THE USE OF RANDOMIZATION FOR LOGIT AND LOGISTIC MODELS

E. M. D. Aris, J. A. P. Hagenaars, M. Croon, and J. K. Vermunt
University of Tilburg [1] [2].

For linear models, randomization of the assignment to the levels of a variable A is a sufficient condition to obtain unbiased estimates of the effect of A on another variable B. For logit and logistic models, it is not. In particular, it has been shown that the omission of a relevant variable Z may result in a biased estimation of the effect of A and in a loss of power. These two phenomena are studied in detail here on several specific simulated cases. By using a logit model with random intercept (which is equivalent to a Directed Loglinear Model with latent variables) it is shown that the bias of the estimated effect can be partly corrected provided the number of categories of the omitted variable is known. Finally, consequences for practical use of logit and logistic models in randomized settings are underlined.

Key words: collapsibility, nonlinear model, randomized setting, simulations.

1 Introduction

In this paper, the consequences of omitting a relevant variable Z in a logit model aimed at measuring the dependency of variable B on variable A is studied (for a description of the logit model see, e.g., Agresti, 1990; Hagenaars, 1990). This study is restricted to the specific cases in which the assignment to the categories of A is randomized.

[1] **Correspondence**: Emmanuel Aris, Department of Methodology, Faculty of Social Sciences, University of Tilburg, P.O. Box 90153, 5000 LE Tilburg, The Netherlands (email: E.M.Aris@kub.nl)

[2] **Acknowledgements:** this research, conducted during the PhD study of the first author, was supported by the Work and Organization Research Center from the University of Tilburg.

If a linear model was considered, as the assignment to the level of A is supposed to be randomized, the estimate of the effect of A on B would not depend of the fact that Z is taken into account in the model or not (see, e.g., Neyman, 1990[1923], Steyer, 1988). If a logit model is considered, this property does not hold anymore and the effect coefficient of A on B may not be the same if it is estimated whether Z is taken into account or not. However, as will be seen later, the test of significance of the effect of A on B is still valid even if performed on the collapsed table.

Most[3] of the following calculations are presented for all three variables A, B, and Z being dichotomous (with values 0 or 1), and dummy coding is used to restrict the parameters. In the continuous or dichotomous cases, there is only one nonzero parameter for the effect of A on B in the logit model on table BAZ ($\alpha^{B|A}$) and in the logit model on table BA ($\beta^{B|A}$).

Let $Logit(B|A,Z)$ be the logit of variable B given variables A and Z, and $Logit(B|A)$ be the logit of variable B given variable A. The complete logit model can be represented by the equation:

$$Logit(B|A=a,Z=z) = log\left(\frac{\pi_{1\ az}^{B|AZ}}{1-\pi_{1\ az}^{B|AZ}}\right) = \alpha^{B} + \alpha^{B|A}a + \alpha^{B|Z}z, \qquad (1)$$

with $log(.)$ being the logarithmic function and $\pi_{1\ az}^{B|AZ}$ being the conditional probability that $B=1$ given that $A=a$ and $Z=z$. Note that the effect of A on B is supposed to be constant across the levels of Z. In order to simplify the calculations, the original effect coefficients are denoted α, α_{A}, and α_{Z}, such that $\alpha = \alpha^{B}$, $\alpha_{A} = \alpha^{B|A}$, and $\alpha_{Z} = \alpha^{B|Z}$. The logit model obtained from the table collapsed over Z can be represented by:

$$Logit(B|A) = \beta^{B} + \beta^{B|A}a.$$

Variable Z is said to be *strongly collapsible* if the effect coefficient of A on B calculated in the collapsed model is equal to the one calculated on the complete model (Ducharme & Lepage, 1986), i.e., if $\beta^{B|A} = \alpha^{B|A}$. Conditions for strong collapsibility for the logit model can be stated as follows:

[3] In the simulation study presented in Section 3, A is supposed to be continuous.

Theorem 1 *(Ducharme & Lepage, 1986; Guo & Geng, 1995)*
Let B, A and Z be three variables, B being categorical, A and Z being either categorical or continuous. Variable Z is said to be strongly collapsible regarding the effect coefficient of A on B if and only if:

$$Z \perp\!\!\!\perp B|A \qquad (i.e.\ \alpha^{B|Z} = 0) \qquad or \qquad Z \perp\!\!\!\perp A|B.$$

Hence, randomization to the level of A is not sufficient to obtain strong collapsibility of all Z variables. However, if variables A and Z are uncorrelated, and the entire population is considered, the value of $\beta^{B|A}$ can be shown to lie between 0 and $\alpha^{B|A}$ (Gail, 1986). Therefore, if the population is considered, although $\beta^{B|A}$ and $\alpha^{B|A}$ always have the same sign, they are not necessarily equal.

In Section 2, the relationship between $\beta^{B|A}$ and $\alpha^{B|A}$ is studied more closely and the differences between the significance tests obtained are underlined. In Section 3, the use of a random effect model is studied in order to correct for the dampening effect obtained by the omitted variable. Finally, the results obtained are summarized and put in perspective in Section 4.

2 Collapsing variables in logit models (randomized setting)

The consequences of estimating the effect of A on B by $\hat{\beta}^{B|A}$ whereas it should be estimated by $\hat{\alpha}_A$ (i.e., $\hat{\alpha}^{B|A}$) are studied first in this section. The variation of the variance of the estimates, and of the power of the test of no effect of A on B in both models, are presented subsequently.

2.1 Value of the effect estimates

Both categories of Z are assumed to be nonempty, otherwise the result is trivial. Hence, the ratio $d = \pi_0^Z/\pi_1^Z$ is a strictly positive real number. Since A is independent from Z, after some calculations, the logit of B conditional on $A = a$ and $Z = z$ obtained from Equation 1 can be written as follows (see calculations in Appendix A):

$$Logit(B \mid A = a, Z = z) \;=\; log\left(\frac{1+d\cdot exp(-\alpha_Z z)+(1+d)\cdot exp(\alpha+\alpha_A a)}{d+exp(-\alpha_Z z)+(1+d)\cdot exp(-\alpha-\alpha_A a-\alpha_Z z)}\right),$$

with $exp(.)$ being the exponential function. Hence, the estimated effect of being in category '1' rather than category '0' of A on the logit of B on the collapsed table can

be shown to be equal to:

$$\begin{aligned}
\beta^{B|A} &= \beta^{B|A}(\alpha, \alpha_{_A}, \alpha_{_Z}, d) \\
&= log\left(\frac{1+d\cdot exp(-\alpha_{_Z})+(1+d)\cdot exp(\alpha+\alpha_{_A})}{1+d\cdot exp(-\alpha_{_Z})+(1+d)\cdot exp(\alpha)} \times \frac{1+d\cdot exp(\alpha_{_Z})+(1+d)\cdot exp(-\alpha)}{1+d\cdot exp(\alpha_{_Z})+(1+d)\cdot exp(-\alpha-\alpha_{_A})}\right) \qquad (2) \\
&= log\left(Q_1(\alpha, \alpha_{_A}, \alpha_{_Z}, d) \times Q_2(\alpha, \alpha_{_A}, \alpha_{_Z}, d)\right).
\end{aligned}$$

It is easy to show that if $\alpha_{_Z} = 0$, then $\beta^{B|A}$ is equal to $\alpha_{_A}$. If $\alpha_{_A}$ is strictly positive, both $Q_1(.)$ and $Q_2(.)$ are strictly higher than 1, and $\beta^{B|A}$ is strictly positive. If $\alpha_{_A}$ is equal to zero, $Q_1(.)$ and $Q_2(.)$ are equal to 1 and $\beta^{B|A}$ is equal to zero. If $\alpha_{_A}$ is strictly negative, both $Q_1(.)$ and $Q_2(.)$ are strictly lower than 1 and $\beta^{B|A}$ is strictly negative. If $\alpha_{_Z}$ grows unboundedly, the asymptotic values of $\beta^{B|A}$ are:

$$\begin{aligned}
L_1 = \lim_{\alpha_{_Z}\to-\infty} \beta^{B|A}(\alpha, \alpha_{_A}, \alpha_{_Z}, d) &= \log\left(\frac{1+(1+d)\cdot exp(-\alpha)}{1+(1+d)\cdot exp(-\alpha-\alpha_{_A})}\right) \\
L_2 = \lim_{\alpha_{_Z}\to+\infty} \beta^{B|A}(\alpha, \alpha_{_A}, \alpha_{_Z}, d) &= \log\left(\frac{1+(1+d)\cdot exp(\alpha+\alpha_{_A})}{1+(1+d)\cdot exp(\alpha)}\right).
\end{aligned}$$

As d is strictly positive, the asymptotic values L_1 and L_2 are equal to zero if and only if $\alpha_{_A} = 0$. If $\alpha_{_A} = 0$, from Equation 2, it is easy to see that $\beta^{B|A}$ is also null. If $\alpha_{_A}$ is different from zero, there is only one root for the partial derivative, obtained for $\alpha_{_Z} = 0$ (which corresponds to the only root of $\partial\beta^{B|A}(.,.,.)/\partial\alpha_{_Z}$). This root corresponds to a maximum for $\alpha_{_A} > 0$, and a minimum for $\alpha_{_A} < 0$, in accordance with what was found in Gail (1986). Given certain values of α, d, and $\alpha_{_A}$, note that the effect value obtained in the collapsed table is a function of $\alpha_{_Z}$ only. For example, in Figure 1, the values of $\beta^{B|A}$ obtained for different values of $\alpha_{_Z}$, given $\alpha_{_A} = d = 1$ and $\alpha = 0$ are shown. Given this example, for high values of $\alpha_{_Z}$, the effect of A on B can be underestimated by 25% or more if calculated on the collapsed table. However, the estimated coefficient remains strictly positive and higher than 50% of its original value.

Note that these underestimations occur for values of $\alpha_{_Z}$ that can be much larger than the values of $\alpha_{_A}$. Suppose, as may often be the case in practice, that the effect of the omitted variable Z on B is smaller (in absolute value) than the one of A on B. Then, using Figure 1, it can be shown that, for an effect of Z less than twice the one of A, omitting Z results in an underestimation of the effect of A on B of less than 2%.

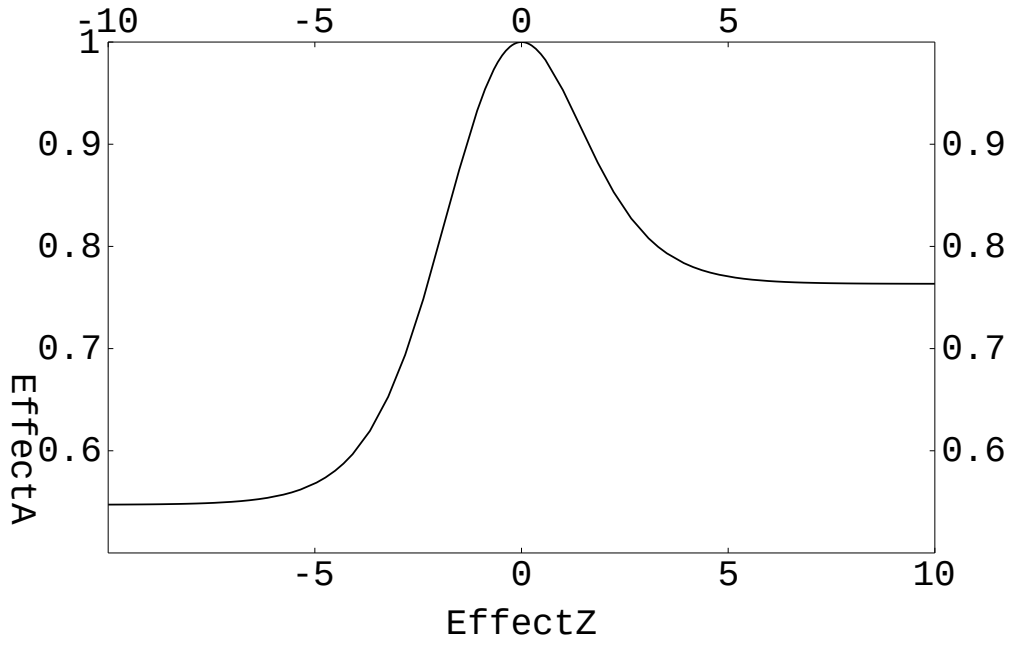

Figure 1: Value of the logit effect $\beta^{B|A}$ of A on B in the collapsed model given the value of $\alpha^{B|Z}$ (with $\alpha = 0$, $\alpha_{A} = 1$ and $d = 1$).

If the effect of Z is of the same magnitude as the one of A, the underestimation of the effect of A is of around 6% or less. Hence, if Z has been omitted in the model, but evidence exists that the effect of Z on B is more than two times smaller than the one of A, the effect of A on B may still be relatively correctly estimated from the collapsed table.

The previous calculations were performed under the assumption that d and α were equal to one and zero respectively. If now d varies from 0.5 to 10, and the effect of omitting Z on the estimate of the effect of A is studied, results shown in Figure 2 can be obtained. The underestimation is not higher than 6% for any value of d between 0.5 and 10 and any value of α_{Z} between -1 and 1. In addition, the further away d from 1, the smaller the underestimation. Hence, as d represents the distribution of Z, if Z tends to be strongly unequally distributed, the dampening of the effect coefficient of A on B is smaller.

If the value of α instead of the one of d is varied, the graphs displayed in Figure 3 can be obtained. Here also the underestimation of α_{A} is not lower than 6% if Z is omitted. Further, it seems that the higher the value of α (in absolute value) the lower

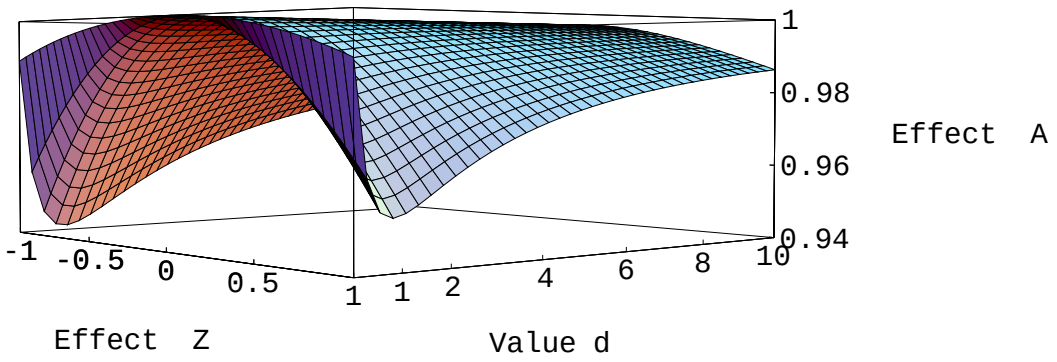

Figure 2: Value of the logit effect of A in the collapsed model (with $\alpha = 0$, $\alpha_{A} = 1$, $0.5 < d < 10$, and an effect of Z not higher than the one of A).

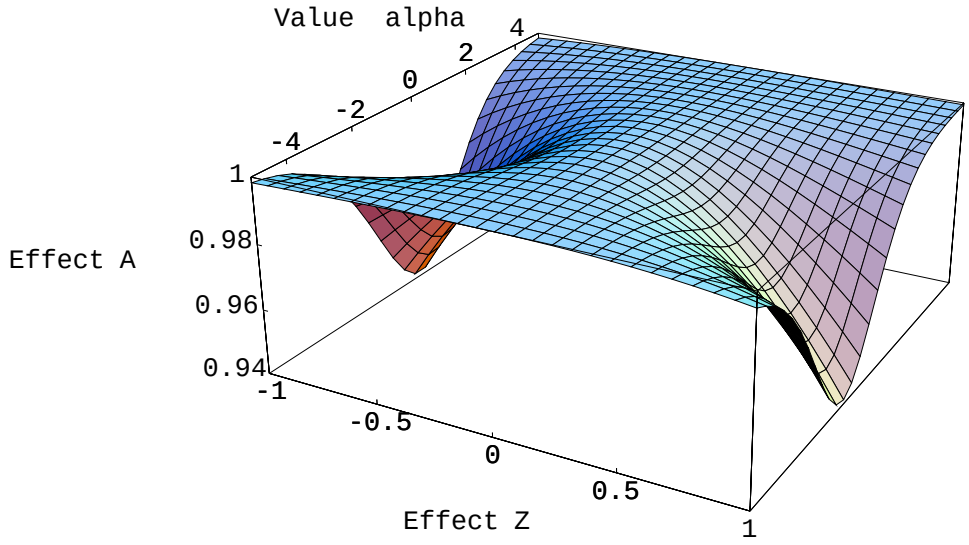

Figure 3: Value of the logit effect of A in the collapsed model (with $d = 1$, $\alpha_{A} = 1 = d$, $-5 < \alpha < 5$, and $|\alpha_{Z}| \leq \alpha_{A} = 1$.).

the underestimation. For example, for a value of α higher than 3 in absolute value, the underestimation is not larger than 2%. Note that high values of α represents variables that are strongly skewed (e.g., with $\alpha = 3$, $\pi_{1\ 00}^{B|AZ}$ is around 95%). Hence, if variable B represents very rare or very frequent phenomena, the dampening effect obtained because of the omission of relevant variables may be small (in a randomized setting).

Until now the effect of A on B was supposed to be independent of the levels of the omitted variable Z. If the effect of A on B depends on the value of Z, the original logit equation on the complete crosstable should then be written

$$Logit(B|A,Z) = \alpha + \alpha_{_A} A + \alpha_{_Z} Z + +\alpha_{_{AZ}} AZ,$$

with $\alpha_{_{AZ}}$ being the interaction effect representing the differential of effect of A on B for class 2 of Z compared to class 1 of Z. With a calculation similar to the one performed for the models without interaction effects, it is possible to write $\beta^{B|A}$ as (see Appendix A)

$$\begin{aligned} \beta^{B|A} &= \beta^{B|A}(\alpha, \alpha_{_A}, \alpha_{_Z}, \alpha_{_{AZ}}, d) \\ &= log\left(\frac{1+d\cdot exp(-\alpha_{_Z}-\alpha_{_{AZ}})+(1+d)\cdot exp(\alpha+\alpha_{_A})}{1+d\cdot exp(-\alpha_{_Z})+(1+d)\cdot exp(\alpha)} \times \frac{1+d\cdot exp(\alpha_{_Z})+(1+d)\cdot exp(-\alpha)}{1+d\cdot exp(-\alpha_{_{AZ}})+(1+d)\cdot exp(-\alpha-\alpha_{_A}-\alpha_{_{AZ}})}\right). \end{aligned}$$

For $\alpha = 0$, $d = 1$, $\alpha_{_A} = 1$, provided that the effect of Z and that the interaction effect are lower (in absolute value) than the effect of A on B, the variation of $\beta^{B|A}$ given $\alpha_{_Z}$ and $\alpha_{_{AZ}}$ is shown in Figure 4.

From this figure, several remarks may be done. In the first place, the variations are much more important than previously: for example, the underestimation is of around 20% if $\alpha_Z = -\alpha_{AZ} = 1$. Further, contrary to the previous case, the effect of A on B found in the collapsed table ($\beta^{B|A}$) can be higher than the original effect α_A, in particular for a high positive interaction effect and a low (first order) effect of Z.

The surfaces shown in Figure 5, are similar to the one of Figure 4, but α_{AZ} and α_Z are allowed this time to vary between -3 and +3. For an effect of Z of -3 and an interaction value of 3, the effect of A on B found in the collapsed table is already close to zero, whereas it was always higher than 0.5 for a model without interaction (see Figure 1).

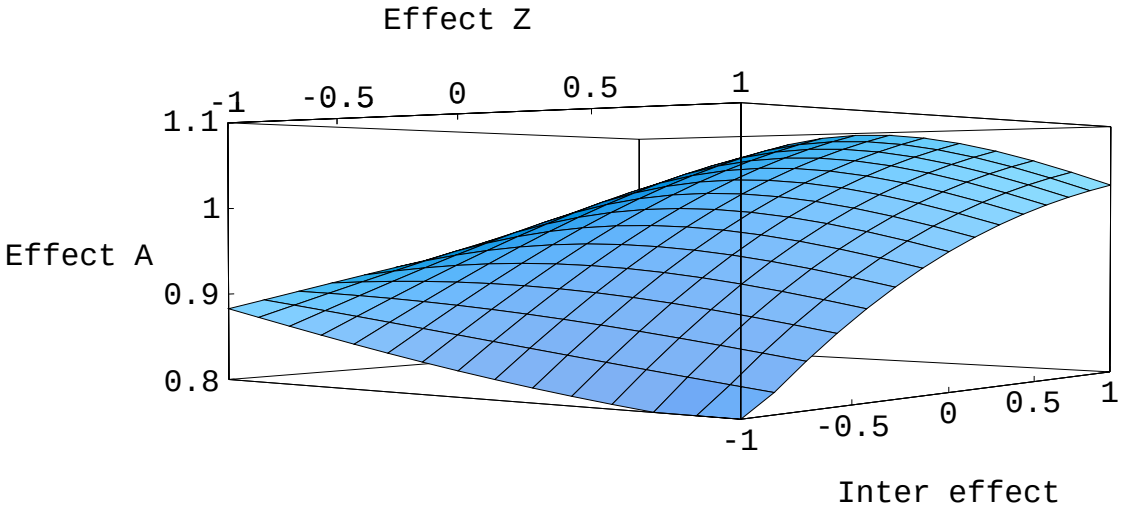

Figure 4: Value of the logit effect of A on B in the collapsed model (with $\alpha = 0$, $\alpha_{A} = 1 = d$, $d = 1$, and with the effect of Z and the interaction effect lower than α_{A}).

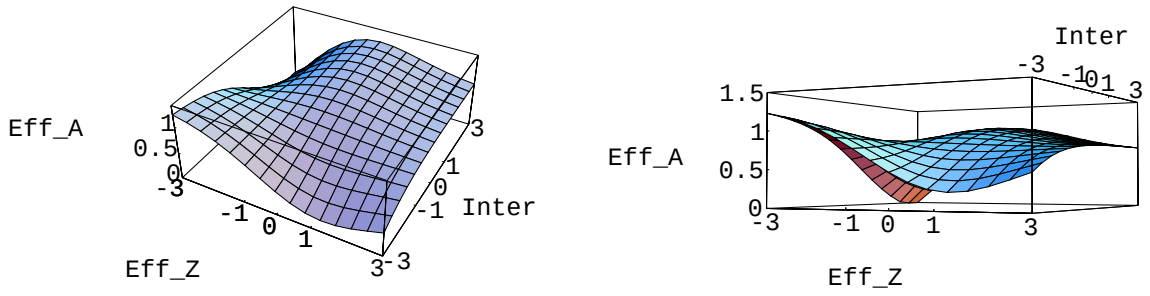

Figure 5: Value of the logit effect of A on B in the collapsed model (with $\alpha = 0$, $\alpha_{A} = 1 = d$, $d = 1$, and with the effect of Z and the interaction effect between -3 and +3).

2.2 Variance and power of null effect tests

As shown by Robinson and Jewell (1991), if one influencing background variable is omitted, the variance of the effect estimates in the logit or logistic models decreases (i.e., $var(\hat{\beta}^{B|A}) \leq var(\hat{\alpha}^{B|A})$), whereas if a linear model was considered, it would have increased. However, they also showed that, in the logit or logistic models, similarly to the linear case, the addition of background variables having an effect on B results in an increase of power for the test of null treatment effect (Robinson & Jewell, 1991). In other words, the test no treatment effect ($\alpha^{B|A} = 0$ or $\beta^{B|A} = 0$) is more powerful if performed on the complete table than if performed on the collapsed table. In order to evaluate this difference in power of the two tests for the logit model, the Asymptotic Relative Efficiency (ARE) of two tests is used. The ARE can be defined as follows.

Definition 1 *The ARE of two tests of null treatment effect ($b = 0$) provided by two estimators $\hat{b}_1$ and $\hat{b}_2$ of b is defined as (Cox & Hinkley, 1974:338):*

$$ARE_{b=0}(\hat{b}_1 \text{ to } \hat{b}_2) = \left[\lim_{b\to 0}\left\{\left(\frac{d}{db}\hat{b}_1\right)\left(\frac{d}{db}\hat{b}_2\right)\right\}\right]^2 \times \lim_{b\to 0}\frac{Var(\hat{b}_2|A)}{Var(\hat{b}_1|A)}.$$

An ARE higher than 1 means that $\hat{b}_1$ has a greater power than $\hat{b}_2$. For the logit model presented previously, the ARE of the two tests of null effect of A on B (on complete or collapsed table) can be shown to be equal to (see Appendix B):

$$ARE_{b=0}(\hat{\alpha}^{B|A} \text{ to } \hat{\beta}^{B|A}) = 1 + \frac{d\cdot exp(\alpha)\left(-1+exp(\alpha_Z)\right)^2}{(1+d)\left((1+exp(\alpha))^2\cdot exp(\alpha_Z)+d\cdot(1+exp(\alpha+\alpha_Z))^2\right)}, \quad (3)$$

or more simply $1 + K(\alpha, \alpha_Z, d)$ with a certain function $K(.,.,.)$ being always strictly positive. Therefore, the test of the hypothesis of no effect of A on B from the complete model has a greater power that the one from the collapsed model, which is in accordance with the results found by Robinson and Jewell (1991). For example, with different values of d (and α equal to zero), the ARE of the tests for the null hypothesis given different values of α_Z are shown in Figure 6.

Note that the different curves shown in Figure 6 are symmetric, hence the loss of power (given $\alpha = 0$ is as important for negative or positive influencing estimates). Further, it can be seen that the lower d, the higher the possible loss of efficiency for sufficiently large values of α_Z. This may be understood as follows. A very large d

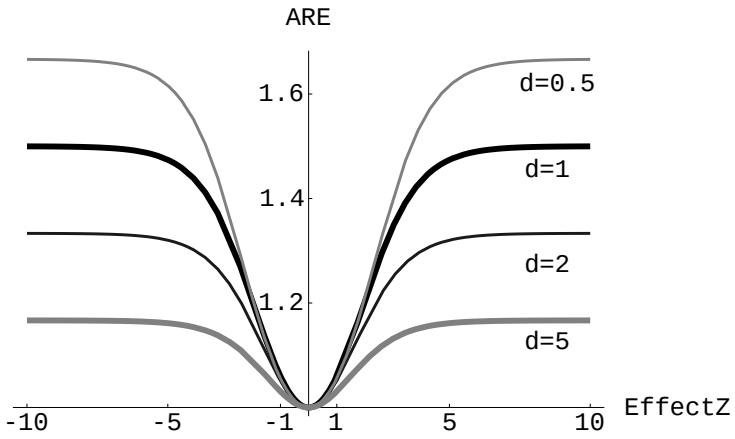

Figure 6: ARE of the null-effect test in the complete versus the collapsed model given the effect of the omitted variable Z for different values of d (with $\alpha = 0$ and $\alpha_{_A} = 1$).

indicates that a large part of the population is in the class for which $Z = 0$, i.e. the reference class. Omitting the effect the effect of $Z = 1$ again $Z = 0$ may not be much important as it is only applied to a small subpopulation (the one for which $Z = 1$). However, if d is small, then a large part of the population has $Z = 1$ and the effect of Z, that should be applied to the large subpopulation from which $Z = 1$ is not accounted for in the model on the collapsed table. Note that for low values of $\alpha_{_Z}$ (between -1 and +1), the loss of power is not much dependent on the values of d.

The variation of ARE given $\alpha_{_Z}$ for different positive values of α (with a d fixed at 1) are displayed in Figure 7. As the curves for $-\alpha$ are symmetric to the ones with $+\alpha$ (see Equation 3), only positive values for α are considered here. The different curves shown are not symmetric anymore given the vertical axis. More specifically, the higher α the lower the loss of power if an omitted Z with a positive effect on B has been omitted, and the higher the loss of power if a covariate Z with a negative effect on B has been omitted. Note that certain values of the ARE obtained by varying α were much higher than the ones obtained by varying d. But here also the differences obtained for different values of α for low values of $\alpha_{_Z}$ are not much pronounced.

Suppose that A is a treatment consisting in taking a certain drug or not (no/yes), and B is the outcome (cured/not cured). Suppose also that the large majority of

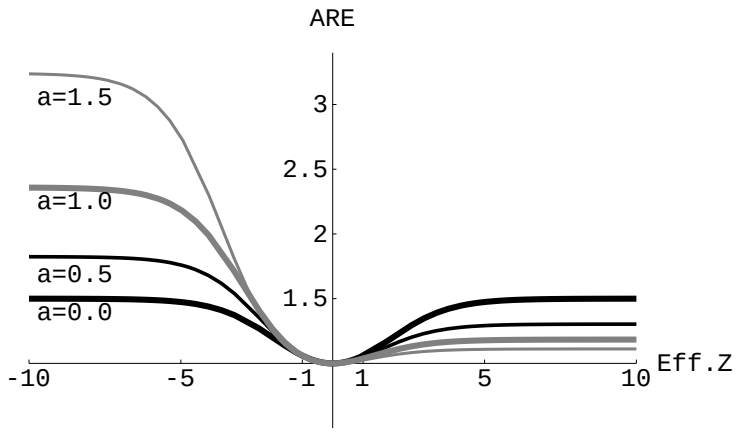

Figure 7: ARE of the null-effect test in the complete versus the collapsed model given the effect of the omitted variable Z for different values of α (with $d = 1$ and $\alpha_{_A} = 1$).

the population recover even without treatment (α is large). The asymmetry found in Figure 7 can be interpreted as follows. The loss of power for detecting a significant effect of the drug remains low if a factor (even strongly) affecting positively the recovery has been omitted, but may be high if a factor negatively affecting the recovery has been omitted.

More generally, note that the limiting values for the ARE values for a logit model without interaction terms are:

$$ARE_1 = \lim_{c\to-\infty} ARE(\alpha, \alpha_{_A}, \alpha_{_Z}, d) \;=\; 1 + \tfrac{exp(-\alpha)}{(1+d)}$$

$$ARE_2 = \lim_{c\to+\infty} ARE(\alpha, \alpha_{_A}, \alpha_{_Z}, d) \;=\; 1 + \tfrac{exp(\alpha)}{(1+d)}$$

It is easy to show that $ARE(\alpha, \alpha_{_A}, \alpha_{_Z}, d)$ is included between 1 and Max(ARE_1,ARE_2). Hence, the loss of power for the test of null effect is bounded. In particular, if the estimated coefficient $\hat{\alpha}_{_Z}$ is significantly different from zero at a certain level, and a certain sample size N, then $\hat{\beta}^{B|A}$ is also significantly different from zero at the same level with a finite sample size N_1 (with $N_1 \geq N$).

3 Use of a random effect model to correct for the dampening effect

Even though the estimation of the logit or logistic effect of A on B may be biased if a relevant variable is omitted, in certain cases and under certain assumptions, it may be possible to correct for the dampening effect resulting from this omission. This is illustrated here by some results from a small simulation study.

Indeed, assume that the logistic effect of a continuous variable A on a dichotomous variable B does not vary across the levels of a dichotomous variable Z.

Suppose that the complete (and adequate) model, can be represented by the equation:

$$Logit(B|A,Z) = \alpha^{B} + \alpha^{B|A}A + \alpha^{B|Z}Z,$$

with A being a continuous variable and Z being a categorical dichotomous variable with values 0 and 1. As seen in section 2.1, even if the assignment to the levels of A is randomized, the effect of A on B ($\beta^{B|A}$) estimated from the equation

$$Logit(B|A) = \beta^{B} + \beta^{B|A}A,$$

is lower or equal (population values) to the correct one ($\alpha^{B|A}$). If now a random effect model is used, the value of β^{B} may vary across the individuals of the population, and may therefore possibly correct for the dampening (the correction being perfect if β^{B} is equal to α^{B} for the subpopulation in which $Z = 0$ and is equal to $\alpha^{B} + \alpha^{B|Z}$ for the subpopulation in which $Z = 1$). This random effect model can be estimated by fitting a Directed Loglinear Model with latent variables (for a description of these models see, e.g., Hagenaars, 1998). Here, the adequate Directed Loglinear Model is constituted by the two hierarchical loglinear models (with sufficient statistics) { A, X} (on table A) and { BA, BX, AX } (on table AB) with X being a latent variable.

In the following, the correction obtained by considering this latent variable is studied whether X is dichotomous, trichotomous, or continuous (normally distributed). Here, only a very large sample size (N=1000000) is considered, in order to evaluate, at the population level, the values of the possible correction. The program ℓEM (Vermunt, 1997) was used to fit these Models. This program contains a modified version of the EM-algorithm (using also Newton-Raphson), that is usually faster than the standard EM algorithm and that allows to fit Directed Loglinear Models with latent variables (Vermunt, 1996:72-73).

Value of $\alpha^{B\|Z}$	Complete model	Collapsed model	Models with latent variable		
			dichot.	trichot.	contin.
0	1.00	1.00	1.00	1.00	1.00
1	1.00	0.90	0.99	1.11	0.90
2	1.00	0.71	1.00	1.01	0.72
3	1.00	0.58	1.02	0.99	0.58
4	1.00	0.52	1.00	0.96	0.51
5	1.00	0.49	1.07	0.94	0.49
6	1.00	0.48	1.02	0.94	0.48
7	1.00	0.47	1.05	0.92	0.47
8	1.00	0.47	1.02	0.94	0.47
9	1.01	0.47	1.03	0.94	0.47
10	1.00	0.47	1.05	0.91	0.47

Table 1: Value of $\hat{\beta}^{B|A}$ given $\alpha^{B|Z}$ for different models with or without correction for the dampening effect ($\alpha^{B|A} = 1$, $\alpha^{B} = 0$, $d = 1$, 1 replication, N=1000000).

Similarly to what was done previously, the value of α^{B} is set to zero and the one of $\alpha^{B|A}$ is set to 1. The value of $\alpha^{B|Z}$ varies from zero to ten by steps of one (due to the symmetry of the curve $\beta^{B|A}(\alpha^{B|Z})$, values of $\beta^{B|A}$ for negative values of $\alpha^{B|Z}$ are not considered). The results obtained are shown in Table 1. The correction of the dampening effect is almost perfect for a dichotomous latent variable, rather important for a trichotomous latent variable and almost null for a continuous normal variable. Hence, if the number of categories of the omitted variable is known and the effect of A on B is the same for each category of Z, the coefficient calculated from the collapsed table can be corrected by using this method. However, if the number of category is unknown and a latent continuous variable is assumed, no correction is obtained by considering a random effect model. Note that finally, although the correction obtained by using a dichotomous latent variable, is satisfactory, several solutions (local maxima) were obtained and only the most satisfactory one (in terms of goodness-of-fit) was presented here. Hence, the chance of obtaining a suboptimal correction is still high if the estimation procedures are run only once, and this even for a very large sample size.

4 Discussion

The problem of collapsing over relevant variables in logit/logistic models in randomized settings has been studied in this paper. The main results are briefly recalled and several consequences for practical use are underlined here.

Suppose that the assignment to the categories of the cause of interest A on the outcome B have been randomized and that both variables are dichotomous. If another dichotomous variable Z, that does not interact with A on B and that has an effect on B lower than the one of A, is omitted in the logit equation, the effect coefficient of A on B, although always lower than the true one, can still be relatively correctly estimated (with an underestimation of less than 6%, for population values). This result holds for different distribution of Z and of B, but note that this dampening decreases if either B or Z have unequal distributions rather than equal distributions. If the effect of A on B varies across the levels of Z, the bias of the effect coefficient of A may be much more important, and the coefficient obtained can even become larger than the original one. If the effect of Z and the interaction effect of Z and A on B are not higher than the effect of A, the obtained coefficient may take values between the original value minus 20% and the original value plus 10%.

If the assignment to the categories of A has been randomized and the effect of A on B is null, the regression of B on A is collapsible over all possible Z provided there is no interaction between Z and A on B. Indeed, this follows from the fact that, at the population level, $\beta^{B|A}$ is included between 0 and α_{A} (e.g., see the graph presented in Figure 1). Hence, provided the effect of A on B is supposed to be zero across the different values of Z, the test for no treatment effect of A on B in the collapsed model is still valid, albeit this test is less powerful than if Z is taken into account. However, if the effect of A on B varies across the categories of Z (and thus is not null), or if the assignment to the categories of A is not randomized, this result is not valid anymore.

Similarly to what was found in Robinson and Jewell (1991) for logistic models, the loss of power for the significance test of an effect of A on B if a relevant variable (Z) has been omitted was shown to be positive and finite for a logit model. In particular, the least observations in the reference class of Z, the higher the loss of power. And if B is skewed on one side, the omission of a variable Z that would "correct" for this skewness (i.e., for which the conditional distributions of B are less skewed than the marginal one) results in a higher loss of power than the omission of a variable Z that accentuates this skewness.

The use of a random effect model in order to correct for the dampening of the coefficient was also presented here. Provided that the latent variable used had the same number of categories than the omitted one (i.e., here two), the correction was satisfactory. However, if the latent variable was supposed to be continuous no correction was obtained. Further, several local maxima were obtained. Hence, although this method can yield satisfactory results, it should be performed with much care.

A Calculation of $\beta^{B|A}(\alpha, \alpha_{_A}, \alpha_{_Z}, d)$

In this section, the calculation of $\beta^{B|A}$ as a function of α, $\alpha_{_A}$, $\alpha_{_Z}$, and d is presented. As,

$$Logit(B = \text{`1'}/\text{`0'}|A = a, Z = a) = log\left(\pi^{B|AZ}_{1\ az}/\pi^{B|AZ}_{0\ az}\right) = \alpha + \alpha_{_A} a + \alpha_{_Z} z,$$

then as B is dichotomous, the following equalities

$$\pi^{B|AZ}_{1\ az} = \frac{exp\left(\alpha+\alpha_{_A} a+\alpha_{_Z} z\right)}{1+exp\left(\alpha+\alpha_{_A} a+\alpha_{_Z} z\right)} \qquad \text{and} \qquad \pi^{B|AZ}_{0\ az} = \frac{1}{1+exp\left(\alpha+\alpha_{_A} a+\alpha_{_Z} z\right)},$$

with $exp(.)$ being the exponential function, can be deduced. Further, since A is independent from Z (from randomization), the following equations:

$$\begin{aligned}
\pi^{B|A}_{b\ a} &= \pi^{B|AZ}_{b\ a0}\pi^{Z|A}_{0\ a} + \pi^{B|AZ}_{b\ a1}\pi^{Z|A}_{1\ a} \\
&= \pi^{B|AZ}_{b\ a0}\pi^{Z}_{0} + \pi^{B|AZ}_{b\ a1}\pi^{Z}_{1} \\
&= \pi^{Z}_0/(1 + exp(-\alpha - \alpha_{_A} a)) + \pi^{Z}_1/(1 + exp(-\alpha - \alpha_{_A} a - \alpha_{_Z})),
\end{aligned}$$

can be deduced for B equal to 0 or 1. After having simplified this previous equation, it is possible to deduce:

$$\frac{\pi^{B|A}_{1\ a}}{\pi^{B|A}_{0\ a}} = \frac{1 + d \cdot exp(-\alpha_{_Z}) + (1 + d) \cdot exp(\alpha + \alpha_{_A} a)}{d + exp(-\alpha_{_Z}) + (1 + d) \cdot exp(-\alpha - \alpha_{_A} a - \alpha_{_Z})}.$$

The logit of B conditional on $A = a$ can then be written as follows:

$$Logit(B|A = a) \;=\; log\left(\frac{\pi^{B|A}_{1\ a}}{\pi^{B|A}_{0\ a}}\right) = log\left(\frac{1+d\cdot exp(-\alpha_{_Z})+(1+d)\cdot exp(\alpha+\alpha_{_A} a)}{d+exp(-\alpha_{_Z})+(1+d)\cdot exp(-\alpha-\alpha_{_A} a-\alpha_{_Z})}\right)$$

So, as the effect parameters are restricted by effect coding:

$$\begin{aligned}
\beta^{B|A} &= \beta^{B|A}(\alpha, \alpha_{_A}, \alpha_{_Z}, d) \\
&= Logit(B|A = 1) - Logit(B|A = 0) \\
&= log\left(\frac{1+d\cdot exp(-\alpha_{_Z})+(1+d)\cdot exp(\alpha+\alpha_{_A})}{1+d\cdot exp(-\alpha_{_Z})+(1+d)\cdot exp(\alpha)}) \times \frac{1+d\cdot exp(\alpha_{_Z})+(1+d)\cdot exp(-\alpha)}{1+d\cdot exp(\alpha_{_Z})+(1+d)\cdot exp(-\alpha-\alpha_{_A})}\right) \\
&= log\left(Q_1(\alpha, \alpha_{_A}, \alpha_{_Z}, d) \times Q_2(\alpha, \alpha_{_A}, \alpha_{_Z}, d)\right).
\end{aligned}$$

Suppose now that the effect of A on B varies given the several levels of Z, then

$$\begin{aligned} Logit(B = 1/0|A = a, Z = z) &= \log\left(\pi_{1\ az}^{B|AZ}/\pi_{0\ az}^{B|AZ}\right) \\ &= \alpha + \alpha_{_A} a + \alpha_{_Z} z + \alpha_{_{AZ}} az. \end{aligned}$$

Precisely the same reasoning as previously can be performed and the formula obtained is:

$$\begin{aligned} Logit(B|A = a) &= \log\left(\frac{\pi_{1\ a}^{B|A}}{\pi_{0\ a}^{B|A}}\right) \\ &= \log\left(\frac{1+d\cdot exp(-\alpha_{_Z}-\alpha_{_{AZ}}a)+(1+d)\cdot exp(\alpha+\alpha_{_A}a)}{d+exp(-\alpha_{_Z}-\alpha_{_{AZ}}a)+(1+d)\cdot exp(-\alpha-\alpha_{_A}a-\alpha_{_Z}-\alpha_{_{AZ}}a)}\right) \end{aligned}$$

therefore,

$$\begin{aligned} \beta^{B|A} &= \beta^{B|A}(\alpha, \alpha_{_A}, \alpha_{_Z}, \alpha_{_{AZ}}, d) \\ &= log\left(\frac{1+d\cdot exp(-\alpha_{_Z}-\alpha_{_{AZ}})+(1+d)\cdot exp(\alpha+\alpha_{_A})}{1+d\cdot exp(-\alpha_{_Z})+(1+d)\cdot exp(\alpha)} \times \frac{1+d\cdot exp(\alpha_{_Z})+(1+d)\cdot exp(-\alpha)}{1+d\cdot exp(^{-\alpha_{_{AZ}}}+(1+d)exp(-\alpha-\alpha_{_A}-\alpha_{_{AZ}})}\right). \end{aligned}$$

B Calculation of the ARE of the two tests for the logit models

As the ARE of two tests of null treatment effect ($b = 0$) provided by two estimators $\hat{b}_1$ and $\hat{b}_2$ of b is equal to :

$$ARE_{b=0}(\hat{b}_1 \text{ to } \hat{b}_2) = \left[\lim_{b\to 0}\left\{\left(\frac{d}{db}\hat{b}_1\right)\left(\frac{d}{db}\hat{b}_2\right)\right\}\right]^2 \times \lim_{b\to 0}\frac{Var(\hat{b}_2|A)}{Var(\hat{b}_1|A)},$$

then it is possible to show that (Robinson & Jewell, 1991):

$$ARE_{b=0}(\hat{b}_1 \text{ to } \hat{b}_2) = \frac{E\left[\pi_{b\ 0z}^{B|AZ}\right]\times E\left[1-\pi_{b\ 0z}^{B|AZ}\right]}{E\left[\pi_{b\ 0z}^{B|AZ}(1-\pi_{b\ 0z}^{B|AZ})\right]}$$

with $E[.]$ denoting the expected value for the possible values $Z = z$. Hence,

$$ARE_{b=0}(\hat{\alpha}^{B|A} \text{ to } \hat{\beta}^{B|A})$$

$$= \frac{\left(\pi_0^Z/(1+exp(-\alpha))+\pi_1^Z/(1+exp(-\alpha-\alpha_{_Z}))\right)\left(\pi_0^Z exp(-\alpha)/(1+exp(-\alpha))+\pi_1^Z\cdot exp(-\alpha-\alpha_{_Z})/(1+exp(-\alpha-\alpha_{_Z}))\right)}{\pi_0^Z(exp(-\alpha)/(1+exp(-\alpha))^2)+\pi_1^Z\left(exp(-\alpha-\alpha_{_Z})/(1+exp(-\alpha-\alpha_{_Z}))^2\right)}$$

which, after simplification can be shown to be equal to

$$\frac{\left(1+d+e^{\alpha}+d\cdot exp(\alpha+\alpha_{_Z})\right)\left(d+(1+(1+d)\cdot exp(\alpha_{_Z})\cdot exp(\alpha)\right)}{(1+d)\left((1+exp(\alpha))^2 exp(\alpha_{_Z})+d(1+exp(\alpha+\alpha_{_Z}))^2\right)},$$

or

$$1+\frac{de^{\alpha}\left(-1+exp(\alpha_{_Z})\right)^2}{(1+d)\left((1+exp(\alpha))^2 exp(\alpha_{_Z})+d(1+exp(\alpha+\alpha_{_Z}))^2\right)}.$$

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics.* London: Chapman and Hall.

Ducharme, G. R., & Lepage, Y. (1986). Testing collapsibility in contingency tables. *J. Roy. Statist. Soc., B 48*, 197-205.

Gail, M. H. (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In S. H. Moolgavkar & R. L. Prentice (Eds.), *Modern statistical methods in chronic disease epidemiology* (p. 3-18). New York: Wiley.

Guo, G. H., & Geng, Z. (1995). Collapsibility of logistic regression coefficients. *J. Roy. Statist. Soc., B 57*, 263-267.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear panel, trend and cohort analysis.* Newbury Park: Sage.

Hagenaars, J. A. P. (1998). Categorical causal modeling: latent class analysis and directed loglinear models with latent variables. *Sociological Methods and Research, 26*, 436-486.

Neyman, J. (1990[1923]). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science, 5*, 465-472.

Robinson, L. D., & Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review, 58*, 227-240.

Steyer, R. (1988). Randomized experiments: some interpretational issues. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research: volume 2, data analysis* (p. 74-93). Hong Kong: The Macmillan Press LTD.

Vermunt, J. K. (1996). *Log-linear event history analysis.* Tilburg: Tilburg University Press.

Vermunt, J. K. (1997). *ℓEM: a general program for the analysis of categorical data.* Tilburg: Technical report, MTO Department, Tilburg University.