模型建立

该模型的建立已在[Cramer，J. S. (2004). Scoring bank loans that may go wrong: A case study. Statistica Neerlandica，58(3)，365–380.]中提出。



2.2. Logistic regression with state-dependent sample selection

Now let us consider the situation where the event $Y_i = 1$ represents a bad credit but it has a low incidence, while the complement $Y_i = 0$ represents a good credit but it is abundant.

Suppose we wish to estimate β from a selected sample, which is obtained by discarding a large part of the abundant zero observations for reasons of convenience. Assume also that the overall sample, hereafter full sample, is a random sample with sampling fraction α and that only a fraction γ of the zero observations, taken at random, is maintained. The probability that the element i has $Y_i = 1$ and it is included in the sample is given by αp_i, but for $Y_i = 0$ it is given by $\gamma\alpha(1-p_i)$, where p_i is calculated from Eq. (2). Then, the probability that an element of the selected sample has $Y_i = 1$ is given by,

$$\tilde{p}_i = \frac{p_i}{p_i + \gamma(1-p_i)}. \qquad (3)$$

The sketch of the proof of Eq. (3) is given in Appendix A.

二、结论

1. 建议在建立信用评分模型时使用平衡样本；
2. 理想情况下是使用平衡样本并使用状态依赖于样本选择的逻辑回归。