A warning concerning random effects and random coefficients in logistic regression models for binary data.

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

Summary

For binary panel data, the introduction of a random respondent effect in a logistic regression model is a useful way of taking respondent heterogeneity into account. More generally, logistic regression models with random coefficients can be used if not only the intercept, but also the coefficients to explanatory variables can be expected to vary from respondent to respondent. However, there are some identifiability problems with these models in the special case where respondents are observed only once. A clarification of these matters can be obtained by studying the probit–linear model rather than the logit–linear model. In practice this change of link function makes very little difference. But the advantage of the probit models is that the identifiability problems — which in the logit models with normal random effects merely result in numerically unstable solutions to the likelihood equations — correspond to mathematically exact overparametrizations in the probit–linear models.

1. Introduction.

Suppose that a coin is flipped 100 times. Consider the following two models for the outcome of this experiment.

(1) *The standard model.* The results are outcomes of 100 independent Bernoulli trials with the same parameter p.

(2) *An "overdispersion model".* The results are outcomes of 100 Bernoulli trials with different parameters $p_1, p_2, \ldots, p_{100}$, which in turn have been generated as i.i.d. observations from a distribution on the unit interval.

It is rather obvious that whatever we do to decide which of the two alternative models we prefer, the data will be of no help whatsoever. There is no way of detecting whether a Bernoulli variable originates from a random or a fixed p. It is a Bernoulli variable anyway.

In more exclusive terms, this means that overdispersion in a model for homogeneous (i.e. with the same p) binomial variables can not be detected when all the binomial totals are 1. This triviality is essentially what the present paper is about. But as we shall see, things become less transparent when models with covariate effects are considered.

The following is a quite general discussion of what happens in a special case when random effects or random coefficients are introduced in logit–linear or probit–linear models for binary data. But just to make things concrete, we refer to the following marketing context. For each of C customers labeled $c = 1, 2, \ldots, C$, a number n_c of purchases of a certain consumers good are recorded. Our responses are

$$y_{ci} = \begin{cases} 1 \text{ if ``our brand'' was preferred,} \\ 0 \text{ if some other brand was preferred,} \end{cases}$$

where $i = 1, 2, \ldots, n_c$ labels the purchases for each customer. In addition, we have a number K of explanatory variables

$$x_{kci}, \;\; c = 1, 2\,, \ldots, C\,, \;\; i = 1, 2, \ldots, n_c\,, \;\; k = 1, 2, \ldots, K\,.$$

Such variables could be the price (or log price, or log price relative to competitors average price) at the time of the purchase, the distance (or inverse distance) to the dealer, dummies for the type of dealer (super market, smaller shop etc.), the customer's exposure to advertisments for our brand at the time of the purchase, demographic variables associated with the customer etc. etc. Some of these variables can have the form x_{kc} (beeing associated only with the customer) whereas others may take the form x_{kci} (beeing associated with the single purchase).

In the following, we consider the case where only one explanatory variable x_{ci} is present. This is just to simplify the notation. The generalization to more than one explanatory variable makes the notation heavier, but conceptually it does not change much.

2. The logit– and probit–linear models.

A standard logistic regression model for this situation would state that

$$P\left(Y_{ci} = 1\right) = \Lambda\left(\alpha_c + \beta x_{ci}\right)$$

where α_c, $c = 1, 2, \ldots, C$ are parameters associated with the customers, β is the parameter determining the effect of the covariate x, and

$$\Lambda(\eta) = \frac{e^{\eta}}{1 + e^{\eta}}$$

is the standard choice of the inverse link function (cfr. McCullagh and Nelder (1989)), the function that transforms the real axis (the natural domain for linear expressions like $\alpha_c + \beta x_{ci}$) to the unit interval (the natural domain for probabilities).

The function Λ can be thought of as a c.d.f. for a continuous distribution, namely the socalled logistic distribution. Accordingly, a logistic regression model has an interpretation in terms of an underlying linear regression model for "latent observations" which goes as follows.

Let L_{ci} be independent normalized logistic variables associated with the purchases. Then we have

$$P(Y_{ci} = 1) = P(L_{ci} \leq \alpha_c + \beta x_{ci})$$

or

$$P(Y_{ci} = 1) = P(0 \leq \alpha_c + \beta x_{ci} - L_{ci}) = P(Y^*_{ci} \geq 0)$$

where $Y^*_{ci} = \alpha_c + \beta x_{ci} - L_{ci}$ are the underlying latent observations. If we could observe the Y^*_{ci}, the relevant model would be an ordinary linear regression model (apart from the somewhat unusual choice of a logistic error distribution rather than a normal, and the fact that the scale parameter for the error term is known and equal to 1). However, what we observe is only the indicators Y_{ci} for the events $Y^*_{ci} \geq 0$. This interpretation of the logistic regression model as a linear regression model combined with an incomplete observation scheme plays a crucial role in the following.

The probit–linear model differs from the logit–linear model only in the choice of link function. In our case, the probit model states that

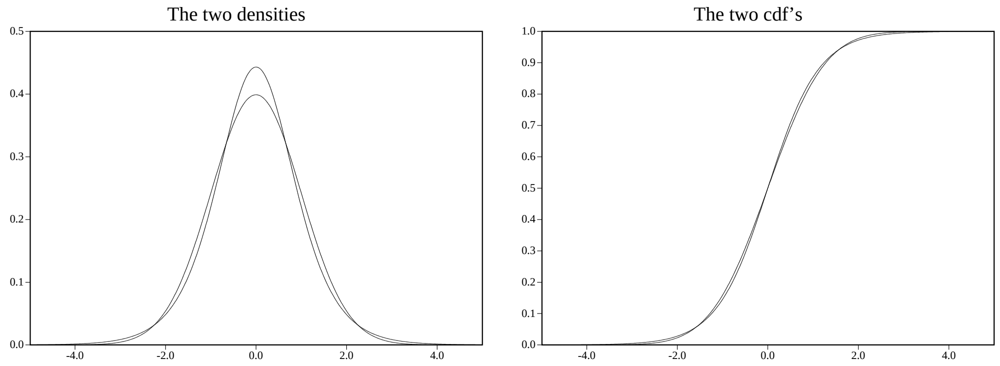
$$P(Y_{ci} = 1) = \Phi(\alpha_c + \beta x_{ci})$$

where

$$\Phi(\eta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\eta} \exp\left(-\frac{z^2}{2}\right) dz$$

is the c.d.f. of the normalized normal distribution. Accordingly, the probit–linear model has an interpretation in terms of an "underlying linear regression model for latent variables" $Y^*_{ci} = \alpha_c + \beta x_{ci} - U_{ci}$ with error terms U_{ci} that are normalized normal. Again, we do not observe the latent variables, only the indicators Y_{ci} of the events $Y^*_{ci} \geq 0$.

In practice it is difficult to distinguish the two models from each other, because the underlying distributions are so very similar. The figure below shows their densities and c.d.f.'s, after suitable normalization of the logistic distribution to ensure comparability. The logistic distribution has variance π, and is therefore rescaled by a scale parameter of $\frac{1}{\sqrt{\pi}}$.



Notice that parameter estimates from logit and probit models are not directly comparable without a similar "$\sqrt{\pi}$–correction".

The logit density (after normalization) is the one with the highest peak, and accordingly the logit c.d.f. is the steeper of the two at the point (0,0.5).

Nowadays, the probit model is rarely used because the algebraic structure of the logit model is much simpler, beeing an exponential family with nice relations to log–linear models for count data and so on. Earlier, probit models were more commonly preferred in cases where the latent variables were believed to have a concrete interpretation, and where the standard choice of a normal distribution for these variables was considered the most natural. However, the choice between the two types is mainly a matter of taste, only for very large data sets is it possible to make a statistical distinction between them.

3. Models with random effects.

The problem with the simple logit– and probit–models in our marketing setup is that they can not measure the effect of covariates that are associated with customers. If the covariate x_{ci} takes the simpler form x_c, the model becomes overparametrized because any change of the coefficient β can be compensated by a change of the customer parameters α_c in such a way that the model fits the (in this case binomial) customer totals $y_{c.}$ exactly. Moreover, the effects of purchase dependent covariates x_{ci} are only measured by the influence they have on change of a customer's behaviour, not by the influence they may have due to different (average) levels for different customers.

It is tempting to solve this problem by use of a model where the individual customer parameters α_c are replaced with a common intercept α, but this can produce very misleading results if the customers are actually different (see Birch and Tjur (forthcoming), Tjur 2002). A conceptually simple (though computationally rather complicated) solution is to think of the customer parameters as drawn from a normal population, that is to let

$$\alpha_c = \alpha + \omega V_c$$

where the V_c, $c = 1, 2, \ldots, C$, are i.i.d. normalized normal. Or, equivalently, the parameters α_c are assumed to be drawn from a normal distribution with mean α and variance ω^2.

In the case where each customer is observed only once ($n_c = 1$), the probit model becomes particularly simple. With an obvious simplification of notation (omitting index i, which now runs from 1 to 1) we get by the latent variables interpretation

$$P(Y_c = 1) = P(\alpha_c + \beta x_c - U_c \geq 0) = P(\alpha + \omega V_c + \beta x_c - U_c \geq 0)$$

$$= P(U_c - \omega V_c \leq \alpha + \beta x_c)$$

which (since we must, of course, assume that the original error terms U_c of the latent model are independent of the random customer parameters $\alpha + \omega V_c$) by the convolution property of the normal distribution equals

$$P\left(Z_c\sqrt{\omega^2+1} \le \alpha + \beta x_c\right) = \Phi\left(\frac{\alpha+\beta x_c}{\sqrt{\omega^2+1}}\right) = \Phi\left(\alpha' + \beta' x_c\right)$$

where $Z_c = \frac{U_c - \omega V_c}{\sqrt{\omega^2+1}}$ is a new set of i.i.d. normalized normal variables. The conclusion is that the model with random customer parameters is equivalent to an ordinary probit–linear model with a common intercept α' and a coefficient β' to x given by

$$\alpha' = \frac{\alpha}{\sqrt{\omega^2+1}} \quad \text{and} \quad \beta' = \frac{\beta}{\sqrt{\omega^2+1}}$$

It follows that the parameters α and β are non–identifiable, only their "randomness–adjusted" versions α' and β' can be estimated. Intuitively, the reason for this is that the only effect of the randomness of the customer parameters is to blur the differences between customers. Without replicates within customers, we cannot distinguish between a weak influence of the covariate and a high variation between customers. It is possible to test the hypothesis $\beta = 0$, since it is equivalent to $\beta' = 0$ in the derived model; and since $|\beta'| \le |\beta|$, it is also possible in some cases to give a positive lower confidence bound or a negative upper confidence bound for β. But estimation of α and β in the usual sense is impossible.

For the logistic regression model with all $n_c = 1$, things are more complicated. From a purely mathematical point of view, we can not say much. Doing the same trick as above with the logit model will result in a model where the variables Z_c are linear combinations of a normal and a logistic variable, and since the two scale parameters can in principle be identified from the shape of the convolved c.d.f., it would be mathematically correct to say that the three parameters α, β and ω^2 can be identified in this model. But the close similarity of the two distributions makes such an identification impossible in practice for moderate data sizes. Even if millions of observations were present and supported the choice of a logistic response curve in favour of a probit curve, it would be somewhat irresponsable to attach any weight to point estimates of α and β that are so sensitive to microscopic changes of the link function and the underlying distribution of the random customer parameters.

4. Models with random coefficients.

An extension of the random effects model, which can be used if also the slope β is believed to vary from customer to customer, goes as follows. Assume that customer intercepts α_c *and* customer specific slopes β_c are drawn from a two–dimensional normal distribution with mean (α, β),

$\mathrm{var}(\alpha_c) = \omega^2$, $\mathrm{var}(\beta_c) = \delta^2$ and $\mathrm{cov}(\alpha_c, \beta_c) = \rho\omega\delta$. For given values of α_c and β_c, we assume that

$$P(Y_{ci} = 1) = \Lambda(\alpha_c + \beta_c x_{ci})$$

In the probit–version of this model with all $n_c = 1$ we get, by the latent variable interpretation (since the variance of the random contribution $U_c - (\alpha_c - \alpha) - (\beta_c - \beta)x_c$ now becomes $\omega^2 + \delta^2 x_c^2 + 2\omega\delta\rho x_c + 1$)

$$\begin{aligned} P(Y_c = 1) &= P(U_c \leq \alpha_c + \beta_c x_c) \\ &= P(U_c - (\alpha_c - \alpha) - (\beta_c - \beta)x_c \leq \alpha + \beta x_c) \\ &= \Phi\left(\frac{\alpha + \beta x_c}{\sqrt{\omega^2 + \delta^2 x_c^2 + 2\omega\delta\rho x_c + 1}}\right). \end{aligned}$$

Also for this model we can show that α and β are non–identifiable. For some constant $k > 1$, rewrite the argument to Φ above as follows.

$$\begin{aligned} \frac{\alpha + \beta x_c}{\sqrt{\omega^2 + \delta^2 x_c^2 + 2\omega\delta\rho x_c + 1}} &= \frac{k\alpha + k\beta x_c}{\sqrt{k^2(\omega^2 + 1) + k^2\delta^2 x_c^2 + 2k^2\omega\delta\rho x_c}} \\ &= \frac{\alpha' + \beta' x_c}{\sqrt{\omega'^2 + \delta'^2 x_c^2 + 2\omega'\delta'\rho' x_c + 1}} \end{aligned}$$

where

$$\alpha' = k\alpha \ , \ \ \beta' = k\beta \ , \ \ \delta' = k\delta$$

$$\omega' = \sqrt{k^2(\omega^2 + 1) - 1} \quad \text{and} \quad \rho' = \frac{k\omega\rho}{\sqrt{k^2(\omega^2 + 1) - 1}}.$$

This means that the replacement of α with $k\alpha$ and β with $k\beta$ can be compensated by a change of the variance/covariance parameters in such a way that the distribution of the data set remains unchanged. Thus, it does not make any sense to talk about estimation of α and β in this case either.

There are, of course, parameter functions that can be estimated. Among these we can mention $\alpha/\sqrt{1 + \omega^2}$, β/δ (the values of Φ's argument for $x = 0$ and $x \to \infty$, respectively) and $-\alpha/\beta$ (the value of x for which the success probability is $\frac{1}{2}$). But it is difficult to see why one should be interested in estimates of these quantities.

In addition to this, the model has some peculiar properties which would make it difficult to interpret the parameters α and β even if we could estimate them. In the special case $\beta = \rho = 0$, for example, the expression for the success probability takes the form

$$P(Y_c = 1) = \Phi\left(\frac{\alpha}{\sqrt{\omega^2 + 1 + \delta^2 x_c^2}}\right).$$

Thus, $\beta = 0$ does *not* imply that the success probability is independent of the covariate. The probability depends on x_c in a rather unusual way. As we can easily see, it converges to $\frac{1}{2}$ (!) as $x_c \to \pm\infty$, and for $x_c = 0$ it takes its maximum or minimum (depending on whether α is positive or negative). The dependence is not even monotone when $\alpha \neq 0$.

Furthermore, the model has a very confusing property, which we illustrate only for the case where $\rho = 0$ because the algebra becomes slightly simpler in this case. Provided that $\delta^2 > 1$ (which, by the way, can always be obtained by the above mentioned transformation by a factor k) and that all values of x_c are positive, we can rewrite the argument to Φ as follows, dividing by x_c in nominator and denominator.

$$\frac{\alpha + \beta x_c}{\sqrt{\omega^2 + 1 + \delta^2 x_c^2}} = \frac{\alpha \frac{1}{x_c} + \beta}{\sqrt{(\omega^2+1)\left(\frac{1}{x_c}\right)^2 + \delta^2}} = \frac{\alpha' + \beta' x'_c}{\sqrt{\omega'^2 + 1 + \delta'^2 x'^2_c}}$$

where

$$\alpha' = \beta \ , \ \beta' = \alpha \ , \ x'_c = \frac{1}{x_c}$$
$$\omega' = \sqrt{\delta^2 - 1} \quad \text{and} \quad \delta' = \sqrt{\omega^2 + 1}\,.$$

This means that if all values of the covariate are replaced by their inverses, while intercept and slope are interchanged (!) and some compensating changes of the two variances are made, we end up with exactly the same distribution of data as before. We can even do this in the case where α and β have the same sign. In this case a seemingly positive (or negative) effect of the original covariate turns into a seemingly positive (or negative) effect of its inverse. This is just algebra, of course, but it means that the usual interpretation of the sign of the regression coefficient β breaks down.

Our conclusion from all this is, of course, a warning against these models in the case where all $n_c = 1$. If they can measure anything at all it is a useless mess of irrelevant functions of the variance/covariance parameters and the parameters α and β that were originally of interest. For the probit models, this follows from what was said above. For the logit models it follows from the fact that they are, in practice, indistinguishable from the probit models.

The problem which originally motivated this work is that some statistics packages can handle these models and "estimate" in them even when all the n_c are 1. Markov Chain Monte Carlo and other formally Bayesian methods have the property that they can produce estimates of non–identifiable or approximately non–identifiable parameters without really discovering it, because the prior distribution in itself holds "information" about these parameters. Be careful when estimates are very sensitive to the choice of prior, even when the priors are very flat.

5. But what do we do then?

The conclusion of all this may appear somewhat destructive. Not only do we reject the possibility of using random effects and random coefficients models in the case where customers are only observed once. The final remarks of the previous section also indicate that the interpretation of the parameters in the random coefficients model can be quite complicated, even when several purchases are observed for each customer.

Our final (and more positive) remark is that the simplest of all logistic regression models, the model

$$P(Y_c = 1) = \Lambda(\alpha + \beta x_c)$$

with only two parameters (or $1 + K$ in case of K covariates) is actually a perfectly respectable and very useful model in the case $n_c = 1$, if it is interpreted correctly. This may appear a bit confusing at the moment. But the reason is that our discussion up to this point has been based on the interpretation of the basic event as "customer c performs his or hers i'th purchase". The correct interpretation of the simple model above should be in terms of basic events of the form "a random customer performs a purchase". In this interpretation, the two sources of randomness (customers are different, and even the same customer may behave differently on different occasions) are pooled. Quite often, this description of customer behaviour is exactly what is needed in the operational use of these models for price setting, advertising decisions etc., where the distinction between the two sources of variation is unimportant. Moreover, as we have seen in section 3, this model actually comes out (exactly in the probit case, approximately in the logit case) as the correct model when a random customer effect is assumed. The fact that the parameters α and β are actually $\alpha/\sqrt{\omega^2+1}$ and $\beta/\sqrt{\omega^2+1}$ from the underlying random effects model need not bother us, when the simple logistic regression model is interpreted in this way.

The drawback is that this will only work if all purchases (or, in practice, almost all purchases) are made by different customers. In principle, the model will produce incorrect results (too narrow confidence intervals for coefficients, too many false significances) when there is customer heterogeneity and some of the n_c are > 1. Other things must be done then, and this is where the random effects models and the random coefficients models come in. See also Birch and Tjur (forthcoming) and Tjur (2002) for a more primitive and easy solution.

Thanks to Jørgen Kai Olsen for drawing my attention to this problem during several interesting discussions.

References.

Birch, K., and Tjur, T. (forthcoming)
Logistic models for single–source data.

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

Olsen, J. K. (2003)
Maksimum likelihood estimation af parametrene i logitmodellen med stokastiske individparametre — et simulationsstudie.
Institut for Afsætningsøkonomi, Handelshøjskolen i København.

Tjur, T. (2002)
Logistic regression models for single–source data — a simulation study
Preprint no. 4, Dept. of Management Science and Statistics, Copenhagen Business School.
(available as http://www.cbs.dk/staff/tuetjur/02-4.pdf)

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