Logistic regression models for single–source data — a simulation study

by
Tue Tjur
Department of Management Science and Statistics
Copenhagen Business School

Summary

For panel (single source) binary data, logistic regression models can produce very misleading results if the variation between respondents is ignored. Various solutions to this problem are discussed. It is argued that the best way of doing it is by a conditional logistic regression model, where the conditioning taking place is on the total number of positive responses for each respondent. However, this model estimates only the within–respondent effects. The between–respondent effects can be estimated by an over–dispersion model with the respondent totals as (overdispersed binomial) responses and within–respondent averages of the original covariates as explanatory variables. The performance of different approaches are analysed by simulation studies.

1. Introduction.

Birch (2002) has argued that logistic regression in longitudinal data, also called panel data, or *single source data* in the marketing context, can produce serious inference errors when heterogeneity between respondents is ignored. What can happen is that an explanatory variable with essentially no effect on the binary responses appears to be strongly significant. Roughly because the reuse of the same respondents again and again results in a phenomenon which — in the most extreme case where the behaviour of the respondents is constant over time — is similar to what happens when all the counts in a 2×2 contingency table are multiplied by some large factor. What happens in this extreme situation is that the χ^2 test statistic (Pearson or –2 log(likelihood ratio)) for independence is multiplied by the same factor, and this will obviously make it "significant" far too often, even when there is independence in the original table.

This problem is wellknown, and from a strictly theoretical point of view there is really not much to say about it, apart from the triviality that an incorrect model can not be expected to produce correct results. From a more practical point of view, there is a little more to say. Logistic regression is a standard tool in this context, and since the conclusions coming out of it quite often appear reasonable and in agreement with common sense (see e.g. Hansen and Hansen 2001), it is of some interest to study how serious the problem is. With the partial purpose to illustrate the difference between logistic regression and some of the alternatives (see section 2 of this paper), Birch (2002) analysed a number of such data sets. However, no unique conclusion came out of this. The purpose of the present paper is to show how simulations can be used to make a quantitative analysis in a concrete case, and to draw some general conclusions from this.

For the kind of marketing data we study, where the response is the indicator for the event that a certain brand of a consumers good is preferred in a purchase, and the explanatory variable is a measure for the consumers exposure to advertisements for that brand, it has been suggested that the problem with heterogeneity can, at least to some extent, be eliminated by the inclusion of "loyalty" as an explanatory variable in the model. A loyalty, in this context, means a measure for the respondents tendency to prefer this specific brand in the near past, for example the relative frequency over the last 5 or 10 purchases. We return to this question in section 4.

It should be emphasized that our use of the phrase "conditional logistic regression" has nothing to do with what Nordmoe and Jain (2000) call the "conditional logit model". Our approach is different because we are only considering a single brand (among others) at a time. Our analogue to Nordmoe and Jain's conditional logit model is the (ordinary, i.e. **un**conditional) logistic regression model, and our solution to the over–dispersion problem (or within–consumer–correlation problem) is quite different from their's.

2. The data and some statistical results.

The data set studied here constitutes a very small corner of a tiny little bit of the British AdLab data base, created by Central Independent Television 1985–90 (see Moseley and Parfitt 1987), kindly made available to us for research purposes by Flemming Hansen, Forum for Advertising Research. The data set, which has been extracted from the data base by Lotte Yssing Hansen and further prepared by Kristina Birch, consists of all purchases of chocolate bars over observation periods of varying lengths, made by 560 households, adding up to a total of 11246 such purchases, i.e. around 20 purchases per household on average. The binary response is defined as 1 if the chocolate bar happens to be a Mars Bar, 0 otherwise. The only explanatory variable considered here is constructed as a weighted average of the counts of television and radio advertisements for Mars Bar that the household was exposed to on day 1, 2, . . . , 28 before the purchase. The weights used in this averaging are proportional to 0.95^d, where d is the number of days passed since the advertisement was seen. A lot of details concerning data stucture, other background variables, the choice of 0.95 as the "retention rate" etc. are ignored here, because they are irrelevant to the general ideas discussed.

Logistic regression. The result of an ordinary logistic regression analysis of this data set of length 11246 with "Mars/not Mars" as the binary response and a logit–linear structure consisting only of a constant term and a linear effect of the above mentioned variable EXPO (ignoring differences between consumers) results in the following conclusions. The estimated probability that a given purchase of chocolate bar results in the choice of a Mars Bar is

$$\frac{\exp(-0.8548 + 0.02007 \times \text{EXPO})}{1 + \exp(-0.8548 + 0.02007 \times \text{EXPO})}.$$

The coefficient 0.02007 to EXPO is significantly positive. The Wald test (i.e. the test produced by any statistical package, where the estimate divided by its approximate standard deviation is evaluated in a standardised normal distribution) reports a two–sided P–value of 0.000005.

Conditional logistic regression. The obvious problem with the analysis presented above is that it does not take into account the fact that consumers have different tastes concerning chocolate bars. One can easily imagine that some buy Mars Bars most of the time, some buy other brands most of time, rather independently of the commercials that happened to be running on their TV set. A simple way of accounting for this is by the introduction of 560 consumer parameters α_c describing these differences. Thus, we assume that the probability of success in the ith purchase for consumer c takes the form

$$P(y_{ci} = 1) = \frac{\exp(\alpha_c + \beta \times \text{EXPO}_{ci})}{1 + \exp(\alpha_c + \beta \times \text{EXPO}_{ci})}.$$

The simplest and best way of estimating this model — since we are not particularly interested in the 560 consumer parameters — is by *conditional logistic regression*. By conditioning on the numbers of successes for each consumer, we obtain an expression for the conditional likelihood, where the consumer parameters α_c cancel out. This analysis results in the following conclusions. The estimated probability that a given purchase of chocolate bar results in the choice of a Mars Bar becomes

$$P(y_{ci} = 1) = \frac{\exp(\alpha_c + 0.00802 \times \text{EXPO}_{ci})}{1 + \exp(\alpha_c + 0.00802 \times \text{EXPO}_{ci})}.$$

However, the coefficient 0.00802 is *not* significantly positive; the Wald test reports a two–sided P–value of 0.31, thus strongly contradicting the results of the simple logistic regression.

An overdispersion model for the consumer totals. There are two ways in which the exposure to Mars commercials could influence the consumers tendency to buy Mars Bars, namely

(1) *Within consumers.* If a consumer is heavily exposed to Mars advertisements, he/she tends to buy more Mars Bars than usual in a period thereafter.

(2) *Between consumers.* Those consumers that watch many Mars commercials buy, on average, more Mars Bars than those who watch few Mars commercials.

These two types of exposure effect correspond closely to "within blocks" and "between blocks" effects in classical analysis of variance. There is no reason, a priori, to expect these two effects to be equal. Nevertheless, a strong argument in favour of the simple logistic regression against the conditional logistic regression is that the conditional model measures only the effect of type (1) above. Any effect of type (2) is confounded with the 560 consumer parameters, and therefore not taken into account by the conditional analysis.

It can be argued that a correct model should involve a random effect of consumers, and perhaps even a seperate EXPO–coefficient for each of the two types of exposure effect. However, a simple way of estimating the "type (2)" effect by an additional analysis, which is very similar to classical standard methods for "recovery of interblock information", goes as follows. Consider, as our new reponses, the 560 sums within consumers of the original binary responses (i.e. the sums we conditioned on before),

$$y_{c.} = \text{ the number of Mars purchases for consumer } c.$$

A naive model might assume that these counts are binomial with totals (indices)

$$n_c = \text{ the total number of purchases for consumer } c$$

and a probability parameter that depends logit–linearly of the consumers' average exposures to Mars commercials. It can be expected, however, that this model will show some over–dispersion due to the differences between consumers mentioned earlier. A simple model that takes this into account is the overdispersion model corresponding to this binomial model, where the expected responses are assumed to be of the same form as in the binomial model, but the binomial variances are modified by a common scale factor (the overdispersion parameter). See McCullagh and Nelder (1989), Tjur (1998). The conclusion of this model is as follows. The estimated expectations of the responses $y_{c.}$ are

$$\mathrm{E}y_{c.} = n_c \frac{\exp(-0.965 + 0.0437 \times \text{EXPO}_c)}{1 + \exp(-0.965 + 0.0437 \times \text{EXPO}_c)}$$

where EXPO_c denotes a suitably defined average over time of consumer c's exposure to Mars advertisements. The test for "no effect of EXPO" shows only weak significance (P $=$ 0.034 two–sided). This test is a T–test, correcting for overdispersion and the estimation error for the overdispersion parameter. There is a strong overdispersion, corresponding to a standard deviation which is almost three times that of the binomial model. A plot of normed residuals against the totals n_c suggests that a model with variance proportional to $n_c^{3/2}$ or n_c^2 rather than n_c would be more realistic. This was tried, and it actually resulted in a more stable behaviour of the normed residuals, but the effect of exposure became completely insignificant ($P > 0.5$).

It can be argued — and we certainly do insist — that the combination of

(1) a conditional logistic regression of the responses y_{ci}, given the consumer sums $y_{c.}$ (for the analysis of type (1) effect), and

(2) an analysis of the consumer sums $y_{c.}$ by a standard binomial overdispersion model, perhaps with a modified variance function (for the analysis of type (2) effect)

together constitute an exhaustive analysis of the data set. Thus, our conclusion is that there is no exposure effect at all in this data set, apart from a vague type (2) tendency based on a P–value of 3.4%.

It has, nevertheless, been argued that the simple logistic regression model may have some element of thruth in that goes beyond what can be found in this way. And it has, in particular, been argued that the inclusion of a "loyalty" variable in this model may be able to account for the heterogeneity in a sufficient manner. This is the topic of section 4. Here, we proceed with a simulation study of the logistic models' tendency to create false significances.

3. Some simulations.

More precisely, we construct 10000 data sets of the same kind as the one considered above in the following way. The total design, including the number of consumers, the lengths of observation periods for consumers and the values of the explanatory variable EXPO are kept fixed and equal to the values we have in the original setup. But the responses y_{ci} are constructed in each of the 10000 cases as follows. First, 560 consumer parameters α_c are drawn from a normal distribution with mean –0.97 and standard deviation 2.0. Then, responses y_{ci} are generated according to a logistic model with these consumer parameters and no exposure effect ($\beta = 0$). Each of these data sets are analysed both by a logistic regression model and a conditional logistic regression model, and the "left sided P–values"

$$R = \Phi\left(\frac{\hat{\beta}}{\sqrt{\mathrm{var}(\hat{\beta})}}\right)$$

are computed. Since the data sets, by the way they are constructed, are guarantied to be without an exposure effect, we would expect these quantites R to follow a uniform distribution on the unit interval — provided that the test performed is a valid one.

The fixed parameter values –0.97 for the mean and 2.0 for the standard deviation in the distribution of the generated consumer parameters were chosen to make the constructed data sets appear as similar as possible to the original data set, in the following sense. The constant –0.97 is simply the estimate of the constant term in the overdispersion model. The value 2.0 for the standard deviation was chosen by trial and error as a value that produced approximately the same degree of overdispersion on average as observed for the original data. In this sense, our simulated data sets are as similar as possible to the original data set, as far as the different sources of variation and the average of the responses are concerned.

The histogram (with division of the unit interval into 100 subintervals of length 0.01) and cumulated cdf. for the R–values that came out of the 10000 conditional models are shown in figure 1, the similar pictures for the ordinary logistic regression models are shown in figure 2.

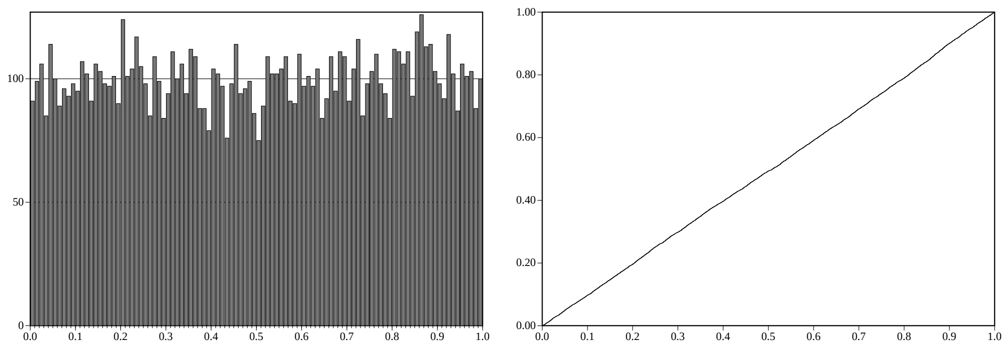

Figure 1. Test for "no effect of exposure" in data without such an effect. Distribution of 10000 P–values in the conditional logistic regression model.

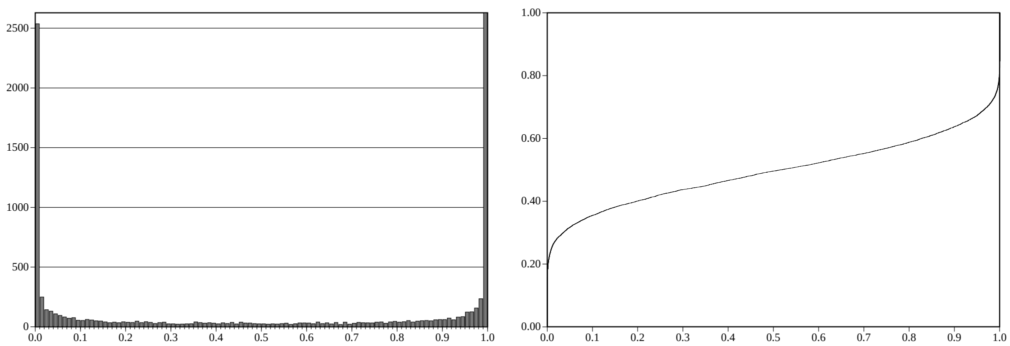

Figure 2. Test for "no effect of exposure" in data without such an effect. Distribution of 10000 P–values in the logistic regression model.

The conclusion from these two figures is obvious. The R–values from the conditional models behave exactly as they should according to standard asymptotic theory. The values from the unconditional model come close to the endpoints far too often. More than half of the values are in the two extreme intervals corresponding to "one–sided significance on level 0.01". What you can not see from this figure is that 1990 of the 10000 values — almost 20 % — are greater than 0.9999975 or smaller than 0.0000025, which means that they appear as more significant than the value observed in the original data set. Thus, the P–value 0.000005 from the logistic regression of the original dataset is totally misleading.

4. The inclusion of a "loyalty" variable.

An idea that may appear promising at first sight is to include an explanatory variable in the logistic regression model which somehow accounts for the fact that consumers have different attitudes to the brand in question. Of course, we can not just use the consumers' average consumptions of Mars (or a transformed version of this) as an explanatory variable, because this would imply, more or less, that we consider the responses as explanatory for themselves. But we can use an idea which is wellknown from the analysis of autoregressive models in time series analysis. It is wellknown that such models can (with little loss of information, if any at all) be handled as ordinary regression models, where "lagged" versions of the response variable occur as explanatory variables. An argument for this is that the likelihood function in the regression model can be interpreted as a conditional likelihood in the autoregressive model, by conditioning on the first few observations. By exactly the same argument, we can handle an ordinary logistic regression model which has, say,

$$\text{LOYAL}_{ci} = \frac{y_{c,i-5} + y_{c,i-4} + y_{c,i-3} + y_{c,i-2} + y_{c,i-1}}{5}$$

(i.e. the releative frequency of Mars purchases among the last five chocolate bar purchases) as an explanatory variable, provided that the first five observations for each consumer (for which the new covariate is undefined) are removed from the data set. In this model, the expression

$$P\left(y_{ci} = 1\right) = \frac{\exp(\alpha + \beta \times \text{EXPO}_{ci} + \gamma \times \text{LOYAL}_{ci})}{1 + \exp(\alpha + \beta \times \text{EXPO}_{ci} + \gamma \times \text{LOYAL}_{ci})}.$$

should be interpreted as a *conditional* probability, given the consumers behaviour in the past. But the likelihood has the same form as the likelihood for a logistic regression, by the multiplication rule for conditional probabilites.

The estimate of β in this model (for the original data set) is 0.01067, with an estimated standard deviation of 0.006283. However, the Wald statistic 1.699=0.01067/0.006283 is insignificant (twosided P = 0.09). This means that the inclusion of the new covariate LOYAL has actually removed the false significance that was present in the analysis without this covariate. The estimate of γ is significantly positive (4.3484, with a standard deviation of 0.09018), which confirms the idea that the term $\gamma \times \text{LOYAL}_{ci}$ does (at least partially) take over the role of the individual parameters α_c.

In a conditional logistic regression analysis of the same data and with the same explanatory variables it could, perhaps, be expected that the effect of LOYAL would be insignificant, because individual parameters have already been accounted for by the conditioning. However, in this model the coefficient γ to LOYAL was estimated to be 1.358 with a standard deviation of 0.1336, thus significantly greater than zero (P=0.000000). The immediate interpretation of this is that the (given number of) Mars purchases for each consumer tend to cluster more than they would if they were spread out at random. But as we will see below, some reservations must be made here, since the corresponding estimates in simulated data sets (that are guaranteed to be without such a clustering) are, quite generally, negative.

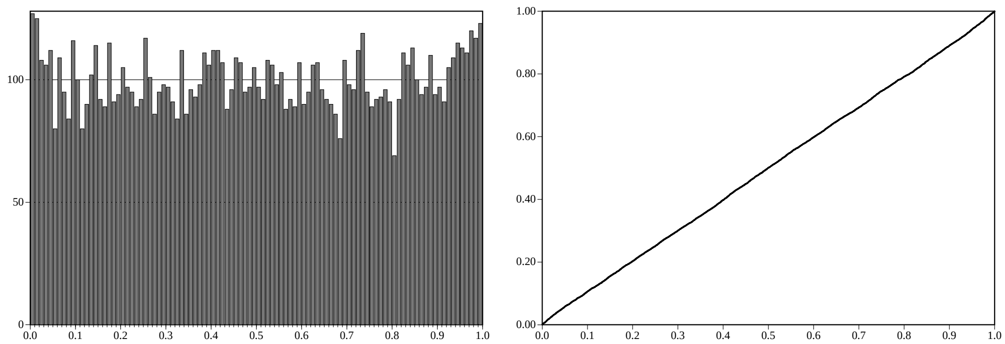

Figure 3. Test for "no effect of exposure" in data without such an effect. Distribution of 10000 P–values in the conditional logistic regression model, including an effect of loyalty.

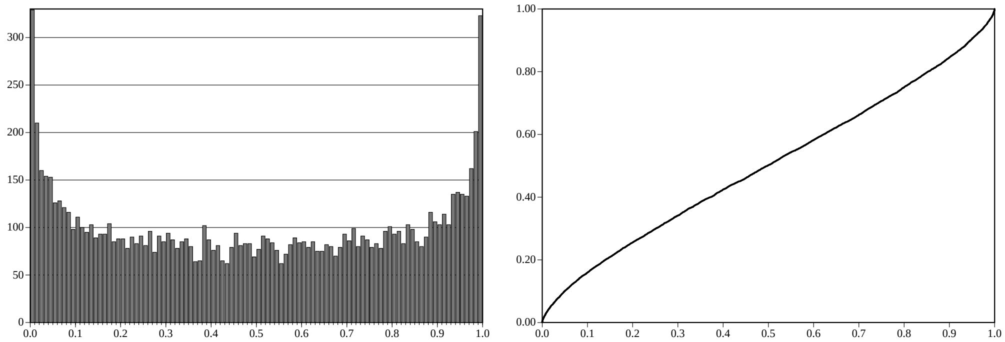

Figure 4. Test for "no effect of exposure" in data without such an effect. Distribution of 10000 P–values in the logistic regression model, including an effect of loyalty.

A simulation study shows a more accurate picture. 10000 data sets were created in exactly the same way as in section 3. The histograms and cdf.s for the resulting "R–values" are shown in figure 3 for the conditional logistic regression and in figure 4 for the ordinary logistic regression, with inclusion of the covariate LOYAL as an explanatory variable.

Figure 3 shows that the R–values from the conditional model are almost uniformely distributed. Figure 4 shows that this is not the case for the ordinary logistic regression. However, the model's tendency to produce false significances is much smaller than for the model without LOYAL. We have, for example, approximately 600 (out of the 10000) values which are formally significant on two–sided level 0.01, where the corresponding number for the model without LOYAL was more than 5000. Without further documentation, we mention that if LOYAL is defined as the average of the 10 (instead of 5) previous responses, the distribution of the R–values becomes almost uniform.

It is of some interest how the estimates of γ (the coefficient to LOYAL) behave for these models. What we would expect is, of course, a strongly positive effect of LOYAL for the logistic regression, because the whole idea is that the term $\gamma \times \text{LOYAL}_{ci}$ is a substitute for the consumer parameter α_c, which should obviously have large values for consumers with a high inclination towards Mars Bars. The distribution of the 10000 estimated values of γ is shown as the first histogram of figure 5. All values are greater than 3.

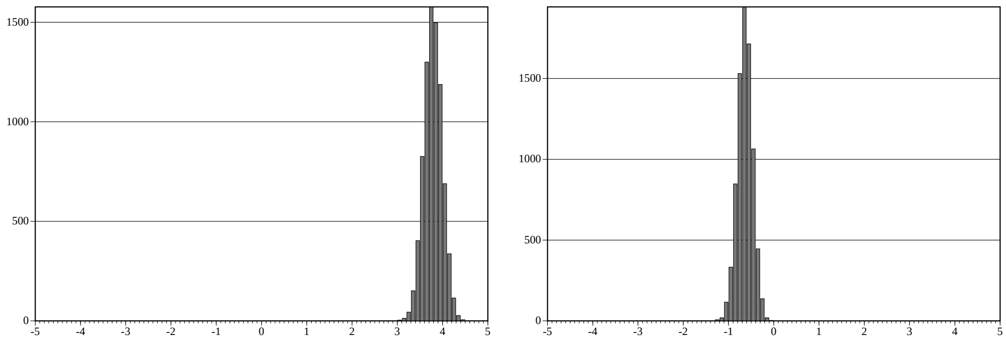

Figure 5. Distributions of the estimates of the coefficient to LOYAL in the 10000 simulations for the logistic (left) and the conditional logistic (right) models.

The second histogram of figure 5 shows the same distribution for the conditional model. Here, all values are negative. This may appear a bit surprising, at first sight, but if one thinks a little about it it is no surprise at all. A large value of LOYAL indicates that the consumer has bought many Mars Bars until now, and since the total number of Mars purchases is fixed (conditioned on) in this model, the number of Mars purchases that remain to be done must be small. It is questionable whether this model makes sense at all. At least, it is quite difficult to understand what is going on when the idea of "conditioning on the past" is applied to a model where the consumer totals $y_{c.}$ are fixed. The (significantly) positive estimate for the same parameter γ in the original data set indicates that the clustering here is even more pronounced than the actual value (whatever that means) of the estimate indicates.

5. Estimation bias.

We may, with some reservations, conclude from the simulation studies of section 4 that the inclusion of a loyalty covariate in the logistic regression model can remove the models tendency to exaggerate the significance of the exposure effect. If the corresponding P–value is extremely small, it is probably safe to conclude that there is an exposure effect, even though the exact P–value should not be taken too seriously. However, correctness of the Wald test is not all we can or should require from the method, correctness of the estimate of the exposure effect (in case it is there) is equally important. To investigate this, we performed another simulation study, involving 10000 data sets constructed as follows. As in section 3 and 4, all design quantities were taken from the original data set, including the values of EXPO, and random consumer parameters were constructed in the same way. But the responses were constructed according to a logistic regression model with the value $\beta = 0.1$ of the coefficient to EXPO, which is about ten times the estimate obtained from the conditional logistic regression in the original data set. In each case, both an ordinary and a conditional logistic regression model with EXPO and LOYAL as explanatory variables was fitted. Figure 6 shows what came out of this. The left histogram shows the distribution of the β–estimates for the simple logistic regression. For some reason these estimates are centered around 0.05, and almost all of them are less than 0.075. Thus, the estimate is seriously biased. For the conditional model (the right histogram), the estimates are centered around the true value 0.1.

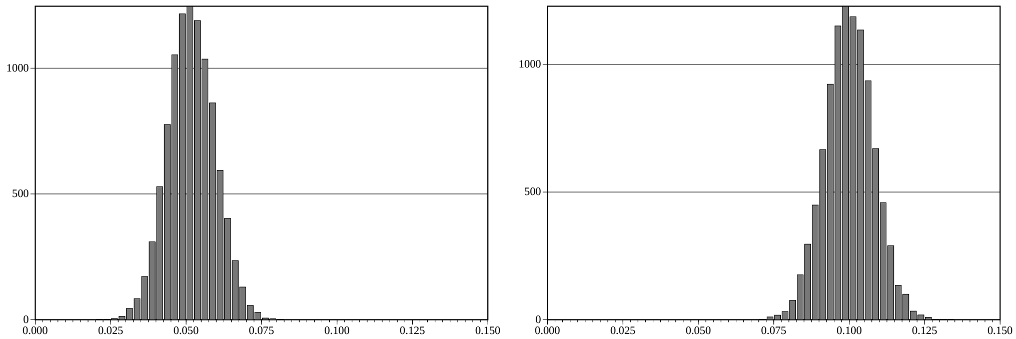

Figure 6. Distributions of the estimates of the coefficient to EXPO in 10000 simulations for the logistic (left) and the conditional logistic (right) models. True value β=0.1, LOYAL included.

A final natural question is whether this bias has to do with the inclusion of the loyalty variable, or it is just an intrinsic property of the logistic regression model when applied to such data sets. To investigate this we did exactly the same simulations once more, but the estimates computed were for the models (unconditional and conditional logistic regression) *without* LOYAL as explanatory variable. To ensure compatibility with the results shown in figure 6, we still excluded the first five observations for each consumer. Here, we would certainly expect the conditional model to work at least as well as it does when LOYAL is included as an explanatory variable. For the unconditional model, all we know is that this model is rather useless when the task is to test $\beta = 0$, but when the problem is to give a point estimate of β we can not be sure what will happen. However, figure 7 shows that it does not make much difference whether LOYAL is included in the model or not. The bias is slightly less pronounced than for the model including LOYAL, but still very strong, with a mean of estimated values only slightly above 0.6. The dispersion has increased, obviously because much of the between–consumer variation that was accounted for by the loyalty effect is now present as a component of the estimation error. The same phenomenon for $\beta = 0$ is an ingredient in the phenomenon studied in section 3 (significance too often when testing $\beta = 0$). For the conditional model, the distribution of the estimates has hardly changed at all.

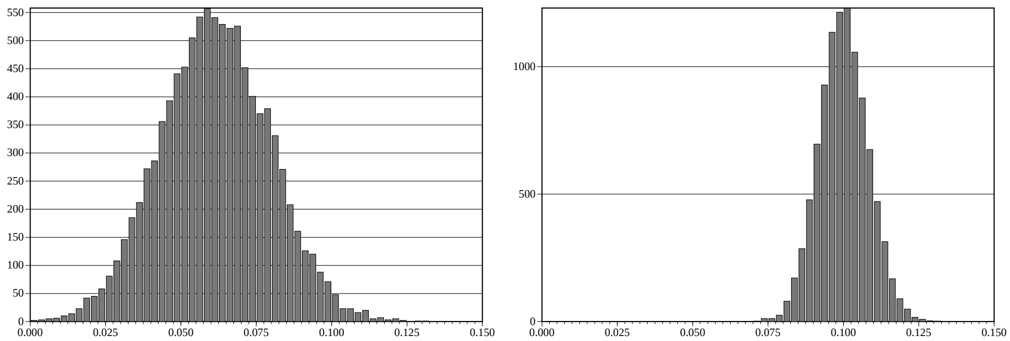

Figure 7. Distributions of the estimates of the coefficient to EXPO in 10000 simulations for the logistic (left) and the conditional logistic (right) models. True value β=0.1, LOYAL not included.

6. Conclusions.

Our general conclusion is a warning against the use of ordinary logistic regression for this kind of data. It may be possible to remove this model's tendency to produce false significances by inclusion of a "loyalty" covariate in the model, but it is not clear at all how and when this works. Moreover, whether this covariate is included or not, the estimates of the parameters of interest are likely to be seriously biased. We recommend a combination of a conditional logistic regression model for the analysis of within–consumer effects and a logit–linear model with the consumer totals $y_{c.}$ as "overdispersed binomial responses" for the analysis of between–consumer effects.

7. A commercial. All computations related to this study were performed by the statistical package ISUW, see `www.mes.cbs.dk/~sttt/`. ISUW has a command `FITCLOGIT` for estimation in conditional logistic regression models, and a command `FITNONLINEAR` which can perform the estimation in generalized linear models with overdispersion.

References.

Birch, K. (2002) Analyzing effects of advertising using conditional logistic regression.
Preprint no. 2, Dept. of Management Science and Statistics, Copenhagen Business School.

Hansen, L.Y. and Hansen, F. (2001)
Advertising and promotion effectiveness — learnings from a five year study.
Research Paper no. 18, Advertising Research Group, Dept. of Marketing, Copenhagen Business School.

McCullagh, P. and Nelder, J. A. (1989).
Generalized Linear Models.
Chapman and Hall.

Moseley, S. and Parfitt, J. (1987)
Measuring advertising effect from single–source data: the first year of the AdLab panel.
Admap, June 1987, pp. 26–33.

Nordmoe, E.D. and Jain, D.C. (2000)
Drawing inferences from logit models for panel data.
Applied Stochastic Models in Business and Industry **16**, pp. 127–145.

Tjur, T. (1998)
Nonlinear regression, quasi likelihood, and overdispersion in generalized linear models. *The American Statistician* **52**, 222–227.

Tue Tjur
The Statistics Group
Copenhagen Business School
Solbjerg Plads 3
DK–2000 Frederiksberg
e–mail tuetjur@cbs.dk