Joint Discussion Paper Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 35-2013

Christian Westphal

Logistic Regression for Extremely Rare Events: The Case of School Shootings

This paper can be downloaded from
http://www.uni-marburg.de/fb02/makro/forschung/magkspapers/index_html%28magks%29

Coordination: Bernd Hayo • Philipps-University Marburg
Faculty of Business Administration and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

Logistic Regression for Extremely Rare Events: The Case of School Shootings

Christian Westphal[a]

This version: July 24, 2013

Abstract

School shootings are often used in public policy debate as a justification for increased regulation, based on qualitative arguments. However, to date, no effort has been made to find valid quantitative evidence for the claims bolstering the regulation recommendations. In defense of this absence of evidence, it is usually argued that the rarity of such events does not allow the employment of quantitative methods. This paper, using a simulation study, shows that, based on the number of shool shootings in the United States and Germany combined, the well-known method of logistic regression can be applied to a case-control study, making it possible to at least test for an association between hypothesized influential variables and the occurrences. Moderate relative risks, explained by an observed variable, would lead to a high power of the appropriate test. A moderate numbers of cases generated by such a variable would suffice to show a significant association.

JEL Classifications: C25; C35; I18; K14
MSC Classifications: 62J12
Keywords: Rare Events; Logistic Regression; Case-Control Studies; School Shootings

[a]University of Marburg, Faculty of Business Administration and Economics, Department of Statistics, christian.westphal@westphal.de, westphal@staff.uni-marburg.de

1 Introduction

The qualitative scientific literature from multiple fields contains a great many claims about what causally leads to the occurrence of school shootings, or, what is at least associated with the occurrence of such tragic events. Some of these claims are employed in public policy debate as a justification for increased regulatory action, and thereby have the potential to influence social welfare, even though these claims, while they may seem "obvious", are not backed up by quantitative evidence. A partial and compact overview of these claims is found in Kleck (1999: 2) and is quoted here in its entirety to illustrate the diversity of claims made:

> guns, "assault weapons", large-capacity ammunition magazines, lax regulation of gun shows; the failure of parents to secure guns, school cliques, and the exclusion of "outsiders"; bullying and taunting in schools, especially by high school athletes; inadequate school security, especially a lack of metal detectors, armed guards, locker searches, and so forth; excessively large high schools; inadequate monitoring of potentially violent students by schools; lazy, uninvolved Baby Boomer parents and correspondingly inadequate supervision of their children; young killers not being eligible for death penalty; a lack of religion, especially in schools; violent movies and television; violent video games; violent material and communications on the World Wide Web/Internet (including bomb-making instructions); anti-Semitism, neo-Nazi sentiments, and Hitler worship; "Industrial" music, Marilyn Manson's music, and other "dark" variants of rock music; Satanism; "Goth" culture among adolescents; and Southern culture.

All of these claims can be modeled as binary variables and the outcome, of course, is binary as well: a school shooting either happens or does not. For the quantitative analyst, it seems obvious to search for a significant association between the events and the hypothesized influencing variables. A theoretical model lending itself to this purpose is given in Robertz (2004) (an excellent book that is, unfortunately, not available in an English translation), where "fantasy" is considered a latent variable, influenced by exogenous variables, and, when pushed too hard, possibly leads to extremely deviant behavior, i.e., a school shooting. Then the "choice" of committing a school shooting depends on the influencing variables; hence we are dealing with a *choice model*, which can be modeled and estimated as a logistic model (see Manski and Lerman, 1977). In epidemiology, these models are called *incidence models* (see Prentice and Pyke, 1979). As King and Zeng (2001*b*) point out, when occurrence (or nonoccurrence) is rare, collecting a random sample with even one occurrence may become prohibitively expensive, which is clearly the case with school shootings as, fortunately, only very

few students choose to kill their peers and teachers. Prentice and Pyke (1979) and Manski and Lerman (1977) show that collecting the occurrences and adding a random sample of nonoccurrences (or vice versa, depending on what is labeled as an occurrence) allows for consistent estimation of the logistic regression parameters from such a *case-control study*. A very good summary of these statistical methods in conjunction with case-control studies can be found in Breslow (1996).

For the problem at hand let us take as our population the "enrolled student years." I define an "enrolled student year" as each year an individual student is enrolled in school. I refrain from specifically stating what schools and which grades should be included; these choices will need to be made at the time of application. With this definition, we can easily measure the number of "enrolled student years that did not lead to a school shooting" and those that did. Following Robertz' (2004) definition of what constitutes a "school shooting," there were 72 cases from 1992 to 2009, committed by male students from 10 to 34 years old in the United States and Germany combined (Robertz and Wickenhäuser, 2010: 14). In the same time frame, there were around 500 million years of education provided to male students and around 1 billion years of education for both sexes, revealing the rarity, indeed, the extreme rarity, of school shootings. The goal of the case-control study is to find a statistically significant association, and better yet, causality, between the occurrence of school shootings and above-mentioned variables.

The method of case-control studies is examined by King and Zeng (2001*b*) (see also an intuitive explanation and application in King and Zeng, 2001*a*) for the case of *rare events*, which King and Zeng define as "dozens to thousands of times fewer ones . . . than zeroes" (King and Zeng, 2001*b*: 138). From the numbers above, I am interested in how these methods perform in finite samples *when the occurrence is millions to tens of millions times more rare* than nonoccurrence; 72 in 500 million would be 1.44 occurrences in 10 million and 72 in 1 billion would be 0.72 occurrences in 10 million.

A viable way to draw a valid inference would be to construct a data set of all cases and controls, with the controls either randomly drawn from the population or artificial controls generated from known population parameters. The next step would be to group all hypothesized variables into two (or more) binary factors, assuming that none of the variables are negatively correlated and that none exhibit coefficients of opposed directions.[1] Next, check whether these factors have a statistically significant association with the outcome. Depending on how factors are constructed ("and" and "or" junctions come to mind), conclusions may be drawn from the test result, factor groups may be ruled out, and a stepwise search for individual variables may be constructed. Given this obvious arbitrary

[1] An assumption that is not contradicted anywhere in the qualitative literature.

interchangeability between individual variables and factors, the terms "variables" and "factors" are used interchangeably below.

This paper contributes to the literature by pointing out an easy-to-use quantitative method for measuring the association of (binary) factors with the occurrence of school shootings (in Section 2) and by examining via a simulation study what sort of relative risk a certain factor, for example, constructed as described above, would have to impose on individuals in order to show positive association in a logistic regression model (in Sections 3 and 4). My core findings are presented in Sections 4.3 and 4.4. A software package designed to repeat the simulation procedure for specific settings is provided, and its use is illustrated for an example setting. The main result shows that for plausible population sizes and overall probability of occurrence, only very few cases would need to be generated by an exogenous factor to find a significant association with the occurrences. Unfortunately, there is no data set, at least to my knowledge, that measures the above-mentioned variables for every school shooting that has ever occurred. Thus, putting the hypotheses to a meaningful test will require retrospectively collecting the data necessary for the cases and the control populations.

2 Methods

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

$$\pi_t = \Pr(y_t = 1|\mathbf{x}_t) = (1 + \exp\{-\mathbf{x}_t\boldsymbol{\beta}\})^{-1} \tag{1}$$

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

$$\log L(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}) = -\sum_{t=1}^{T} \log\big(1 + \exp\{(1 - 2y_t)\mathbf{x}_t\boldsymbol{\beta}\}\big) \tag{2}$$

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

$$L(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) = \prod_{t=1}^{T} \Pr(\mathbf{x}_t|y_t) \tag{3}$$

of retrospective (choice-based) sampling yields the same estimator $\hat{\boldsymbol{\beta}}$, except for the intercept β_0. The intercept estimated from this likelihood is consistent for

$$\beta_0 + \log\Big[\Big(\frac{\bar{y}}{1-\bar{y}}\Big)\Big(\frac{1-\mathscr{E}[y_t]}{\mathscr{E}[y_t]}\Big)\Big] \tag{4}$$

and therefore, with knowledge of $\mathscr{E}[y_t]$, can and should be easily corrected for (see King and Zeng, 2001*b*: 144 and Section 6.2).

Using the corrected version of $\hat{\boldsymbol{\beta}}$ for estimating probabilities for some $\mathbf{x}_f$ via $(1+\exp\{-\mathbf{x}_f\hat{\boldsymbol{\beta}}\})^{-1}$ results in consistent but biased estimates due to two problems pointed out by King and Zeng (2001*b*: 145–150). First, there is a bias in $\hat{\boldsymbol{\beta}}$, which can be estimated using the following bias estimation from King and Zeng (2001*b*), which is based on McCullagh and Nelder (1989: 119–120,455–456):

$$\widehat{\text{bias}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\xi} \tag{5}$$

with $\xi = 0.5\text{tr}(Q)\big[(1+w_1)\hat{\pi}_t - w_1\big]$, tr being the trace operator, w_t being $w_1 = \mathscr{E}(y_t)/\bar{y}$ for cases, $w_0 = (1-\mathscr{E}(y_t))/(1-\bar{y})$ for non-case,s and $\hat{\pi}_t$ being the estimated probabilities of occurrence for unit t from $\hat{\boldsymbol{\beta}}$. $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{W}$ 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 $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \widehat{\text{bias}}(\hat{\boldsymbol{\beta}})$ (see King and Zeng, 2001*b*: 147,161).

Second, when probabilities are then estimated from $\tilde{\boldsymbol{\beta}}$ via

$$\tilde{\pi}_f = \Pr(y_f = 1|\mathbf{x}_f, \tilde{\boldsymbol{\beta}}) = (1+\exp\{\mathbf{x}_f\tilde{\boldsymbol{\beta}}\})^{-1} \tag{6}$$

it must be kept in mind that changes in $\tilde{\boldsymbol{\beta}}$ usually do not affect $\tilde{\pi}_f$ symmetrically and hence do not cancel out. The probability calculation can be corrected for this problem by considering the distribution $f_{\tilde{\boldsymbol{\beta}}}$ of $\tilde{\boldsymbol{\beta}}$:

$$\Pr(y_f = 1|\mathbf{x}_f) = \int_{\mathbb{D}(\boldsymbol{\beta})} \Pr(y_f = 1|\mathbf{x}_f, \tilde{\boldsymbol{\beta}}) f_{\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}}) \mathrm{d}\tilde{\boldsymbol{\beta}} \tag{7}$$

which can be estimated by using an estimation of the distribution $f_{\tilde{\boldsymbol{\beta}}}$ and can furthermore be approximated (see King and Zeng, 2001*b*: 149,161–162) by

$$\Pr(Y_f = 1|\mathbf{x}_f) \approx \tilde{\pi}_f + C_f \tag{8}$$

$$C_f = (0.5 - \tilde{\pi}_f)\tilde{\pi}_f(1-\tilde{\pi}_f)x_0\mathscr{V}(\tilde{\boldsymbol{\beta}})\mathbf{x}_0' \tag{9}$$

where $\mathbf{x}_0$ are the exogenous values for some arbitrarily chosen comparison group

and $\mathcal{V}(\cdot)$ is the covariance matrix. Using the estimated distribution of $\tilde{\boldsymbol{\beta}}$, Equation (8) becomes a Bayesian estimator (see King and Zeng, 2001*b*: 149).

Equations (4) and (5) are implemented in Imai, King and Lau (2012). The correction in Equatin (9) is easily made by using, for example, the `R` function `fitted.values()`.

3 Simulation

3.1 Software

For the simulation, I wrote an `R`-package named `reccsim` (Westphal, 2012), standing 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 and given a set of parameters, describing how factors/variables are distributed among the population.
2. Creating a pseudo-random case-control study `data.frame` from that `PopulationAtRisk` that then may be used for model estimation, for example with Imai, King and Lau (2012).

3.2 Parameters for Simulation

Assume an event's probability of occurrence to be 1 in 10 million, which is somewhere between the observed frequency of school shootings committed by "male enrolled student years" and "all student years," as set out in Section 1. Also consider two assumed factors, for example, an individual's access to "Guns" and an individual's consumption of violent computer "Games," influencing the individual probability $\Pr(y_t = 1|Guns, Games)$; note that for my analysis, it does not matter what two factors are assumed and, indeed, if one wishes to be as abstract as possible, a simple A and B will suffice. The issues that arise from these assumptions involve, first, that Equation (8) is not proven to be uniformly superior over the other estimators reported above. How do the bias corrections behave for extremely rare events and for different quantities of interest (QIs) discussed below? More importantly, what relative risk – given a population size and overall probability of occurrence – is needed to identify influencing factors? What happens when the model is not correctly specified? How does increasing the size of the control group relative to the case group affect the results? As shorthand for this last question, I will use the term *controls-to-case ratio* (as in Hennessy et al., 1999), abbreviated by CTC. To aid in answering these questions, I give an example distribution of the variables among the population in Table 1. The assumed factors of influence are two binary variables "Guns" and "Games." There is slight association between "Guns" and "Games." I will search for relative risks

necessary to identify these variables' (factors') influence for population sizes of 100 million, 200 million, 500 million, and 1 billion, the latter two figures approximating the real-world setting (see the Introduction). Based on these populations, $10, 20, 50$, and 100 cases, respectively are expected from the aggregated binomial experiment. For multiple hypotheses testing, the type-I-error is set to 0.1 and a power of 0.98 for the test is considered sufficient. Note that the test's power requirement is specified very conservatively to protect my results from weak claims about necessary conditions for the method to function as intended. The marginal frequency for *Guns* in Table 1 is a very rough computation based on household gun ownership density in the United States and Germany under the assumption of independence between household gun ownership and school children. The marginal frequency for *Games* is simply a guess based on personal experience, and conveniently symmetrical to the marginal frequencies of gun availability. The joint frequency between both variables is, frankly, an arbitrary choice.

I will evaluate the correctly specified model – given here in R's `formula` notation – $Shooting \sim Guns + Games$, as well as the underspecified model $Shooting \sim Guns$, but leave the discussion of interaction effects to future research, seeing as the arguably necessary "explicit theory" in Berry, Meritt and Esarey (2010: 261-262) is yet to be posited.

Guns/*Games*	0	1	$\sum$
0	0.50	0.20	0.70
1	0.20	0.10	0.30
$\sum$	0.70	0.30	1.00

Table 1: Distribution of the population for simulation with assumed factors *Guns* and *Games*

Thus, the groups are as follows: "0" – the group having neither guns nor playing games; "Guns" – the group only having guns; "Games" – the group only playing violent games; and "Guns:Games" – the group having guns and playing violent video games.

I varied the relative risks rr_i as follows. $\pi_{Guns}/\pi_0 = rr_{Guns}$ from 1 to 10 in increments of 0.2. $\pi_{Games}/\pi_0 = rr_{Games}$ was $\in \{1, 2, 5, 10\}$ for each value of π_{Games}/π_0. Because for reasonably small probabilities, the odds ratio approximates the relative risk, we can compute

$$\begin{aligned} rr_{Guns,Games} &= \pi_{Guns,Games}/\pi_0 \approx OR_{Guns,Games} \\ &= \exp\{\beta_{Guns} + \beta_{Games}\} = OR_{Guns} \cdot OR_{Games} \approx \pi_{Guns}/\pi_0 \cdot \pi_{Games}/\pi_0, \end{aligned} \quad (10)$$

with π_i being the probability of occurrence in group i, when there is no interac-

tion. p_i is group i's proportion of the population (notation as in King and Zeng, 2002). There was a restriction of

$$10^{-7} = \pi = p_0\pi_0 + p_{Guns}\pi_{Guns} + p_{Games}\pi_{Games} + p_{Guns,Games}\pi_{Guns,Games} \quad (11)$$

to account for the aforementioned occurrence of 1 in 10 million. For each set of parameters, the model estimation was repeated 10,000 times with a random case-control study generated each 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 nonexistence of empty groups*. This restriction can be easily satisfied in applications by restricting analysis to situations where cases are observed from all groups and increasing the CTC until there are controls from all groups, if necessary.

4 Results

In this section, I set out the simulation results. Unless otherwise noted, figures in the text refer to the population of 1 billion and a controls-to-cases ratio of $CTC = 5$. Results for different population sizes and different CTCs can be found in Tables 2, 3, and 4. Increasing the CTC does not change the results much. Varying the population size has a notable impact, as the number of cases generated varies. For a population of 100 million, the effects could not be found with a high enough power. The power of the test for β_{Guns} maxes out at 0.86 for a population of 100 million in the case of underspecification and at 0.79 for correct model specification. This is in accordance with the results of Peduzzi et al. (1996); there are simply not enough *events per variable*.[2] My requirements for the power are much stricter than the powers reported in Vittinghoff and McCulloch (2007: 715) and therefore my results, when interpreted in terms of *events per variable* (see Vittinghoff and McCulloch, 2007), differ, too.

4.1 Correctly Specified Model

4.1.1 Point Estimates

King and Zeng's theoretical results of $\tilde{\boldsymbol{\beta}}$ having less bias *and* less variance show in my results where $\tilde{\beta}_0$ has up to a 10% smaller RMSE[3] than $\hat{\beta}_0$ and $\tilde{\beta}_A$ has up to a 7% smaller RMSE[4] than $\hat{\beta}_A$. The RMSE ratios depending on rr_{Guns} are illustrated

[2] Also note how Westphal (2012) could easily be applied to re-study Peduzzi et al.'s topic.
[3] 18% and 24% for populations of 500 and 200 million, respectively.
[4] 14% and 21% for populations of 500 and 200 million, respectively.

in Figure 1. They look similar for different population sizes.

The *average absolute difference in bias* between both methods for all parameter sets is around eight times as high as the *average absolute difference in variance*.

My findings differ from King and Zeng when it comes to the estimation of probabilities. Using King's $\tilde{\boldsymbol{\beta}}$ increases the RMSE of $\tilde{\pi}_0$ up to 12% over simple prior correction.[5] Using King and Zeng's Bayesian method increases RMSE by 30%.[6] This increase of RMSE approaches zero for increasing $\pi_0$and likely will completely disappear or even reverse for larger π_0 than I simulated. Evidence for the latter conjecture is found in King and Zeng (2001*b*: Figure 6), where an X of 2.3 approximately represents a relative risk of 10 between the "groups" $X = 0$ and $X = 2.3$. That Figure clearly shows, that much higher relative risks are needed to find the Bayesian estimator superior. The same cannot be said about $\tilde{\pi}_{Guns}$. While the RMSE of $\tilde{\pi}_{Guns}$ itself seems to improve with increasing π_{Guns}, it becomes worse for the Bayesian estimator. It therefore appears that some caution is advisable when applying King's methods to extremely rare events in an effort to determine the probability estimations for the groups.

When estimating relative risks, using $\tilde{\boldsymbol{\beta}}$ shows huge improvement in variance and bias over using $\hat{\boldsymbol{\beta}}$ (see Figure 1 (a)–(d), population size: 1 billion, $CTC = 5$). Obvious improvement is achieved by using King and Zeng's Bayesian correction in mean squared error; however its magnitude seems to be negligible (maximum ratio observed: $\sqrt[2.5\cdot 10^7]{10}$).

Another quantity of interest is the power of the test. Due to its lower bias and variance, King and Zeng's estimator $\tilde{\boldsymbol{\beta}}$ is preferable to $\hat{\boldsymbol{\beta}}$ in terms of the test's power. The interesting section of the approximate power curve for the 1 billion population is shown in Figure 1 (e). Figure 1 (f) clearly shows that King and Zeng's estimator $\tilde{\boldsymbol{\beta}}$ is superior in specifity and sensitivity to $\hat{\boldsymbol{\beta}}$ in this setting.

4.1.2 Confidence Intervals

Confidence intervals for the quantities of interest (i) coefficients β_j where $j \in \{Guns, Games\}$, (ii) probabilities $\pi_i, i \in \{Guns, Games, (Guns, Games)\}$, and (iii) relative risks rr_i can be simulated. Imai, King and Lau (2012) provide the function `sim()` for conducting this simulation. Due to the number of simulations needed, I used the method described by King, Tomz and Wittenberg (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, King and Lau (2012) for simulating 1,000 draws from each of the $\boldsymbol{\beta}$ estimators' posteriors, mimicking `sim()`'s behavior. I set the nominal level of coverage at 90% for all simulations.

[5] 33% and 100% for populations of 500 and 200 million, respectively.

[6] 80% and 350% for populations of 500 and 200 million, respectively.

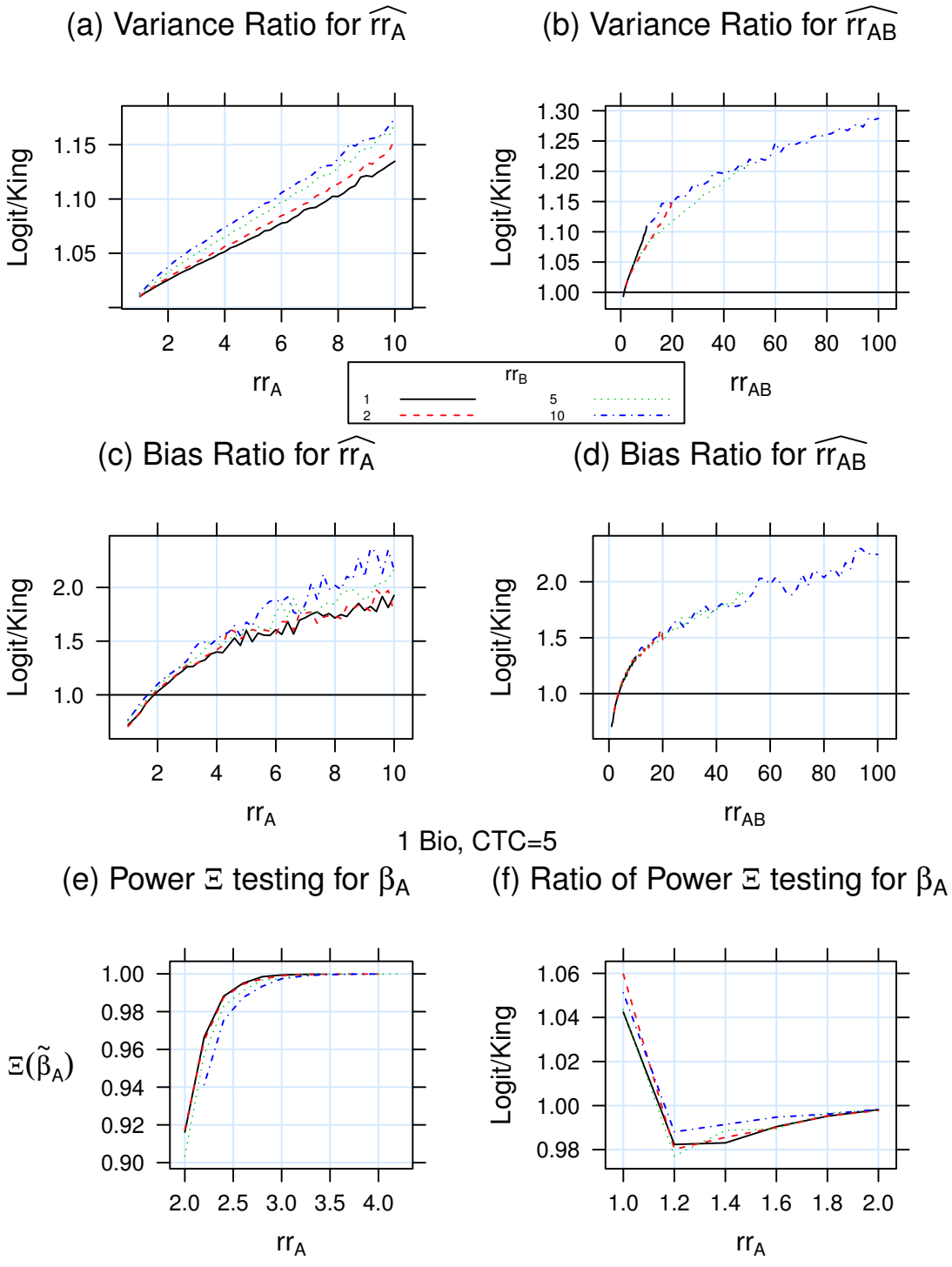

Figure 1: $\hat{\beta}$ vs. $\tilde{\beta}$, population 1 billion, $CTC = 5$

As to relative risks, Figure 2 (a) shows that neither the logit estimator with prior correction nor King's corrected estimator dominate when the model is specified correctly. When misspecified, however, King and Zeng'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 the underspecification and way too high (starting at 93% and reaching up to 100%) for the correct specification.

4.2 Varying Population Size

Varying the population size from 100 million to 200 million, 500 million, or 1 billion does not change the direction of the results. The relative difference between the RMSEs of relative risk estimation appear to increase quadratically. Therefore, King and Zeng's correction is the more important the smaller the population/the rarer the event. 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 $\tilde{\beta}$.

4.3 Quantities of Interest

For the specific application of school shootings and possible contributing factors, there are multiple *quantities of interest*, set out for populations of 200 million, 500 million, and 1 billion in Table 2, 3, and 4. Below, I briefly discuss these quantities of interest.

The Maximum rr_{Guns} Needed to Reach a Power of 0.98. Which was the largest rr_{Guns}, unconditional on rr_{Games}, that yielded at least a simulated power of 0.98? Additionally, in the appropriate table rows, the value of rr_{Games} under which this value was found is given. The meaning of this value is that to achieve a test power of 0.98, and given all rr_{Games} I simulated, rr_{Guns} of the figure given in the table, or larger, will lead to a rather powerful test. The hypothesis test I conducted is two sided. Hence, a possible criticism is that, possibly, my power (i.e., the rejection of the null hypothesis) is being erroneously bolstered by a percentage of significant *negative* coefficient estimates. However, for relative risks ≥ 2.4, 2 out of 1.56 million simulation results exhibit this characteristic. Therefore, this potential problem seems of little concern.[7]

The Minimum rr_{Guns} Needed to Reach a Power of 0.98. Which was the smallest rr_{Guns}, unconditional on rr_{Games}, that yielded at least a simulated power of 0.98? In the appropriate table rows, the value of rr_{Games} under which this value was found is given. The meaning of this value is, that to achieve a test

[7] The figures are of similar negligible size for cases other than a population size of 1 billion, $CTC = 5$ and the respective relative risks reported in Tables 2 and 3.

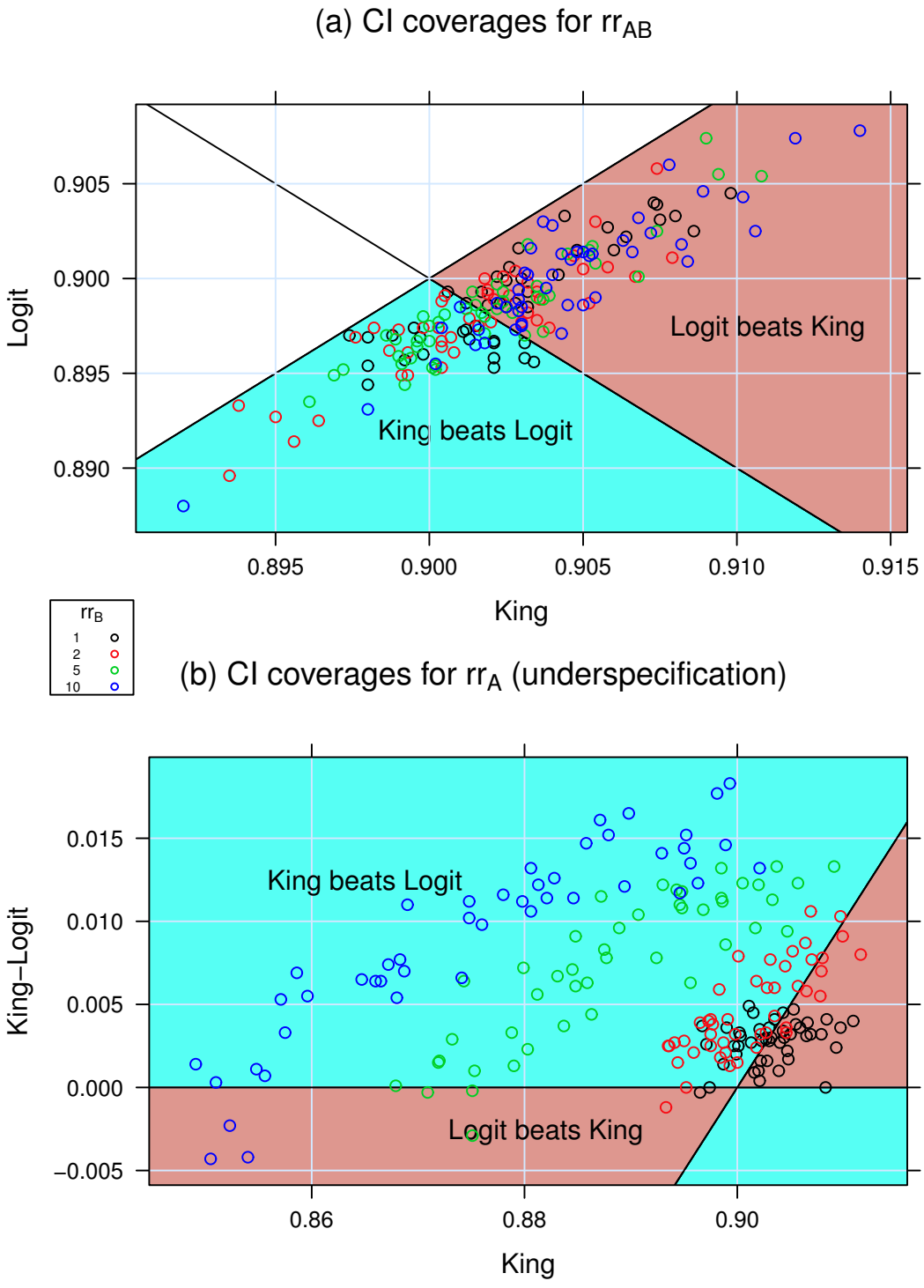

Figure 2: Comparison of confidence interval coverage for a nominal coverage of 0.9

power of 0.98, rr_{Guns} smaller than this value never resulted in a power of ≥ 0.98.

Cases Attributable to a Factor. The quantities of interest (QI) (c) and (d) in Tables 2 and 3 require some explanation: Given the different probabilities of occurrence in the (four) different groups there is a *baseline* probability of π_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' π_i would switch to π_0. The groups would still generate cases, but at a lower probability. Therefore, the difference in probabilities between π_i and π_0 multiplied by the size of the subpopulation in group i tells us how many additional cases group i is responsible for, 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. In that situation, CAG_i is fully attributable to this factor and therefore can be written as *cases attributable to factor* j conditional on the group i: $CAF_{j|i}$. When $rr_{k\neq j} > 1, k \in \{Guns, Games\}$, $rr_{j,k}$ is computed as in Equation (10). Therefore, not all $CAG_{Guns,Games}$ can be attributed to a single factor. I split them between the factors using the weights of groups "Guns" and "Games" relative risks logarithm in the logarithm of the relative risk of group "Guns,Games":

$$CAG_{Guns,Games} = population \cdot p_{Guns,Games} \cdot (\pi_{Guns,Games} - \pi_0) \tag{12}$$

$$CAF_{j|Guns,Games} = CAG_{Guns,Games} \cdot \frac{\log(rr_j)}{\log(rr_{Guns,Games})} \tag{13}$$

This measure meets the following requirements for $a > 1, b \geq 1$: (i) For $a > b \quad \log(a)/\log(ab) > 0.5$. (ii) For $a < b \quad \log(a)/\log(ab) < 0.5$. (iii) For $a = b \quad \log(a)/\log(ab) = 0.5$. (iv) For $b = 1 \quad \log(a)/\log(ab) = 1$. In each case, a and b may be substituted by rr_{Guns} and rr_{Games}.

It is interesting that in a case where a second factor imposes a high relative risk, *fewer* cases are attributable to the first variable under the minimum identification requirement. I conjecture that the explanation for this can be found in Equation (10): under the assumption of no interaction between the linear terms, an additional variable with a relative risk > 1 leads to a multiplicative effect for the relative risk and therefore has an multiplicative effect on the number of cases exhibiting the factor relative to the number of cases *not* exhibiting the factor. The *CAG* can be computed by using the function `add.cases()` in Westphal (2012).

4.4 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 (2001*b*: 141) state, 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

QI	Population size in million (Expected no. of cases)	200 (20)	500 (50)	1000 (100)
(a)	Max. rr_{Guns} needed to reach a power of 0.98 ($rr_{Games} = 10$)	NA	3.8	2.4
(b)	Min. rr_{Guns} needed to reach a power of 0.98 ($rr_{Games} = 1$)	8.2	3.4	2.4
(c)	$CAF_{Guns\|\cdot}$ is X out of Y (X/Y) expected cases given $rr_{Games} = 1$ and QI (b) from this table	14/20	22/50	30/100
(d)	$CAF_{Guns\|\cdot}$ is X out of Y (X/Y) expected cases given $rr_{Games} = 10$ and QI (a) from this table	NA	14/50	17/100

Table 2: Power of testing for β_A for different population sizes.

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 (2001*b*: 153–157) undertake their analysis by *dropping* a percentage of controls from the data; I *add* some controls to the data. Hence, I approach the problem from the opposite direction: that is, King and Zeng start with 100% of controls, I start with none. Table 3 sets out the results for a range of "zeroes dropped," which is different from King and Zeng (2001*b*), who drop, at most, 90% of the non-cases.

As to be expected, adding more controls necessarily reduces variance. These effects are also shown in Table 3. Unfortunately, increasing the number of controls is costly in two ways. Obviously, research costs increase due to having to collect a larger control sample. Not so obviously, the cost of learning about the estimators' behavior increases because 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 CTC of 500 on a state-of-the-art personal computer without any parallelization. Tables 2 and 3, QIs (a), (b), (e), and (f), show that the marginal returns measured in indentifying influential variables at lower relative risks depend on population size and the numbers of cases expected to be generated.

		CTC (expected % of non-cases dropped)		
		5	10	50
Pop. Size	QI	(99.99994)	(99.999989)	(99.99949)
200 mio.	(a)	NA	8	6.2
	(b)	8.2	6.8	6.2
	(c)	14/20	13/20	12/20
	(d)	NA	9/20	8/20
	(e)	7.6	6.6	6.2
	(f)	6.4	6	5.4
	Max. MSE $\widetilde{rr_{Guns}}$	20.5	18.6	16.2
	(rr_{Guns}, rr_{Games})	(10,1)	(10,10)	(10,1)
	Max. MSE $\widetilde{rr_{Guns,Games}}$	4007	3681	2992
	(rr_{Guns}, rr_{Games})	(10,10)	(10,10)	(10,10)
500 mio.	(a)	3.8	3.4	3.0
	(b)	3.4	3.2	3.0
	(c)	22/50	20/50	19/50
	(d)	14/50	12/50	11/50
	(e)	3.4	3.2	3.0
	(f)	3.0	2.8	2.8
	Max. MSE $\widetilde{rr_{Guns}}$	21.3	20.0	18.8
	(rr_{Guns}, rr_{Games})	(10,2)	(10,2)	(10,2)
	Max. MSE $\widetilde{rr_{Guns,Games}}$	4804	4219	3334
	(rr_{Guns}, rr_{Games})	(9.6,10)	(9.8,10)	(9.8,10)
1 bio.	(a)	2.4	2.4	2.2
	(b)	2.4	2.4	2.2
	(c)	30/100	30/100	27/100
	(d)	17/100	17/100	15/100
	(e)	2.4	2.4	2.2
	(f)	2.2	2.0	2.0
	Max. MSE $\widetilde{rr_A}$	11.5	9.65	9.42
	(rr_A, rr_B)	(10,10)	(10,10)	(10,5)
	Max. MSE $\widetilde{rr_{AB}}$	2740	2270	1874
	(rr_A, rr_B)	(10,10)	(10,10)	(10,10)

Table 3: Effects of increasing the CTC for different quantities of interest (QI), (e) and (f) explained in table 4.

4.5 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 necessarily increases probability of occurrence for all groups. From King and Zeng (2001*b*: Equation (6)), we know that variance for $\hat{\boldsymbol{\beta}}$ decreases with increasing π:

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}) &= \Big[\sum_{t=1}^{T} \pi_t(1-\pi_t)\mathbf{x}_t'\mathbf{x}_t\Big]^{-1} \\ \partial(\pi_t - \pi_t^2)/\partial\pi_t &= 1-2\pi_t > 0 \forall \pi_t \in (0, 0.5) \\ &\text{and thereby decreasing its inverse.} \end{aligned} \tag{14}$$

Therefore, under the assumption of $\hat{\boldsymbol{\beta}}$'s bias not increasing with π – for example, when doubling π and cutting population size in half – a lower relative risk will be needed to find 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 (2001*b*: Figure 4) and maybe shown from Equation (5).

4.6 Underspecified Model

Between *Guns* and *Games* there is a phi coefficient of $\phi = \frac{1}{21}$, i.e., a very weak association. Nevertheless, despite how weak this association seems to be, its effect when underspecifying the model as $Shooting \sim Guns$ is notable when looking at the power of the test in Table 4. Group "Guns'" effect is now found sooner for high relative risks in group "Games." 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 "Games" onto "Guns" (see Lee, 1982: 207, Proposition 2). Finding a "Guns'" effect sooner for a non-influential "Games" when the model is specified as $Shooting \sim Guns$ is due to reduction in variance, which itself is due to, in this case correct, model building.

QI	Population size in million	200	500	1000
(e)	Max. rr_A needed to reach a power of 0.98 ($rr_B = 1$)	7.6	3.4	2.4
(f)	Min. rr_A needed to reach a power of 0.98 ($rr_B = 10$)	6.4	3.0	2.2

Table 4: Power for different population sizes, underspecified model, $CTC = 5$

Moreover, King and Zeng's coefficient bias correction now has the most influence on the relative risk bias when the influence of "Games" is lowest instead of highest. Apart from that, results change neither in direction nor (much) in effect size.

5 Conclusion

This paper shows that even for *extremely* rare events with binary exogenous variables, the logistic regression model is well worth to study in attempting (a) find association and (b) estimate 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 simply an elaborate test for proportions.

This study revealed under what exogenous parameter settings confirmatory data analysis can be used to evaluate hypotheses derived from qualitative case studies of extremely rare events. King and Zeng's methods are very helpful, but must be applied selectively, depending on the researcher's quantities of interest. The reduction in mean squared error for the relative risk estimation compared to that achieved by the logistic regression maximum likelihood estimator is remarkable when used in the context of extremely rare events – even for a population size of 1 billion. When estimating relative risks or when searching for significance, there is no reason not to apply this correction (implemented in Imai, King and Lau, 2012) when dealing with case-control studies. Although its power does not improve dramatically, it will always offer some improvement due to the decreased bias and variance.

Based on the current paper and the work of King and Zeng (2001*b*,*a*), I suggest the following rules of thumb:

1. Effects can be found even for extremely rare events under moderate requirements for the relative risks imposed by the explanatory factors.
2. For different quantities of interest under different parameters, different methods have to be applied.
3. The more one factor's influence is hidden by another factor's influence, the more important become Equations (5) and (7).

Moreover, Westphal (2012) can be used to easily compare the methods described in Section 2 of this paper across plausible parameter sets, given a real world research problem. Indeed this should be a valid method for studying school shootings and, if properly conducted, may result in some actual quantitative evidence that may help society more effectively deal with this tragic problem.

References

Berry, William D., Jacqueline H.R. De Meritt, and Justin Esarey. 2010. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?"

American Journal of Political Science, 54(1): 248–266.

Breslow, Norman E. 1996. "Statistics in Epidemiology: The Case-Control Study." *Journal of the American Statistical Association*, 91(443): 14–28.

Genz, Alan, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. 2012. "mvtnorm: Multivariate Normal and t Distributions." R package version 0.9-9992.

Hennessy, Sean, WarrenB. Bilker, Jesse A. Berlin, and Brian L. Storm. 1999. "Factors Influencing the optimal Control-to-Case Ratio in Matched Case-Control Studies." *American Journal of Epidemiology*, 149(2): 195–197.

Imai, Kosuke, Gary King, and Olivia Lau. 2012. "Zelig: Everyone's Statistical Software." R package version 3.5.3.

King, Gary, and Langche Zeng. 2001*a*. "Explaining Rare Events in International Relations." *International Organization*, 55(3): 693–715.

King, Gary, and Langche Zeng. 2001*b*. "Logistic Regression in Rare Events Data." *Political Analysis*, 9: 137–163.

King, Gary, and Langche Zeng. 2002. "Estimating risk and rate levels, ratios and differences in case-control studies." *Statistics in Medicine*, 21: 1409–1427.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science*, 44(2): 347–361.

Kleck, Gary. 1999. "There Are No Lessons to Be Learned from Littleton." *Criminal Justice Ethics*, 18: 60–63.

Lee, Lung-Fei. 1982. "Specification Error in Multinomial Logit Models." *Journal of Econometrics*, 20: 197–209.

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

McCullagh, P., and J.P. Nelder. 1989. *Generalized Linear Models.* . 2. ed., Chapman & Hall.

Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. 1996. "A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis." *Journal of Clinical Epidemiology*, 49(12): 1373–1379.

Prentice, R. L., and R. Pyke. 1979. "Logistic disease incidence models and case-control studies." *Biometrika*, 66(3): 403–411.

Robertz, Frank J. 2004. *school shootings : Über die Relevanz der Phantasie für die Begehung von Mehrfachtötungen durch Jugendliche*. Verlag für Polizeiwissenschaft.

Robertz, Frank J., and Ruben Wickenhäuser. 2010. *Der Riss in der Tafel*. Springer.

Silvapulle, Mervyn J. 1981. "On the Existence of Maximum Likelihood Estimators for the Binomial Response Models." *Journal of the Royal Statistical Society*, 43(3): 310–313.

Vittinghoff, Eric, and Charles E. McCulloch. 2007. "Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression." *American Journal of Epidemiology*, 165(6): 710–718.

Westphal, Christian. 2012. "reccsim: Simulation of Rare Events Case-Control Studies." R package version 0.9-1.