Predictive Performance of the Binary Logit Model in Unbalanced Samples

J.S. Cramer*

July 1998

Subject to revision. Do not quote without informing author

Abstract

In a binary logit analysis with unequal sample frequencies of the two outcomes the less frequent outcome always has lower estimated prediction probabilities than the other one. This effect is unavoidable, and its extent varies inversely with the fit of the model, as given by a new measure that follows naturally from the argument. Unbalanced samples with a poor fit are typical for survey analyses of the social sciences and epidemiology, and there the difference in prediction probabilities is most acute. It affects two common diagnostics, the within-sample 'percentage correctly predicted' and the identification of outliers. Partial remedies are suggested.

*Affiliation: Tinbergen Institute, Amsterdam; address: Baambrugse Zuwe 194, 3645 AM Vinkeveen, the Netherlands, e-mail *marsa@axxel.nl.* I thank Bas van der Klaauw, Ruud Koning, Francois Laisney, Geert Ridder, Jan Sandee and Frank Windmeijer for helpful comments on earlier versions and Bas van der Klaauw for material assistance with the computation of the PART results. I am much indebted to many authors who kindly provided me with their data sets and results. Frank Bakker, Ronald Brand, Jaap Dronkers, Guus Hart, Francois Laisney, Dinand Webbink and Karin Wittebrood took the trouble to prepare diskettes and answer my questions. I also thank their co-authors who acquiesced in the use of their data and results: H. Bartelink, J. Borger, Ann R. Braun, Anne Davies, J. van Dongen, Siegfried Gabler, Janet Jackson, H. Kemperman, Martha E. Klein, J. Lebesque, Michael Lechner, Nora C.Mesa, Hessel Oosterbeek and H. Sillevis Smitt.

1 Preliminaries

The setting of this paper is a standard binary logit regression that has been estimated by Maximum Likelihood (ML). Leaving the parameter estimates $\hat{\theta}$ aside we at once proceed to the estimated within-sample probabilities $\hat{P}_i$ of the outcome $Y_i = 1$. These probabilities are arranged in a vector $\hat{p}$, with complement $\hat{q}$; the outcomes are likewise recorded in y, with complement vector z. The *crude residuals* are defined as

$$e_i = Y_i - \hat{P}_i \tag{1}$$

or the vector $e = y - \hat{p}$. The vector of complement residuals $z - \hat{q}$ is identical with reverse sign.

The sample consists of n observations, m with $Y_i = 1$ and $n - m$ with $Y_i = 0$, or $Z_i = 1$; the relative shares are α and $1 - \alpha$. Whenever the two sample shares are unequal, α is by convention the larger share and the corresponding outcome is labelled $Y_i = 1$.

The regressor matrix of the *full* model is X, and the ML estimates $\hat{p}$ of the logit model satisfy

$$X^T(y - \hat{p}) = X^T e = 0. \tag{2}$$

X is always taken to include a unit constant, so that in particular

$$i^T(y - \hat{p}) = i^T e = 0, \tag{3}$$

or, in other terms,

$$\bar{P} = \alpha, \tag{4}$$

where $\bar{P}$ is the overall mean of the elements of $\hat{p}$. This property of the estimated probabilities will be called *equality of the means*.

In addition to $\hat{P}_i$ and $\hat{Q}_i$ we shall make use of the estimated probability of the observed outcome $Pr(i)$

$$Pr(i) = Y_i\hat{P}_i + Z_i\hat{Q}_i. \tag{5}$$

Note that the maximum of the loglikelihood function is

$$\log \hat{L} = \log L(\hat{\theta}) = \sum \log Pr(i). \tag{6}$$

The *null model* has the unit constant as the sole regressor; it is nested in the full model with richer X. In this model $\hat{P}_i$ and $\hat{Q}_i$ are constant and equal to α and $1-\alpha$ respectively, with loglikelihood

$$logL_0 = m\log\alpha + (n-m)\log(1-\alpha). \tag{7}$$

This is the lower limit of $\log\hat{L}$ of (6). On average, therefore, the $Pr(i)$ are at least equal to their null values of α for $Y_i = 1$ and of $1-\alpha$ for $Y_i = 0$, but it is of course hoped that they are substantially higher. This leads us to consider the ratio of $Pr(i)$ to its null value

$$Pf(i) = Y_i(\hat{P}_i/\alpha) + Z_i(\hat{Q}_i/(1-\alpha)). \tag{8}$$

$Pf(i)$ reflects the improvement of the full model over the null model in predicting the i'th outcome; it is an *index of performance* for that particular observation. It is not a probability; it is nonnegative, and its overall level or average should exceed 1. Upon taking logarithms and summing we find

$$\sum \log Pf(i) = \log\hat{L} - \log L_0. \tag{9}$$

Doubling this gives LR, the common likelihood ratio statistic for the significance of the full model,

$$2\sum \log Pf(i) = LR. \tag{10}$$

The geometric mean of the $Pf(i)$ is

$$\widetilde{Pf} = \exp(LR/2n). \tag{11}$$

LR is nonnegative and $\widetilde{Pf}$ is never smaller than 1.

2 Prediction Probabilities in Unequal Sample Shares

In most survey data of the social sciences and epidemiology the sample shares of the two outcomes are unequal, and values of α of .7 or .8 are much more common than equal parts. Upon fitting a logit model it is then invariably found that the estimated prediction probabilities $Pr(i)$ are quite high for $Y_i = 1$, the outcome with the greater share, and very low for the outcome with the lesser share[1]. If we distinguish two subsets among the $\hat{P}_i$, with $\hat{P}_i^+$

[1]Inequality of sample proportions of the outcomes thus by itself leads to a high overall level of Pr(i) and to high loglikelihoods.

for $Y_i = 1$ and $\hat{P}_i^-$ for $Y_i = 0$, and likewise for $\hat{Q}_i$, the $\hat{P}_i^+$ have a much higher overall level than the $\hat{Q}_i^+$. This asymmetry in the prediction of $Y_i = 1$ and $Y_i = 0$ is well known to practitioners. Yet there is no clear reason why a rare outcome should be badly predicted; a good prediction must be simply a matter of choosing the right regressors. This is indeed so, and even quite rare outcomes can in principle have estimated probabilities all the way up to 1; but whatever value they attain, on average the other, prevalent outcome will always be predicted even better. The *extent* of this systematic difference varies with the fit of the model; and since outside controlled experiments the fit is usually mediocre, a great contrast between the poor prediction of rare states and the good prediction of prevalent states is the rule.

The argument that establishes this result is somewhat unusual but really almost trivial. Consider the averages of $\hat{P}_i$ for the two subsets of observations with $Y_i = 1$ and $Y_i = 0$ already mentioned. The first, which refers to the outcome with the larger share, is

$$\bar{P}^+ = \hat{p}^T y/m \tag{12}$$

and the other

$$\bar{P}^- = \hat{p}^T z/(n-m). \tag{13}$$

The overall mean $\bar{P}$ is their weighted average, or, with (4),

$$\alpha\bar{P}^+ + (1-\alpha)\bar{P}^- = \bar{P} = \alpha. \tag{14}$$

If the fitted model has any explanatory power, $\bar{P}^+$ exceeds $\bar{P}^-$, and the two will lie on either side of their (weighted) average α. Since they are mean probabilities, they are both constrained to the interval $(0,1)$; $\bar{P}^-$ lies in $(0,\alpha]$, and $\bar{P}^+$ in $[\alpha,1)$. This suggests writing $\bar{P}^+$ as a linear combination of α and 1 with nonnegative weights $(1-\lambda)$ and λ, or

$$\bar{P}^+ = (1-\lambda)\alpha + \lambda = \alpha + \lambda(1-\alpha). \tag{15}$$

By (14) this gives

$$\bar{P}^- = \alpha(1-\lambda) \tag{16}$$

so that $\bar{P}^-$ is a linear combination of 0 and α with the same weights λ and $1-\lambda$.

Similar expressions hold for the $\hat{Q}_i$. The subset means are

$$\bar{Q}^+ = \hat{q}^T z/(n-m) \tag{17}$$

and

$$\bar{Q}^- = \hat{q}^T y/m. \tag{18}$$

Since $\hat{Q}_i$ is the complement of $\hat{P}_i$ we have

$$\bar{P}^+ + \bar{Q}^- = 1, \quad \bar{P}^- + \bar{Q}^+ = 1.$$

and this gives

$$\bar{Q}^+ = (1-\alpha) + \lambda\alpha \tag{19}$$

and

$$\bar{Q}^- = (1-\lambda)(1-\alpha) \tag{20}$$

$\bar{Q}^+$ and $\bar{Q}^-$ are linear combinations like $\bar{P}^+$ and $\bar{P}^-$ with the same weights λ and $(1-\lambda)$.

The upshot is that all four means are determined by two parameters, the share α and the weight λ, with

$$0 < \alpha < 1, \quad 0 \leq \lambda < 1. \tag{21}$$

The limits of α are self-evident; as for λ, it is zero for the null model with

$$\bar{P}^+ = \bar{P}^- = \bar{P} = \alpha,$$

but it can not attain its upper bound since this would imply $\bar{P}^+ = 1$, $\bar{Q}^+ = 1$, and such perfect prediction is beyond logit probabilities or their estimates[2].

Figure 1 shows how $\bar{P}^+$ and $\bar{Q}^+$ vary with λ for a given α and with α for a given λ. In the first panel α is .8, which is a quite common value. Provided λ is high enough, $\bar{Q}^+$ can reach quite high values, but $\bar{P}^+$ will always be even higher. In the second panel we need only look at the right-hand half since outcomes have been so labelled that $\alpha \geq .5$. Here $\bar{P}^+$ again always surpasses $\bar{Q}^+$, and as α increases towards 1 the one goes up and the other goes down. This effect is the more marked the lower λ: here it is only .2.

[2]Negative λ can be ruled out as this would mean that the loglikelihood of the full model is less than that of the null model

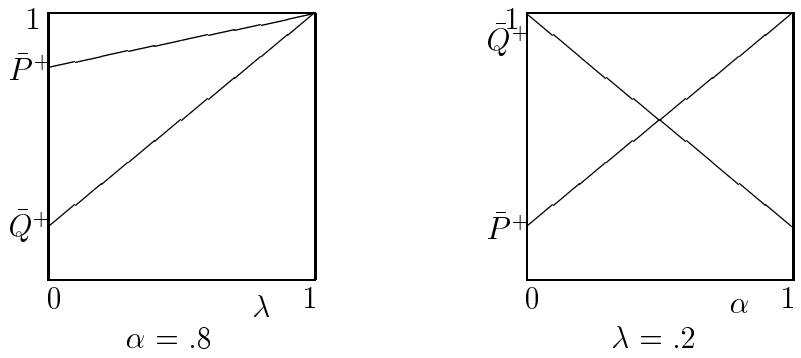

Figure 1. Behaviour of $\bar{P}^+$ and $\bar{Q}^+$ with α and λ.

Combining (15) and (19) as

$$\bar{P}^+ - \bar{Q}^+ = 2(\alpha - .5)(1 - \lambda) \tag{22}$$

we have the answer to the initial question why the level of predicted probabilities varies with the sample share. Unless $\alpha = .5$, $\bar{P}^+$ exceeds $\bar{Q}^+$, and this excess varies inversely with λ. Large values of λ would therefore limit its extent; but in practice this is of little help, as λ is usually quite small. In the the illustrative nonexperimental studies of the next section it ranges between .07 and .33.

In these conditions estimated probabilities are a poor measure of within-sample predictive performance: they would lead to the absurd conclusion that success is predicted very well while failure is predicted badly, as if one can at the same time predict survival with precision but death not at all.

From (15) and (16), (19) and (20) we also find

$$\bar{P}^+ - \bar{P}^- = \bar{Q}^+ - \bar{Q}^- = \lambda. \tag{23}$$

λ can therefore be seen as a crude measure of fit since it indicates the discrimination of $\hat{P}_i$ (and of $\hat{Q}_i$) between the two observed outcomes. This interpretation is further explored in section 4.

3 Illustrations

We illustrate these arguments by eight studies from the social sciences and epidemiology. They are listed by increasing α in Table 1, with their abbreviated name, subject, sample size n and number of regressors K. Four deal

with economic, educational and medical issues in Holland; one refers to the French labour market; one is a study of British rapists; and two have been taken from American textbooks of biostatistics. Sample size n and the number of regressors (including the intercept) K vary widely. Section 7 provides thumbnail sketches with further details of the analyses and their adaptation to present purposes. All studies have been published in respectable journals or are otherwise available.

Name	Subject	α	n	K
PART	labour market participation of married women in France, 1979. Gabler et al. (1993)	.520	3658	21
CAR	private car ownership of Dutch households, 1980. Cramer (1991)	.642	2820	6
FIBRO	fibrosis after breast cancer surgery, Holland, 1979-88. Borger et al. (1994)	.720	332	12
EDUA	performance of Dutch schoolchildren, 1965. Dronkers (1993)	.777	699	8
ICU	intensive care unit performance, Massachusetts, 1983. Lemeshow et al. (1988)	.800	200	9
EDUB	educational choice of Dutch schoolchildren, 1982. Oosterbeek and Webbink (1995)	.802	1706	12
DEPRI	depression in Los Angeles, 1979. Afifi and Clark (1990)	.830	294	6
RAPE	antecedents of rapists in Great Britain, 1965-93. Davies et al. (1997)	.843	210	13

Table 1. Illustrative binary logit studies

In six out of eight samples α lies between .70 and .85; only two samples are more evenly balanced. A cursory inspection of the literature suggests that among surveys of these types unbalanced samples of this order are the

rule, and a fair 50/50 division is an exception. One can easily find examples with a much skewer distribution of the outcomes than reported here.

Name	α	$\bar{P}^+$	$\bar{Q}^+$	diff	λ
PART	.520	.627	.596	.031	.223
CAR	.642	.759	.568	.191	.327
FIBRO	.720	.793	.467	.326	.260
EDUA	.777	.823	.385	.438	.208
ICU	.800	.872	.487	.385	.359
EDUB	.802	.816	.256	.560	.072
DEPRI	.830	.838	.208	.630	.046
RAPE	.843	.872	.312	.560	.183

Table 2. Prediction probabilities for eight examples

In Table 2 we repeat α and show the two subset means $\bar{P}^+$ and $\bar{Q}^+$, their difference, and the value of λ from (23). It is quite clear that with unequal sample shares the less frequent outcome systematically has a (much) lower average prediction probability than the other. This is equally due to low values of λ as to high values of α, as by (19) $\bar{Q}^+$ must exceed both $(1-\alpha)$ and λ. As the table shows, λ varies widely between the various studies. The highest value is is .36 for ICU, based on a nonrandom subset from a larger sample, selected for textbook use. The lowest value occurs for DEPRI, where all six regressors turn out to be rather crude categorical variables. The values between .07 and .33 for the other samples seem to set the norm for survey studies.

All the illustrative studies are from epidemiology and the social sciences. Much more extreme cases of rare outcomes can be found in marketing and in financial analyses, such as the response to large indiscriminate direct mail campaigns or the incidence of take-over bids or bankruptcies. In contrast, the

problem hardly arises in controlled experiments like the classic bio-assay of the effect of pesticides on strictly homogeneous batches of organisms. Most samples are about equally balanced as a matter of design, and the analyses have a substantially better fit. In a handful of such controlled experiments we found values of λ between .4 and .7[3].

4 λ as a measure of fit

In (15) λ was introduced as a weight that pulls up $\bar{P}^+$ from its null model value (for zero λ) towards the unattainable limit $\bar{P}^+ = 1$ of perfect prediction (for unit λ). By (23) it is the difference between $\bar{P}^+$ and $\bar{P}^-$. Both interpretations suggest that it is a measure of fit, whether the sample shares are unequal or not.

This suggestion is strengthened by the relation of λ to the index of prediction performance $Pf(i)$ of (8). Its arithmetic mean over the entire sample is

$$\overline{Pf} = 1/n \sum Pf(i) = \alpha(\bar{P}^+/\alpha) + (1-\alpha)(\bar{Q}^+/(1-\alpha)), \tag{24}$$

or, by (15) and (19),

$$\overline{Pf} = 1 + \lambda. \tag{25}$$

There is more. In ordinary least squares regression with a continuous regressand c the residuals $e^c = c - \hat{c}$ satisfy

$$\imath^T e^c = 0, \quad c^T e^c = 0.$$

Similar properties hold for the crude residuals (1) of the logit model. First, as in (3),

$$\imath^T e = 0;$$

as for the second property, it can be shown that

$$\hat{p}^T e/n \xrightarrow{p} 0. \tag{26}$$

[3]We used the data of two examples from Finney(1947), viz. the case of *vasoconstriction* of Gilliatt (1947) (Finney p.184) and of *Tribolium castaneum* of Hewlett (1969) (Finney p.260) as well as six analyses of mite eggs from Bakker et al (1993).

This holds under quite general conditions for *any* consistent estimate $\hat{p}$ of p for the $(0,1)$ outcomes y of *any* discrete model (see Cramer (1997)). For finite samples it is of course only an approximation

$$\hat{p}^T e/n \approx 0 \tag{27}$$

or equivalently

$$\hat{p}^T y/n \approx \hat{p}^T \hat{p}/n \tag{28}$$

and likewise for $\hat{q}$ and z. We shall refer to this as the *orthogonality* property of $\hat{p}$ - orthogonality to the residuals.

Name	$R_{\hat{p},e}$
PART	.0021
CAR	.0090
FIBRO	-.0135
EDUA	-.0192
ICU	.0096
EDUB	.0008
DEPRI	-.0003
RAPE	.0003

Table 3. Correlation of $\hat{p}$ and e.

By (28) $\hat{p}$ and e are approximately uncorrelated, and Table 3 bears out that for the present illustrations the correlations are indeed quite close to zero. Inspection shows that $\hat{p}$ comes near to a linear combination of X, which by (2) is orthogonal to e. This need not always be so, but the correlation of $\hat{p}$ and e is easily established in any particular instance.

The common decomposition of the sum of squares of $\hat{P}_i$ in (27) gives

$$\hat{p}^T y/n \approx \hat{p}^T \hat{p}/n = \alpha^2 + \sigma_p^2 \tag{29}$$

where σ_p^2 is the variance of the column $\hat{p}$ over its entire length. Substitution in (12) yields

$$\bar{P}^+ \approx \alpha + \sigma_p^2/\alpha, \tag{30}$$

and equating this to (15) gives

$$\lambda \approx \frac{\sigma_p^2}{\alpha(1-\alpha)}. \tag{31}$$

For $\lambda = 1$ or perfect prediction σ_p^2 attains its maximum of $\alpha(1-\alpha)$; $\lambda = 0$ implies $\sigma_p^2 = 0$, or constant $\hat{P}_i$, which is the null model.

Since $\hat{P}_i$ and e_i are (nearly) uncorrelated we have the familiar decomposition of the sum of squares of Y_i

$$SS_y \approx SS_e + SS_p. \tag{32}$$

Clearly,

$$SS_p = n\sigma_p^2, \quad SS_y = n\alpha(1-\alpha)$$

so that by (31)

$$\lambda \approx SS_p/SS_y = 1 - SS_e/SS_y. \tag{33}$$

Thus λ is supported by (32), which is a straightforward analysis of variance of y as advocated by Efron (1978). Efron strongly stresses the need for a Pythagorean relation which permits a simple additive decomposition; (32) satisfies this, if only approximately (or asymptotically). By (33), λ resembles R^2 since it indicates the proportion of the total dependent variation that has been 'explained'.

It is tempting to use (33) to construct a F statistic for the overall significance of the full model, but there is no need for this. There is a perfectly good Likelihood Ratio test of this issue, and by (11) and (25) both LR and λ are based on the same $Pf(i)$, although the one uses the geometric mean and the other the arithmetic mean. Hence λ does not add any new information. The only merit of F would be to show that quite low values of λ are compatible with a significant relation, provided the sample is large enough - just as in the case of R^2. This is the usual situation in the analysis of survey data in the social sciences.

To sum up, λ varies between zero for the null model and 1 for perfect prediction; it reflects the differences between $\bar{P}^+$ and $\bar{P}^-$; it measures the proportion of the total variation of y that is 'explained'; and it turns up in various other measures and decompositions. In short, it uncommonly resembles R^2 of linear regression. Like R^2, however, it is merely a descriptive measure with immediate intuitive appeal rather than a proper statistic with a known distribution.

5 Unfortunate effects for unbalanced samples

We return to unbalanced samples with widely different levels of $\hat{P}_i^+$ and $\hat{Q}_i^+$ and examine the effect on two common diagnostics.

Percentage correctly predicted

Many statistical computer packages show the *percentage correctly predicted* in the sample. Estimated $0, 1$ attributes $\hat{Y}_i$ are assigned to the observations according to whichever is the greater of $\hat{P}_i$ and $\hat{Q}_i$, or

if $\hat{P}_i \geq .5$, $\hat{Y}_i = 1$:
if $\hat{Q}_i > .5$, $\hat{Y}_i = 0$.

This criterion or 'cut-off' point of .5 is optimal if either form of misclassification carries the same loss. The $\hat{Y}_i$ are then set off against the observed values in a 2 by 2 table, shown for EDUB as Table 4.

	$Y_i = 1$	$Y_i = 0$	total
$\hat{Y}_i = 1$	1358	334	1692
$\hat{Y}_i = 0$	10	4	14
total	1368	338	1706

Table 4. Predicted and observed states for EDUB
(cut-off point .5)

In this case the number of correct predictions is $1358 + 4 = 1362$, and the success rate would be blithely reported as $1363/1706 = 79.8\%$. But this result reflects the composition of the sample rather than the performance of the model; it is due to an α of .8, coupled with a poor fit ($\lambda = .072$). This leads to a low level of the $\hat{Q}_i^+$ and thereby to a high *overall* level of $\hat{P}_i$, so that for all but 14 observations the prediction is $\hat{Y}_i = 1$, while of course 80% of all observations actually have $Y_i = 1$. Underneath, the scores for the two outcomes are very different: the success rate is $1358/1368 = .99$ for $Y_i = 1$ but only $4/338 = .01$ for $Y_i = 0$.

This incongruous result is linked to the 'cut-off' point of .5. An alternative is a prediction that is optimal in the sense that, for given $\hat{P}_i$, it maximizes $Pf(i)$ of (8), and hence the fit of $\hat{y}$ to the given $\hat{p}$. This is achieved by a 'cut-off' point of α, or

if $\hat{P}_i \geq \alpha$, $\hat{Y}_i = 1$:
if $\hat{Q}_i > (1 - \alpha)$, $\hat{Y}_i = 0$.

Table 5 shows that with this procedure the overall success rate drops to .63, but that it is much more evenly spread over the two alternatives: it is now .62 for $Y_i = 1$ and .68 for $Y_i = 0$. This is a more sensible result.

	$Y_i = 1$	$Y_i = 0$	total
$\hat{Y}_i = 1$	841	110	951
$\hat{Y}_i = 0$	527	228	755
total	1368	338	1706

Table 5. Predicted and observed states for EDUB
(cut-off point α)

Admittedly EDUB, because of its poor fit, is one of the worst examples of the damage uneven sample shares can do, but Table 6 shows that similar results hold for the other analyses with unbalanced samples. Clearly for such samples the conventional 'percentage correctly predicted' does not mean a thing. With the alternative criterion (a cut-off point of α), the overall success rate is lower, but successful prediction is much more evenly spread over the two outcomes. Moreover the percentage correctly predicted now reflects fit rather than sample proportions: among the 8 cases of Table 6 the conventional percentage varies with α and the alternative with λ.

Name	at .5: overall	$Y_i = 1$	$Y_i = 0$		at α: overall	$Y_i = 1$	$Y_i = 0$
PART	.71	.71	.71		.70	.68	.73
CAR	.79	.88	.62		.77	.79	.74
FIBRO	.77	.90	.44		.73	.72	.75
EDUA	.80	.94	.31		.71	.69	.78
ICU	.88	.99	.43		.77	.77	.75
EDUB	.80	.99	.01		.63	.62	.68
DEPRI	.83	1.00	.00		.54	.49	.78
RAPE	.86	.99	.15		.62	.60	.76

Table 6. Fraction correctly predicted at cut-off points of .5 and α.

Detection of outliers

Outliers or atypical observations are conventionally identified by their contribution to the fit of the model, measured by the effect of their deletion. In linear regression large absolute values of the residual indicate an outlier, in discrete models small values of $Pr(i)$ of (5) (the estimated probability of the observed outcome) do so. Pregibon (1981) makes use of

$$d_i^2 = -2 \log Pr(i)$$

and

$$\chi_i^2 = \frac{(1 - Pr(i))^2}{Pr(i)(1 - Pr(i))}$$

which indicate the contribution of observation i to the deviance (minus twice the loglikelihood) and to the Pearson chi-square fit statistic respectively. Both criteria are equivalent to $Pr(i)$, with low values indicating outliers. This stamps observations that are highly unlikely as outliers.

Pregibon uses an evenly balanced sample with a very good fit as an illustration[4]. With unequal sample proportions, however, the two outcome sets do not have an equal chance of yielding outliers: since the prediction probabilities of the less frequent outcome $Y_i = 0$ are substantially lower, these observations are much more readily branded as outliers. The outliers are therefore heavily concentrated among the less frequent outcome.

This is demonstrated in Table 7. The number of observations ranked as outliers is 1% of the sample, with a minimum of 10. The next columns show the share of the rare outcome in the sample, among the conventional outliers, and among outliers according to an alternative criterion.

Name	# outliers	share of $Y_i = 0$ in sample	in outliers by $Pr(i)$	in outliers by $Pf(i)$
PART	37	.48	.89	.84
CAR	28	.36	.75	.43
FIBRO	10	.28	.90	.30
EDUA	10	.22	.80	.70
ICU	10	.20	1.00	.60
EDUB	17	.20	1.00	.94
DEPRI	10	.17	1.00	.80
RAPE	10	.16	1.00	.40

Table 7. Share of $Y_i = 0$ in sample and among supposed outliers by two criteria

There is no hard rule that outliers must reflect the sample distribution of the outcomes. Recording errors or other anomalies may systematically occur

[4]This is the *vasoconstriction* case of Gilliatt(1947) and Finney(1947,1971); α is .51 and λ .52.

more frequently with one outcome than with the other. This may account for the PART result, an evenly balanced sample with the outliers concentrated in one outcome subset. For the other examples columns 2 and 3 demonstrate clearly that an uneven division of the sample leads to an uneven distribution of the outliers in the opposite sense: in four out of six cases *all* outliers refer to the outcome with the smaller sample share.

The alternative is to define outliers according to the observations' contribution to the fit as measured by λ, or - by (25) - to rank the observations by $Pf(i)$ instead of $Pr(i)$. The result is shown in the last column of Table 7. There is some improvement, notably for CAR and FIBRO, but in other cases the imbalance persists. EDUB and DEPRI are particularly disappointing; but this may be due to their very poor fit, which makes any attempt to identify outliers illusory.

Clearly, in unbalanced samples the outliers detected by the conventional criterion should be viewed with reserve, but the proposed remedy does not always work. In case of a very poor fit the very notion of outliers may be out of place.

6 Generalization to other models

The present analysis has been conducted entirely in terms of the vectors $\hat{p}$ and y and their complements, and we have made use of only two properties of the estimated probabilities, viz. the *equality of the means* of (4) and the *orthogonality* of (26)). For a linear probability model with simple regression estimation both properties hold exactly. For logit models with ML estimation, as considered here, the first holds exactly and the second asymptotically. For any other binary probability model both properties hold asymptotically, provided $\hat{p}$ is a consistent estimate of $\mathcal{E}y = p$ (see Cramer (1997)). The present argument therefore holds asymptotically for a wide class of analyses. Whether the two key properties are an acceptable approximation in a finite sample must be verified in each particular instance.

The extension to *multivariate* probability discrete models with $S > 2$ states labelled s is not immediate. The zero mean property as well as the orthogonality again hold asymptotically for any pair $\hat{p}_s$ and y_s, but the easy symmetry of $\hat{p}$ and $\hat{q}$ of the present approach breaks down.

7 Notes on sources

Eight published analyses have been used as illustrations. We once more acknowledge our debt to five authors who have generously provided their data and sometimes their calculations for this purpose. A short description of these sources (and their adaptation, where applicable) is given below in alphabetical order of their acronyms. Unless stated otherwise, the original analysis is a plain binary logit regression with ML estimation; in some cases we have performed this analysis on data originally used in a different manner. For a full appreciation of all quoted analyses the reader must turn to the original publications.

CAR is an analysis of the ownership of private cars by Dutch households in the budget survey of 1980. The presence of a business car in the household is a particularly effective regressor, and this accounts for the good fit. The data set is used extensively for illustrative purposes in Cramer (1991).

DEPRI is based on a 1979 survey of the incidence of depression among 1 000 persons in Los Angeles, reported by Frerichs et al (1981). A subset of 294 observations has been published in the textbook by Afifi and Clark (1990) and has also found its way to the exercises in the Manual of the BMDP computer package of Dixon (1992). The original survey reports a great variety of possible regressors, 37 in number, but only five have been used in exercise LR3 of the BMDP manual, and this is the analysis that has been replicated here. Since the regressors are all categorical variables - sex, marital status, and the like - the sample consists of clusters of several observations with the same regressor values and hence the same $\hat{P}_i$.

EDUA has been taken from a study of the Dutch educational system; the issue is whether a major school reform has indeed improved the performance of the system. Dronkers (1993) compares various changes in status of pupils as a result of schooling in two cohorts before and after the reform. Since he controls for background variables the cohort variable will capture the reform's effect. The present example refers to the transition from low teacher assessment to low achievement score, that is the first top cell of Dronkers' Table 1. We have re-estimated the logit concerned.

EDUB is taken from a study by Oosterbeek and Webbink (1995) of educational choice of Dutch high-school pupils in their final year; the issue is whether or not they intend to continue in higher education (the majority does). This choice is related to 11 regressor variables like the student's social background and school record and economic variables such as the cost and benefits of further schooling. We have re-estimated the logit for the 1982 sample of 1706 students, Table 2 of the article quoted.

FIBRO refers to a study of fibrosis after breast conservation cancer therapy by Borger et al. (1994). The sample consists of Dutch patients treated between 1979 and 1988, for which the degree of fibrosis six years after treatment was established on a four-degree scale by expert inspection. In the original study these four states are analysed by means of a restricted logit model, with much attention being paid to nonlinear effects of the characteristics of diagnosis and treatment. Table 3 of the quoted article lists seven variables that have been retained, with categorical subdivisions increasing the number of estimated

coefficients to 11 (apart from three intercepts). The same data set has here been used for the ML estimation of a simple binary logit model for fibrosis of the most severe degree as a function of the same variables, that is 11 covariates and one intercept.

ICU is based on a study by Lemeshow et al (1988) of survival and death of intensive care patients. The data of the original study refer to a fairly homogeneous sample of 737 patients admitted to an Intensive Care Unit in Springfield in 1983. The covariates are characteristics of the patients prior to admission; we here only use the 8 regressors retained by the authors in their preferred specification of table 3 of the article. The present analysis is based on a subset of 200 observations from the original sample of 737, selected for didactic purposes: it is reproduced in the textbook of Hosmer and Lemeshow (1989) and freely available to third parties at http://www-unix.oit.ukase.edu/-statdata.

PART refers to a sample of 3658 households from the French INSEE household budget survey of 1979. These data have been used extensively for analyses of the labour market participation of married women (in 1979 a minority worked), and they have served as an example in the article on semi-nonparametric (SNP) estimation by Gabler et al (1993). The authors employ 20 regressor variables that have proved useful in earlier research, and compare the results of SNP estimation of the binary choice model with ML estimation of a straightforward probit. We here use a plain logit estimated by ML from the same data set.

RAPE is taken from an investigation into the antecedents of stranger rapists on the basis of various aspects of their *modus operandi* by Davies et al. (1997). The data consist of 210 records collected from British police forces over 28 years from 1965 to 1993; in the study, twelve aspects of the rape and the rapists' behaviour were further investigated. In the present analysis these serve to explain whether he has a previous conviction or not. The data have been put at our disposal by the authors.

References

Afifi, A.A., and V. Clark (1990) *Computer-aided Multivariate Analysis, 2nd edition.* Lifetime Learning Publ., Belmont.

Bakker, Frank M., Martha E. Klein, Nora C. Mesa and Ann R. Braun (1993) Saturation deficit tolerance spectra of phytophagon mites and their phytoseiid predators on cassava. *Experimental & Applied Acarology,* **17**, 97-113.

Borger, J., H. Kemperman, H. Sillevis Smitt, A. Hart, J. van Dongen, J.Lebesque, H. Bartelink (1994) Dose and volume effects on fibrosis after breast-conservation therapy. *International Journal of Radiation, Oncology, Biology and Physics* **30**, 1073-1081.

Cramer, J.S. (1991) *The Logit Model - an introduction for economists.* Edward Arnold, London.

Cramer, J.S. (1997) Two properties of predicted probabilities in discrete regression models. *Tinbergen Institute discussion paper* TI 97-044/4.

Davies, A., K. Wittebrood and J.L. Jackson (1997) Predicting the criminal antecedents of the stranger rapist from his offence behaviour. *Science and Justice* forthcoming.

Dixon, W.J. (ed.) (1992) *Biomedical Data Processing Manual.* California-Princeton Fullfillment Services, Princeton.

Dronkers, J. (1993) Educational reform in the Netherlands: did it change the impact of parental occupation and education ? *Sociology of Education* **66**, 262-277.

Efron, Bradley (1978) Regression and ANOVA with zero-one data: measures of residual variation. *Journal of the American Statistical Association* **73**, 113=121.

Finney, D.J. (1947, 3d edition 1971) *Probit Analysis.* Cambridge University Press, Cambridge.

Frerichs, R.R., C.S. Aneshensel, and V. A. Clark (1981) Prevalence of depression in Los Angeles city. *American Journal of Epidemiology*, **113**, 691-699.

Gabler, Siegfried, Francois Laisney and Michael Lechner (1993) Semi-nonparametric Estimation of Binary-Choice Models With an Application to Labor-Force Participation. *Journal of Business and Economic Statistics* **11**, 61-80.

Gilliatt, R.W. (1947) Vaso-constriction in the finger following deep inspiration. *Journal of Physiology* **107**, 76-88.

Hewlett, P.S. (1969) The toxicity to *Tribolium castaneum* of mixtures of pyrethrins and piperonylbutoxide: fitting a mathematical model. *Journal of Storing Products Research* **5**, 1-9.

Hosmer, David H., and Stanley Lemeshow (1989) *Applied Logistic Regression.* Wiley, New York.

Lemeshow, Stanley, Daniel Teres, Jill Spitz Avrunin, and Harris Pastides (1988) Predicting the Outcome of Intensive Care Unit Patients. *Journal of the American Statistical Association* **83**, 348-356.

Oosterbeek, Hessel, and Dinand Webbink (1995) Enrolment in higher education in the Netherlands. *De Economist* **143**, 367-380.

Pregibon, D. (1981) Logistic regression diagnostics. *Annals of Statistics* **9**, 705-724.