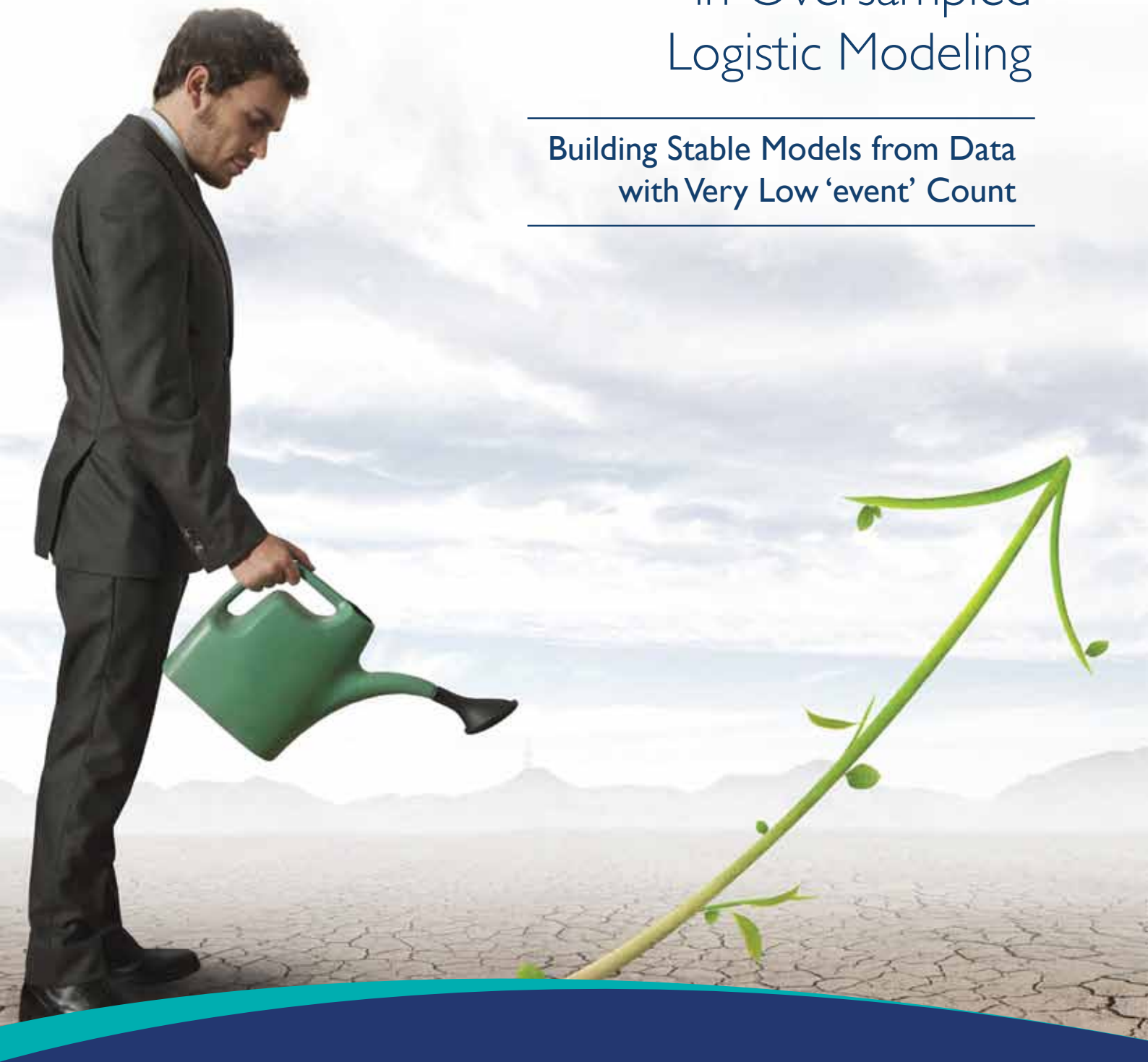
Correcting Sample Bias in Oversampled Logistic Modeling

Building Stable Models from Data with Very Low 'event' Count



ABSTRACT

In binary outcome regression models with very few 'bads' or 'minority events', it becomes difficult to build a classification model directly. Instead it may seem feasible to build a model by increasing the number of events through oversampling from the set of 'events' to build a new data set for developing the model. However, the coefficients must be translated to capture the values that would come if no oversampling was done for the estimation. This paper provides a standard formula for this purpose and a derivation for the same. The technique outlined finds immense use in modeling credit defaults or modeling to identify responders in a marketing campaign, where such events are very rare.

Such scorecards created using the new and improved formula accounting the 'rare' events find immense application in the arena of Marketing Analytics when there is a dearth in survey responses, such that a representative sample is hard to come by. These scorecards will create a consistent estimator of change of opinions, reflecting a more accurate depiction of customers' wants, needs and opinions.

Such scorecards also find applicability in specific emerging economies such as the ones in the Middle East where the occurrence of a default event is very low such that there is a sample bias problem. A consistent estimator would provide a more accurate scorecard (behavioral / application) that would help banks in more efficiently interpreting the characteristics of customers.

This whitepaper aims to benefit statisticians and managers delving into data to bring out key insights and patterns be it in the area of risk and/or marketing. It presents and derives a simple formula attributed to Manski and Lerman (1977) that helps them develop and validate classification models when the occurrence of the event of interest (e.g. default, purchase of a product) is low.

TABLE OF CONTENTS

1. BACKGROUND

Binary models are extremely valuable when it comes to classifying observations in a population. These models are geared at identifying an 'event' from a set of 'non-events'. For example, from a population of active credit card holders, it makes sense to keep a note of people who have defaulted on payments in the past. Default, in this case may be referred to as an 'event', and as a corollary, non-default qualifies for a 'non-event'. An entire class of statistical models have been developed and effectively deployed for such problems involving only two types of outcomes and are known as binary-choice models. Logistic regression is one of the most frequently used binary-choice models.

In real-life problems when the number of events is very few, it becomes difficult to develop a reliable profile of an observation / customer on any set of available variables that eventually makes the observation / customers translate to an 'event'. To do this, there is a need to'over-sample' the 'events' from the population to build a fairly representative sample on which to base the development of a classification model.

However, the coefficients of the key mutually exclusive drivers for an 'event' derived from such a sample need to be 'scaled back', so that a realistic strength of each driver is calculated that is valid outside the over-sampled modeling data.

2. SOLUTION

It has been proved that oversampling of the 'events' does not change the consistency of the coefficients that relate to independent variables. However an adjustment needs to be applied to the intercept term in a logistic regression to keep it consistent.

Let βo denote the intercept estimated by the model after oversampling then the following correction needs to be performed on it:

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

Where:

τ: Fraction of 'events' in population

$\bar{y}$: Fraction of 'events' in sample

Clearly this corrected intercept estimate equals an uncorrected intercept only when the sample had been randomly sampled from the population, indicating the absence of any oversampling exercises.

3. PROOF OF THE SOLUTION

Let

F: Occurrence of an event

α: Unconditional Probability of an event in sample

τ: Conditional Probability of an event in sample

β: Unconditional Probability of an event in population

μ: Conditional Probability of an event in population

S: Occurrence of a non-event

1-α: Unconditional Probability of a non-event in sample

1-β: Unconditional Probability of a non-event in population

θ: P(X | S) / P(X | F)

X: Data

Applying Bayes formula we have for the sample:

$$
\begin{aligned}
& p(F \mid X) \\
& = p(X \mid F) * p(F) / p(X) \\
& = P(X \mid F) * P(F)[P(X \cap F) + P(X \cap S)]^{-1} \\
& = P(X \mid F) * P(F)[P(X \mid F) * P(F) + P(X \mid S) * P(S)]^{-1} \\
& = [1 + (P(X \mid S) * P(S)) / P(X \mid F) * P(F)]^{-1} \\
& = [1 + P(X \mid S) * (1-\alpha) / P(X \mid F)\alpha]^{-1} \\
& = [1 + \theta(1-\alpha) / \alpha]^{-1}
\end{aligned}
$$

Or

$$\theta = [\alpha / (1-\alpha)] * [1 - p(F \mid X)] / p(F \mid X)$$

$$\theta = [\alpha / (1-\alpha)] * [1 - \tau] / \tau$$

Equating the two expressions for θ we get

$$\ln(\alpha / (1-\alpha)) - \ln(\beta / (1-\beta)) = \ln[(1-\mu)\tau / (1-\tau)\mu]$$

$$\ln(\beta / (1-\beta)) = \ln(\alpha / (1-\alpha)) - \ln[(1-\mu)\tau / (1-\tau)\mu]$$

Thus from the above equation, to arrive at the Unconditional log-odds of the population from the Unconditional log-odds of the sample, only a constant amount that is a function of the Conditional log-odds must be deducted.

It must be noted that log-odds is same as the cross-product of the Coefficient Vector and Data Matrix. Additionally, the above result being just a shift in levels, it does not alter rank-ordering of the model.

However changes if any in the scores of a model due to induced changes in 'event' rates must be evaluated through changes in Unconditional log-odds between the sample created and the population. This is because both the sample and the population may be snapshots of two processes evolving over time and hence the distinction between conditional and unconditional parameters for both the sample and the population. Recognizing this is characteristic to any Bayesian analysis.

4. CONCLUSIONS

The technique outlined finds immense application in any situation where the event under study is very rare.

In the area of Credit, incidences of default on loan payments is a rarity, especially so in regions such as the Middle East or many other emerging economies. From the Marketing arena, response to marketing campaigns is at times very low that they do not provide a reasonable size of the population of 'events' or 'purchasers' to enable development of dynamically valid and robust statistical models. Due to the generality of this issue, the presently outlined solution is applicable in any industry wherever there is a behavioral aspect at the customer level that we are attempting to predict and which is rare.

5. BIBLIOGRAPHY

Logistic Regression in Rare Events Data, Gary King and Langche Zeng, Society for Political Methodology, 2001

Manski, Charles F., and Steven R. Lerman. 1977. "The Estimation of Choice Probabilities from Choice Based Samples." Econometrica 45(8): 1977 – 1988