Logistic Regression for Extremely Rare Events

Christian Westphal*

April 24, 2013

Abstract

Objectives: The quantitative analysis of extremely rare events and factors influencing these events poses some difficulties. The objective of my paper is to evaluate logistic regression for events millions times more rare than non-events.

Methods: Based on former theoretical and experimental results a simulation study is conducted. A specialized software is developed and supplied with this paper.

Results: With a power of 0.98 logistic regression is capable of identifying increased chances of occurrence for not too large relative risks when occurrence is extremely rare (on in ten millions).

Conclusion: Logistic regression may be used to study extremely rare events. For generic research questions the software provided with this paper may be used to study how logistic regression will perform in a specific setting.

1 Introduction

The occurrence of binary events is often attributed to influencing factors. In econometrics we usually speak of choice models (see Manski and Lerman, 1977). In epidemiology these models are called incidence models (see Prentice and Pyke, 1979). As King and Zeng (2001b) point out, when occurrence (or non-occurrence) is rare, collecting a random sample with even one occurrence may

*westphal@staff.uni-marburg.de
become prohibitively expensive. Prentice and Pyke (1979) and Manski and Lerman (1977) have shown that collecting the occurrences and adding a random sample of non-occurrences (or vice versa) allows for consistent estimation of the logistic regression parameters from this case-control study. A very good summary about statistical methods in conjunction with case-control studies can be found in Breslow (1996). These methods have been used by King and Zeng (2001b), accompanied by an intuitive explanation and application in King and Zeng (2001a), for the case of rare events, where King and Zeng describe these as "dozens to thousands of times fewer ones [...] than zeroes" (King and Zeng, 2001b: 138). I am interested in how these methods behave in finite samples when the occurrence is millions to tens of millions times more rare than non-occurrence. This is for example the case with extremely rare delinquency decisions in business crime or in "targeted school violence" as defined by Vossekuiil et al. (2002) when observed over many years and under suitable definition of the population and certain further assumptions.\(^1\)

Therefore in section 2 the statistical methods are summarized. Simulation is the chosen method for studying the problem and the simulation carried out is described in section 3. Discussion of the results can be found in section 4 with a conclusion and an outlook in sections 5 and 6. Logistic regression shows to be promising for the quantitative analysis of even extremely rare events.

\(^1\)For methods other than choice based sampling this extreme has been studied by Owen (2007).
2 Methods

For a binary random variable $Y = [y_1 \ y_2 \ \ldots \ y_T]'$ denoting the occurrence $y_t = 1$ or non-occurrence $y_t = 0$ for sample member $t = 1, 2, \ldots, T$ of an event influenced by some exogenous variables $x_t = [x_{1,t} \ x_{2,t} \ \ldots \ x_{K,t}]$ and thereby $X = [x'_1 \ x'_2 \ \ldots \ x'_T]'$ the logistic regression model

$$\pi_t = \Pr(y_t = 1|x_t) = (1 + \exp(-x_t^\beta))^{-1} \quad (1)$$

with $\beta = [\beta_1 \ \beta_2 \ \ldots \ \beta_K]'$ can be used to estimate and test for the effects $\beta$. Under random sampling from the population at risk – that is, every unit $t$ that has a chance of developing occurrence – maximum likelihood methods allow for consistent and asymptotically normal estimation of $\beta$ with the log-likelihood

$$\log L(\beta|Y, X) = -\sum_{t=1}^{T} \log(1 + \exp((1 - 2y_t)x_t^\beta)) \quad (2)$$

yielding the estimator $\hat{\beta}$. It can be shown (see Prentice and Pyke (1979: 404-407), McCullagh and Nelder (1989: 111-114)) that maximizing the likelihood

$$L(\beta|X, Y) = \prod_{t=1}^{T} \Pr(x_t|y_t) \quad (3)$$

of retrospective (choice based) sampling yields the same estimator $\hat{\beta}$, except for the intercept $\beta_0$. The intercept estimated from this likelihood is consistent for

$$\beta_0 + \log \left[ \left( \frac{\bar{y}}{1 - \bar{y}} \right) \left( \frac{1 - E[y_t]}{E[y_t]} \right) \right] \quad (4)$$
and therefore under knowledge of \( \mathcal{E}[y_t] \) can and should be easily corrected for (see King and Zeng (2001b: 144) and King and Zeng (2001b: section 6.2)).

Using the corrected version of \( \hat{\beta} \) for estimating probabilities for some \( x_f \) via 
\[
(1 + \exp(-x_f \hat{\beta}))^{-1}
\]
gives consistent but biased estimates due to two problems pointed out by King and Zeng (2001b: 145-150). First there is a bias in \( \hat{\beta} \) which can be estimated by the following bias estimation given by King and Zeng (2001b) and based upon McCullagh and Nelder (1989: 119-120,455-456):

\[
\text{bias}(\hat{\beta}) = (X'WX)^{-1}X'W\xi
\]  
\[ (5) \]

with \( \xi = 0.5\text{tr}(Q)[(1 + w_1)\hat{\pi}_t - w_1] \), \( \text{tr} \) being the trace operator, \( w_t \) being \( w_1 = \mathcal{E}(y_t)/\bar{y} \) for cases, \( w_0 = (1-\mathcal{E}(y_t))/(1-\bar{y}) \) for non-cases and \( \hat{\pi}_t \) being the estimated probabilities of occurrence for unit \( t \) from \( \hat{\beta} \). \( Q = X(X'WX)^{-1}X' \) and \( W \) is the diagonal matrix constructed from the \( \hat{\pi}_t(1-\hat{\pi}_t)w_t \). Applying this correction also reduces variance for the bias corrected estimator \( \tilde{\beta} = \hat{\beta} - \text{bias}(\hat{\beta}) \) (see King and Zeng, 2001b: 147,161).

Second, when probabilities are then estimated from \( \tilde{\beta} \) via

\[
\tilde{\pi}_f = \text{Pr}(y_f = 1|x_f, \tilde{\beta}) = (1 + \exp(x_f \tilde{\beta}))^{-1}
\]  
\[ (6) \]

one has to consider that changes in \( \tilde{\beta} \) usually do not affect \( \tilde{\pi}_f \) symmetrically and hence do not cancel out. The probability calculation can be corrected for this by considering the distribution \( f_{\tilde{\beta}} \) of \( \tilde{\beta} \):

\[
\text{Pr}(y_f = 1|x_f) = \int_{D(\tilde{\beta})} \text{Pr}(y_f = 1|x_f, \tilde{\beta}) f_{\tilde{\beta}}(\tilde{\beta})d\tilde{\beta}
\]  
\[ (7) \]
which can be estimated by using an estimation of the distribution $f_{\tilde{\beta}}$ and can furthermore be approximated (see King and Zeng, 2001b: 149,161-162) by

$$\Pr(Y_f = 1 | x_f) \approx \hat{\pi}_f + C_f$$  \hspace{2cm} (8)

$$C_f = (0.5 - \hat{\pi}_f) \tilde{\pi}_f (1 - \tilde{\pi}_f) x_0 V(\tilde{\beta}) x_0'$$  \hspace{2cm} (9)

where $x_0$ are the exogenous’ values for some arbitrarily chosen comparison group and $V(\cdot)$ being the covariance matrix. Using the estimated distribution of $\tilde{\beta}$ (8) becomes a Bayesian estimator (see King and Zeng, 2001b: 149).

Up to (5) these methods are implemented in Imai et al. (2012). The correction in (9) can be easily done by hand e.g. by using the R (R Development Core Team, 2011) function `fitted.values()`.

3 Simulation

3.1 Software

For simulation I wrote an R-package named reccsim (Westphal, 2012) for rare events case-control study simulation. The package’s main functionalities are

1. building a PopulationAtRisk object. This object describes how the cases come to happen under a specific hypothesis.

2. creating a pseudo-random case-control study data.frame from that PopulationAtRisk which then may be used for model estimation for example by Imai et al. (2007).
3.2 Parameters

Assume a probability of occurrence of an event of one in ten millions. Also consider two assumed factors $A$ and $B$ influencing the individual probability $\Pr(y_i = 1|A, B)$. The questions coming up are: (i) (8) is not proven to be uniformly superior over the other estimators reported above. How do the corrections behave for extremely rare events and for different quantities of interest (QI)? (QIs being the parameters $\beta$ themselves, the probabilities of occurrence for different exogenous values and the relative risks between different exogenous values as well as the corresponding confidence interval coverage.) (ii) What relative risk – given a population size – is needed to identify influencing factors? (iii) What happens when the model is not correctly specified? (iv) How does an increase in the size of the control group relative to the case group affect the results? For this I will use the term controls-to-case ratio (as in Hennessy et al., 1999), abbreviated by $CTC$. For studying these problems I give an example population in table 1. Assumed factors of influence are binary variables $A$ and $B$, population size is unspecified, there is slight association between $A$ and $B$. Population size will usually be given for a selected problem. I will search needed relative risks to identify factors’ influence for population sizes of one hundred millions, two hundred millions, five hundred millions and one billion, with numbers of $10, 20, 50, 100$ expected cases from the binomial experiment. For multiple hypothesis testing $\alpha$ is set to $0.1$ and a power of $0.98$ of the test is considered sufficient.

I will evaluate the correctly specified model – given here in R’s formula notation – $Y \sim A + B$ as well as the underspecified model $Y \sim A$ and forgo interaction
effects for the benefit of future research and due to the argued "necessity of an explicit theory" in Berry et al. (2010: 261-262).

Table 1: Distribution of the population for simulation with assumed factors A and B.

<table>
<thead>
<tr>
<th>A/B</th>
<th>0</th>
<th>1</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.50</td>
<td>0.20</td>
<td>0.70</td>
</tr>
<tr>
<td>1</td>
<td>0.20</td>
<td>0.10</td>
<td>0.30</td>
</tr>
<tr>
<td>∑</td>
<td>0.70</td>
<td>0.30</td>
<td>1.00</td>
</tr>
</tbody>
</table>

The groups therefore are as follows: "0": the group having neither "A" nor "B"; "A": the group exhibiting only attribute "A"; "B": the groups exhibiting only attribute "B"; "A:B": the group exhibiting both attributes.

I varied the relative risks $r r_i$ as follows. $\pi_A/\pi_0 = r r_A$ from 1 to 10 by 0.2. $\pi_B/\pi_0 = r r_B$ was $\in\{1, 2, 5, 10\}$ for each value of $\pi_A/\pi_0$. For reasonably small probabilities the odds ratio approximates the relative risk: we can compute

$$r r_{AB} = \pi_{AB}/\pi_0 \approx OR_{AB} = \exp{\beta_A + \beta_B} = OR_A \cdot OR_B \approx \pi_A/\pi_0 \cdot \pi_B/\pi_0,$$

with $\pi_i$ being the probability of occurrence in group $i$, when there is no interaction. $p_i$ is group "i"'s proportion of the population (notation as in King and Zeng, 2002). There was a restriction of

$$10^{-7} = \pi = p_0 \pi_0 + p_A \pi_A + p_B \pi_B + p_{AB} \pi_{AB}$$

10^{-7} = \pi = p_0 \pi_0 + p_A \pi_A + p_B \pi_B + p_{AB} \pi_{AB} (11)

to account for the aforementioned occurrence of one in ten millions. For each set of parameters the model estimation was repeated 10000 times with a random case-control study generated every time. To ensure the existence of the maximum
likelihood estimator (see Silvapulle, 1981), generated case-control studies with empty groups among either the cases or the controls were rejected. Therefore my results are estimations of theoretical properties of the estimators conditional on the non-existence of empty groups. This restriction can easily be satisfied in applications by restricting analysis to situations where cases are observed from all groups and increasing the $CTC$ if necessary.

4 Results

Unless otherwise specified, the presented results are from the simulations with a population of one billion and $CTC = 5$ with the corresponding values for 500 millions and 200 millions in parentheses. For a population of 100 millions the effects could not be found with a high enough power. The power of the test for $\beta_A$ maxes out at 0.86 for a population of 100 millions in the case of underspecification and 0.79 for correct model specification. This is in accordance with the results of Peduzzi et al. (1996), as there are not enough events per variable. Also note how Westphal (2012) could easily be applied to re-study Peduzzi et al.'s topic. My requirements for the power are much stricter than the powers reported in Vittinghoff and McCulloch (2007: 715) and therefore our results, when interpreted in terms of events per variable (see Vittinghoff and McCulloch, 2007), differ.

4.1 Correctly specified model

4.1.1 Point Estimates

King's theoretical results of $\tilde{\beta}$ having less bias and less variance shows in my results with $\tilde{\beta}_0$ having up to a 10% (18%, 24%) smaller RMSE than $\hat{\beta}_0$ and
\( \tilde{\beta}_A \) having up to a 7% (14%, 21%) smaller RMSE than \( \hat{\beta}_A \). The RMSE ratios depending on \( rr_A \) all look like the curves in figure 1. Source code for generating these graphs for many parameters of interest is supplied with this article. For \( rr_B = 10 \) King's estimator actually produces worse results (minimal MSE ratio of 0.68) than the simple prior correction estimator in the case of a population size of 200 million.

The average absolute difference in bias between both methods for all parameter sets is around eight times (four times, two times) as high as the average absolute difference in variance.

My findings differ for the estimation of probabilities. Using King's \( \tilde{\beta} \) increases the RMSE of \( \tilde{\pi}_0 \) up to 12% (33%, 100%) over simple prior correction. Using King and Zeng (2001b)'s bayesian method increases RMSE by 30% (80%, 350%). This increase of RMSE approaches zero for increasing \( \pi_0 \) and likely will fully vanish and even reverse for larger \( \pi_0 \) than I simulated. Evidence for the latter conjecture is found in King and Zeng (2001b: Figure 6) where an \( X \) of 2.3 approximately represents a relative risk of 10 between the "groups" \( X = 0 \) and \( X = 2.3 \). The figure clearly shows, much higher relative risks are needed to find the bayesian estimator to be superior. The same cannot be said about \( \tilde{\pi}_A \). While the RMSE of \( \tilde{\pi}_A \) itself seems to improve with increasing \( \pi_A \) it gets worse for the bayesian estimator. It therefore seems some caution is advisable when applying King's methods to extremely rare events when you are interested in the probability estimations for the groups.

When estimating relative risks, using \( \tilde{\beta} \) shows huge improvement in variance and bias over using \( \hat{\beta} \) (see figure 1 (a)-(d), population size: one billion, \( CTC = 5 \)). Improvement by using King and Zeng's bayesian correction in mean squared
error can be clearly visualized. However its magnitude seems to be negligible (maximum ratio observed: $2.8 \times 10^7$).

Another quantity of interest is the power of the test. Due to the lower bias and variance of King and Zeng’s estimator $\tilde{\beta}$ it is preferable to $\hat{\beta}$ in terms of the test’s power. The interesting section of the approximate power curve for the one billion population is shown in (e). Figure 1 (f) clearly shows King’s estimator $\tilde{\beta}$ being superior in specificity and sensitivity to $\hat{\beta}$.

4.1.2 Confidence Intervals

Confidence intervals for the quantities of interest (i) coefficients $\beta_j, j \in \{A, B\}$ (ii) probabilities $\pi_i, i \in \{A, B, AB\}$ and (iii) relative risks $rr_i$ can be simulated. Imai et al. (2012) provides the function sim() for conducting this simulation. Due to the number of simulations needed, I applied the method described by King et al. (2000: 349-350) and King and Zeng (2002: 1419) directly by using Genz et al. (2012) and the saved point estimates and estimated coefficients' covariance matrices from the output generated by Imai et al. (2012) for simulating one thousand draws from each of the $\beta$ estimators' posteriors. I set the nominal level of coverage at 90% for all simulations.

When we are interested in the relative risks, figure 2 (a) shows neither the logit estimator with prior correction nor King’s corrected estimator dominate another, when the model is specified correctly. When misspecified, King’s corrected estimator clearly beats the logit estimator with prior correction (figure 2 (b)). Each point in figure 2 represents one set of relative risks with $rr_B$ indicated by the point’s color. For the probability estimation confidence interval coverage for both estimators is far too low (in the region of 40%) for underspecification
Figure 1: $\hat{\beta}$ vs. $\tilde{\beta}$, population one billion, \textit{CTC} of five.
RESULTS

(a) CI coverages for $r_{AB}$

(b) CI coverages for $r_A$ (underspecification)

Figure 2: Comparison of confidence interval coverage for a nominal coverage of 0.9
and way too high (starting at 93% and going up to close to 100%) for correct specification.

4.2 Varying population size

Varying population size from 100 million to 200, 500 million and one billion does not change the direction of the results. The relative difference between RMSEs of relative risk estimation seemingly increases quadratically. King and Zeng's correction therefore is the more important the smaller the population / the rarer the events are. The population size of 100 million did not yield high enough powers. For all other population sizes table 2 shows some pivotal characteristics of the power of the test for $\beta$.

| Table 2: Power of testing for $\beta_A$ for different population sizes. |
|-----------------------------|-----------------------------|-----------------------------|
| QI                          | Population size in millions (Expected no. of cases) | 200  | 500  | 1000 |
| (Expected no. of cases)     |                              | (20) | (50) | (100) |
| (a) Max. $rr_A$ needed to reach a power of 0.98 ($rr_B = 10$) | NA | 3.8 | 2.4 |
| (b) Min. $rr_A$ needed to reach a power of 0.98 ($rr_B = 1$) | 8.2 | 3.4 | 2.4 |
| (c) $CAF_{A|I}$ is X out of Y ($X/Y$) expected cases given $rr_B = 1$ and QI (b) from this table | 14/20 | 22/50 | 30/100 |
| (d) $CAF_{A|I}$ is X out of Y ($X/Y$) expected cases given $rr_B = 10$ and QI (a) from this table | NA | 14/50 | 17/100 |

The quantities of interest (QI) (c) and (d) in table 2 and table 3 require some explanation: Given the different probabilities of occurrence in the (four) different groups there is a baseline probability of $\pi_0$ for units without exposure
to risk influencing factors. So if one could remove the probability increasing factors from the non-zero groups, these groups’ $\pi_i$ would switch to $\pi_0$. They would still generate cases, but with a lower probability. Therefore the difference in probabilities between $\pi_i$ and $\pi_0$ multiplied with the size of the subpopulation in group $i$ tells us how many additional cases group membership of group $i$ is responsible, from now on called cases attributable to group $i$: $CAG_i$. Attributing $CAG_i$ to a factor is easy when only one factor increases the relative risk of group $i$. Then $CAG_i$ are fully attributable to this factor and therefore can be written as cases attributable to factor $j$ conditional on the group $i$: $CAF_{ji}$. When $rr_i > 1, i \in \{A, B\}$, $rr_{AB}$ is computed as in (10). Therefore not all $CAG_{AB}$ may be attributed to a single factor. I split them between the factors using the weights of groups’ “A” and “B” relative risks’ logarithm in the logarithm of the relative risk of group “AB”:

$$CAG_{AB} = population \cdot p_{AB} \cdot (\pi_{AB} - \pi_0) \quad (12)$$

$$CAF_{j|iAB} = CAG_{AB} \cdot \frac{\log(rr_j)}{\log(rr_{AB})} \quad (13)$$

This measure fulfills the following properties for $a > 1, b \geq 1$: (i) For $a > b$ \ \ \ \ \ $\log(a)/\log(ab) > 0.5$. (ii) For $a < b$ \ \ \ \ \ $\log(a)/\log(ab) < 0.5$. (iii) For $a = b$ \ \ \ \ \ $\log(a)/\log(ab) = 0.5$. (iv) For $b = 1$ \ \ \ \ \ $\log(a)/\log(ab) = 1$. Where $a$ and $b$ may be substituted by $rr_A$ and $rr_B$.

The $CAG$ can be computed by using the function add.cases() in Westphal (2012).
4.3 Increasing the controls-to-case ratio

The original $CTC$ was set at five times as many controls as cases (in accordance with Hennessy et al., 1999) in each case-control study. As King and Zeng (2001b: 141) clearly note, for rare events most information lies in the cases and not in the controls. In my setting initially there are no controls in the data and I use the number of cases to determine the number of controls. Obviously when there are very few cases compared to the population size, this method generates very few controls compared to the population size. King and Zeng (2001b: 153-157) undertake their analysis by dropping a percentage of controls from the data; I add some controls to the data. This is the same, but we start at different ends: King and Zeng start at 100% of controls I start at 0% of controls. As table 3 shows, I give results for a complete different range of “zeroes dropped” compared to King and Zeng (2001b) who drop at most 90% of the non-cases.

As is to be expected adding more controls necessarily reduces variance. The effects are also shown in table 3. Unfortunately increasing the number of controls is costly in two ways: Obviously when conducting the research costs increase by having to collect a large control sample. Not so obviously, the cost for learning about the estimators’ behaviour increases as simulations take longer. The simulations I conducted for a single population size took about two days for a CTC of 5, about as long for a CTC of 10, twice as long for a CTC of 50 and would have taken around 60 days for a CSF of 500 on a state of the art personal computer without any parallelization.
Table 3: Effects of increasing the CTC for different quantities of interest (QI), (e) and (f) explained in table 4.

<table>
<thead>
<tr>
<th>Pop. Size</th>
<th>QI</th>
<th>CTC (expected % of non-cases dropped)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(99.99994)</td>
</tr>
<tr>
<td>200 mio.</td>
<td>(a)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>(b)</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>(c)</td>
<td>14/20</td>
</tr>
<tr>
<td></td>
<td>(d)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>(e)</td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td>(f)</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>Max. MSE $\hat{r}<em>{r_A}$ ($r</em>{r_A},r_{r_B}$)</td>
<td>20.5</td>
</tr>
<tr>
<td></td>
<td>Max. MSE $\hat{r}<em>{r</em>{AB}}$ ($r_{r_A},r_{r_B}$)</td>
<td>(10,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(10,10)</td>
</tr>
<tr>
<td>500 mio.</td>
<td>(a)</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>(b)</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>(c)</td>
<td>22/50</td>
</tr>
<tr>
<td></td>
<td>(d)</td>
<td>14/50</td>
</tr>
<tr>
<td></td>
<td>(e)</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>(f)</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>Max. MSE $\hat{r}<em>{r_A}$ ($r</em>{r_A},r_{r_B}$)</td>
<td>21.3</td>
</tr>
<tr>
<td></td>
<td>Max. MSE $\hat{r}<em>{r</em>{AB}}$ ($r_{r_A},r_{r_B}$)</td>
<td>(10,2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4804</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(9.6,10)</td>
</tr>
<tr>
<td>1 bio.</td>
<td>(a)</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>(b)</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>(c)</td>
<td>30/100</td>
</tr>
<tr>
<td></td>
<td>(d)</td>
<td>17/100</td>
</tr>
<tr>
<td></td>
<td>(e)</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>(f)</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>Max. MSE $\hat{r}<em>{r_A}$ ($r</em>{r_A},r_{r_B}$)</td>
<td>11.5</td>
</tr>
<tr>
<td></td>
<td>Max. MSE $\hat{r}<em>{r</em>{AB}}$ ($r_{r_A},r_{r_B}$)</td>
<td>(10,10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2740</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(10,10)</td>
</tr>
</tbody>
</table>
4.4 Increasing the probability

My simulations did not vary overall probability of occurrence. However it is easy to see, that given constant relative risks, increasing overall probability of occurrence has to increase probability of occurrence for all groups: From King and Zeng (2001b: equation (6)) we know that variance for $\hat{\beta}$ decreases with increasing $\pi$:

$$V(\hat{\beta}) = \left[ \sum_{t=1}^{T} \pi_t (1 - \pi_t) x_t' x_t \right]^{-1}$$

$$\frac{\partial (\pi_t - \pi_t^2)}{\partial \pi_t} = 1 - 2 \pi_t > 0 \forall \pi_t \in (0, 0.5)$$  \hspace{1cm} (14)

and thereby decreasing its inverse.

Therefore under the assumption of $\hat{\beta}$'s bias not increasing with $\pi$ – for example when doubling $\pi$ and cutting population size in half – a lower relative risk will be needed for finding the influence of a factor when the number of (expected) cases remains the same. The aforementioned assumption can be justified by Peduzzi et al. (1996: Figure 2) in combination with King and Zeng (2001b: Figure 4).

4.5 Underspecified model

Between $A$ and $B$ there is a phi coefficient of $\phi = \frac{1}{21}$, i.e. very low association. Despite how low association seems to be, its effect when underspecifying the model as $Y \sim A$ is noticeable when looking at the power of the test in table 4. Group "A"'s effect is now found sooner for high relative risks in group "B". Of course, while the test result is correct in a binary choice fashion, the improved power is not due to the test somehow becoming more sensitive but
due to falsely loading explanatory power from “B” onto “A” (see Lee (1982: 207, Proposition 2), also Yatchew and Griliches (1985)). Finding “A”’s effect sooner for a non-influential “B” is due to the reduction in variance due to – in this case correct – model building.

<table>
<thead>
<tr>
<th>Q1</th>
<th>Population size in millions</th>
<th>200</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>(e)</td>
<td>Max. $r_{rA}$ needed to reach a power of 0.98 ($r_{rB} = 1$)</td>
<td>7.6</td>
<td>3.4</td>
<td>2.4</td>
</tr>
<tr>
<td>(f)</td>
<td>Min. $r_{rA}$ needed to reach a power of 0.98 ($r_{rB} = 10$)</td>
<td>6.4</td>
<td>3.0</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Also King and Zeng’s coefficient bias correction now has most influence on the relative risk bias when “B”’s influence is lowest instead of highest. Apart from that, results do neither change in direction nor much in effect size.

5 Conclusion

We see that even for extremely rare events – with the rarity assumption fulfilled and fixed exposed proportions as is required (see Greenland and Thomas, 1982: 552)) – with binary exogenous variables, the logistic regression model is well worth being studied for (a) finding effects and (b) estimating relative risks when a serious effect from some factor is conjectured. Also note that for binary exogenous variables no belief in the logistic form has to be held. It is an elaborate test for proportions.

We learnt under what exogenous parameter settings confirmatory data analysis can be used for extremely rare events for evaluating hypotheses derived from case studies by qualitative research. King and Zeng’s methods are very helpful and have to be applied selectively for certain settings of rare and rarest events.
and depending on the researcher's quantities of interest. Its reduction in mean squared error for the relative risk estimation over the logistic regression maximum likelihood estimator is remarkable when used for extremely rare events – even for a population size of one billion. When estimating relative risks or when searching for significance there is no reason not to apply this correction (implemented in Imai et al., 2012) when dealing with case-control studies. Although its power does not improve dramatically, it must be uniformly more powerful due to the decreased bias and variance.

A few general rules of thumbs from my work and King and Zeng (2001b,a) can be found:

1. Effects can be found even for extremely rare events.
2. The stronger a factor's influence on the probability of occurrence, the more important become (5), (7).
3. For different quantities of interest under different parameters different methods have to be applied.
4. The more clouded one factor's influence is by another factor's influence, the more important become (5), (7).

As a further contribution, Westphal (2012) may be used to easily compare the methods from section 2 for plausible parameter sets, given a real world research problem.

6 Outlook

While throughout my simulation I assumed \( \pi \) known to be one in ten millions, in applications \( \pi \) will usually not be known. There is a simple remedy to unknown overall probabilities given in King and Zeng (2004: 5-6) and a more detailed
discussion in King and Zeng (2002).

Westphal (2012) may very well be used to also study interaction terms in case-control studies. Any researcher undertaking this endeavor should be aware of how interaction effects must be interpreted and computed. This is summed up comprehensively by Ai and Norton (2003) and Norton et al. (2004).

Furthermore for applications where confounders are present, it might be worthwhile to conduct a similar study with an additional step of propensity scoring as evaluated in Cepeda et al. (2003).
REFERENCES

References


REFERENCES


