FACULTY OF
ACTUARIAL SCIENCE
&
 econometrics



# A & E NOTE

NOTE AE N4/85

*OUTLIERS IN PROBABILITY MODELS*

J.S. Cramer

University of Amsterdam

# OUTLIERS IN PROBABILITY MODELS

Note AE N4/85

J.S.Cramer

Instituut voor Actuariaat en Econometrie, Jodenbreestraat 23,
1011 NH Amsterdam

Abstract I propose a likelihood ratio criterion for identifying outliers among individual observations to which a probability model has been fitted. Outliers are found to occur especially among the rare sample categories, and further analysis suggests that the fit of a probability model is systematically biased in favour of the predominant categories. I intend to pursue the analysis with drawing further conclusions from Figure 1, and this may possibly lead to a proposal for (re)weighting observations in the estimation of a probability model.
The analysis is throughout illustrated by the application of a multinomial logit model to four automobile ownership modes of some 3 000 households from a budget survey.

## 1. A logit analysis of automobile ownership

We have studied the ownership of (private) automobiles among 2819 households of the Dutch CBS household budget survey of 1980 by means of a standard multinomial logit analysis. Four ownership classes or categories are distinguished, viz.

NONE - no automobile owned.

USED - one automobile, bought second-hand

NEW - one automobile, first owner

MORE - two or more automobiles

These ownership classes all refer to private automobiles only, in contrast to business cars as defined below.

Let s = 1, 2, .. S denote these categories or states, with S here equal to 4, $A_s$ the corresponding index sets, and i = 1, 2, ... n with n = 2819 indicate the sample households. The logit model then is

$$P(i \in A_s) = p_{si} = \frac{\exp(x_i' \beta_s)}{\sum_t \exp(x_i' \beta_t)} \qquad (1)$$

We normalize the parameter vectors by writing

$$\beta_1 = 0 \qquad (2)$$

where 1 refers to the class NONE.

The vector $x_i$ consists of a unit constant and four regressor variables. These are

BUSC - a business car dummy for the presence of a business car, a company car, an expense account car or a tax deductible car; 1 if such a car is present, 0 otherwise.

URBA - degree of urbanization, graded in six classes from 1 (cities) to 6 (country).

LPER - household size, measured by log of number of persons.

LYPP - income level, measured by log of disposable household income per person.

With four states, the normalization (2), and five coefficients in $\beta_s$, we have fifteen coefficients. These have been estimated by maximizing the loglikelihood function

$$\log L(\theta) = \sum_s \sum_{i \in A_s} \log p_{si}(\theta) \qquad (3)$$

Here $p_{si}$ is taken from (1) and $\theta$ is the fifteen element vector comprising $\beta_2$, $\beta_3$ and $\beta_4$. The (asymptotic) variances of the estimates follow from the information matrix.

To facilitate the interpretation we give a somewhat redundant presentation of the results. By (1) we have

$$\log(p_{si}/p_{ti}) = x_i'(\beta_s - \beta_t) \qquad (4)$$

These regression coefficients of the log of the odds ratio has been reported in the upper half of Table 1 for all pairs (s, t). We have also calculated the partial derivatives of the probabilities of the four states in respect of the regressor variables, or

$$\frac{\partial p_s}{\partial x_j} = p_s \left\{ \beta_{sj} - \sum_t \beta_{tj} p_t \right\} \qquad (5)$$

These derivatives have been evaluated at the sample mean frequencies. They are shown in the bottom half of Table 1. All estimates are accompanied by their (asymptotic) t-ratio's, given in absolute value.

- see here Table 1 -

The estimated coefficients have quite small standard errors, and the results make sense. The presence of an expense account car (BUSC) discriminates only between NONE and any form of private automobile ownership, and so does URBA. Household size, represented by LPER, has a strong effect on the number of cars owned, but none on the choice between USED and NEW. Income (LYPP) once again works in favour of all types of ownership, but particularly in favour of new cars. All this is as it should be.

We can also assess the model's performance by the equivalent of the F test of a regression equation. When we ignore all the regressor variables, but retain the constant and its intercept coefficient, this is equivalent to equating twelve parameters to zero. The Maximum Likelihood estimates of the $p_{si}$ in this case correspond to the sample frequencies, or

$$\hat{p}_{si} = f_s = n_s/n \qquad (6)$$

The loglikelihood of this primitive base-line model is therefore

$$\log L_0 = \sum_s n_s \log f_s = \sum_s n_s \log n_s - n\log n \qquad (7)$$

This must of course be less than the maximum of (3), since we have imposed twelve restrictions, and we may test these by a likelihood ratio test. In the event we find

$$\log L_0 = -3527, \qquad \log L(\hat{\theta}) = -2954.$$

Twice the difference is 1146, and this is of course highly significant for a chi-square variable with 12 degrees of freedom.

By the usual standards of this rather superficial inspection the present analysis is reasonably successful. This should be borne in mind in the sequel.

Table 1. Estimates of logit model parameters, n=2819
(absolute values of t-ratio's in brackets)

regression coefficients of equ.(4)

| (s/t) | BUSC | URBA | LPER | LYPP |
|---|---|---|---|---|
| USED/NONE | -2.89 | -0.14 | +2.90 | +2.10 |
| | (15.29) | (4.53) | (16.12) | (13.06) |
| NEW/NONE | -3.03 | -0.10 | +2.90 | +3.15 |
| | (14.32) | (3.09) | (14.84) | (17.59) |
| MORE/NONE | -3.56 | -0.08 | +6.81 | +4.55 |
| | (9.57) | (1.44) | (19.60) | (15.38) |
| NEW/USED | -0.14 | +0.03 | +0.00 | +1.05 |
| | (0.58) | (1.12) | (0.02) | (6.78) |
| MORE/USED | -0.67 | +0.06 | +3.92 | +2.46 |
| | (1.76) | (1.08) | (12.24) | (8.95) |
| MORE/NEW | -0.53 | +0.02 | +3.91 | +1.41 |
| | (1.37) | (0.42) | (11.91) | (5.13) |

partial derivatives of equ.(5)

| state | BUSC | URBA | LPER | LYPP |
|---|---|---|---|---|
| NONE | +0.69 | +0.03 | -0.75 | -0.63 |
| | (18.87) | (4.27) | (19.71) | (18.12) |
| USED | -0.32 | -0.02 | +0.27 | +0.11 |
| | (8.07) | (3.70) | (8.16) | (4.00) |
| NEW | -0.27 | -0.01 | +0.20 | +0.34 |
| | (7.28) | (1.29) | (6.56) | (12.89) |
| MORE | -0.10 | -0.00 | +0.29 | +0.17 |
| | (4.84) | (0.09) | (15.90) | (11.19) |

## 2. Identification of outliers

The 2819 individual household observations may be expected to contain a certain number of outliers, or atypical observations. Such observations may be due to errors of observation or to eccentric behaviour of the household concerned; in either case we should like to eliminate the observation from the analysis, whether it materially affects the parameter estimates or not. We wish to identify these observations, and to gauge their effect on the estimates.

A probability model does not yield residuals, like a regression equation; the nearest analogy to a large residual, and prima facie evidence of an outlier, is severe misclassification, or a very small value of the predicted probability of the state that actually occurs. This is given by

$$p_j = p_{s(j)j}(\hat{\theta}) \tag{8}$$

where s(j) denotes the observed state of the j'th observation.

We shall indeed use this criterion to delete observations. To delete the j'th observation is equivalent to the introduction of a (0,1) dummy $D_j$ in (1), which is 1 for the j'th observation and 0 elsewhere. The extended model is

$$p_{si} = \frac{\exp(x_i'\beta_s)}{\sum_t \exp(x_i'\beta_t)} + \gamma_s D_j \tag{9}$$

As in (2) we have

$$\beta_1 = 0$$

while moreover

$$\sum_s \gamma_s = 0 \tag{10}$$

This last condition (10) must be imposed in order to preserve the properties of probabilities for the left-hand side of (9).

Inspection will show that in the extended model (9) the $\hat{\gamma}$ can be adjusted to ensure a perfect fit for the j'th observations, regardless of the values of the $\hat{\beta}_s$ or $\hat{\theta}$. At the maximum of the loglikelihood we will therefore have

$$p_j(\tilde{\theta}, \tilde{\gamma}) = p_{s(j)j}(\tilde{\theta}, \tilde{\gamma}) = 1, \qquad \log\hat{p}_j = 0.$$

The j'th observation therefore does not contribute to logL, regardless of the values of the $\hat{\beta}_s$, and the addition of three new parameters - four $\gamma_s$ subject to (10) - is equivalent to deleting a single observation. It is easy to see that this holds for any specification of the probability model, and that the removal of an observation is always equivalent to the addition of (S - 1) extra parameters.

The effect of deleting one observation on the maximum of the loglikelihood function is to omit one term from the summation of (3), and to change the parameter estimates from $\hat{\theta}$ to $\tilde{\theta}$. As a result the loglikelihood increases by

$$\Delta\log L(j) = -\log p_j(\hat{\theta}) + \sum_s \sum_{i \in A_s} \left\{ \log p_{si}(\tilde{\theta}) - \log p_{si}(\hat{\theta}) \right\} \qquad (11)$$

The increase in logL upon the removal of the j'th observation can thus be decomposed in (i) a part which is directly due to the removal of an awkward observation, and (ii) a part which reflects the improved fit to the remaining observations. Note that both terms are nonnegative.

By the argument given above, the same result can be obtained by the introduction of (S - 1) new parameters, and the hypothesis that these are all zero, or that the j'th observation is not an outlier, is amenable to a likelihood ratio test. Under the null twice $\Delta\log L(j)$ is (asymptotically) chi-square distributed with

(S - 1) degrees of freedom. In the present case, with S = 4, the j'th observation is a significant outlier at the 5% level if $\Delta\log L(j)$ exceeds 7.815/2 = 3.907. As both terms of (11) are non-negative, the j'th observation may register significance by the first term alone; it is sufficient that

$$-\log p_j(\hat{\theta}) > 3.907$$

that is

$$p_j(\hat{\theta}) < .0201 \qquad (12)$$

This is an easy criterion for identifying outliers. There is no need for repeated estimation; all we have to do is to calculate the n values

$$p_j(\hat{\theta}) = p_{s(j)j}(\hat{\theta})$$

for a single set of estimates from the analysis under review. As the derivation above shows, the critical minimum (here .0201) depends only on the number of categories S, not on the fit of the model or on any characteristic of the analysis. As S increases, this critical value declines rapidly, as can be seen from Table 2.

- see here Table 2 -

The present criterion primarily identifies atypical observations, and we can only hope that these include the observations that are influential in determining the parameter estimates, for these are of course the observations we are after. As the second term of (11) is ignored, influential observations that do not show up by their eccentric position are overlooked. This is a drawback of the method. If we really wish to concentrate on influential outliers, we must look at the second term, and there is nothing for it but to re-estimate $\theta$ with each of the n observations omitted in turn.

Table 2. Critical minimum value of $\hat{p}_j$ for various S

| S | critical value of $\hat{p}_j$ at significance levels of | |
|---|---|---|
| | .05 | .01 |
| 2 | .1465 | .0362 |
| 3 | .0500 | .0100 |
| 4 | .0201 | .0034 |
| 5 | .0087 | .0013 |
| 6 | .0039 | .0005 |
| 7 | .0018 | .0002 |
| 8 | .0009 | .0001 |
| 9 | .0004 | - |
| 10 | .0002 | - |

- less than .00005

## 3. An application

When we apply the criterion (12) to the logit analysis of automobile ownership we at first find 16 significant outliers among the 2819 observations. This is far less than the number we would normally expect to exceed the 5% significance level, and it reflects the fact that we use a lower limit of the true test statistic since part of (11) is neglected. It may also suggest that the right-hand tail of the distribution of the $p_i$ is abnormally thin; this would mean that the outliers are indeed far out, and indicate genuine freak observations.

Upon deleting the 16 observations from the analysis, the loglikelihood increases by 83.9. When we partition this overall effect along the lines of (11), the major part (80.3) is due to the removal of awkward observations, and only 3.6 reflects the improved adjustment to the remaining observations. It is clear that the effect on the estimates is small; yet they still change enough for the same criterion (12) to yield another six outliers at the revised estimates. When we delete these in turn, one new outlier appears. Deleting this one yields no new outliers.

The effect of omitting the 23 outliers on the estimates of the parameters can be gauged by comparing Table 3 with Table 1. While there is a further overall improvement in the t-ratios, the results are not very much different from what we had before; but it should be realized that the change is due to the removal of less than one per cent of the sample observations.

- see here Table 3 -

We should of course also find out by inspection in what sense the outliers diverge from the norm, or from the model. The first thing we observe is that their distribution over the four ownership categories is very uneven; we find the following incidence

Table 3. Estimates of logit model parameters after removing outliers (absolute values of t-ratio's in brackets)

regression coefficients of equ.(4)

| (s/t) | BUSC | URBA | LPER | LYPP |
|---|---|---|---|---|
| USED/NONE | -3.15 | -0.15 | +3.22 | +2.39 |
| | (16.00) | (4.92) | (16.99) | (14.07) |
| NEW/NONE | -3.38 | -0.12 | +3.27 | +3.53 |
| | (15.07) | (3.45) | (15.81) | (18.61) |
| MORE/NONE | -4.22 | -0.12 | +7.78 | +5.40 |
| | (10.09) | (2.13) | (20.32) | (16.64) |
| NEW/USED | -0.23 | +0.03 | +0.05 | +1.14 |
| | (2.54) | (1.12) | (0.25) | (7.16) |
| MORE/USED | -1.07 | +0.03 | +4.57 | +3.01 |
| | (2.54) | (0.49) | (12.96) | (10.06) |
| MORE/NEW | -0.84 | -0.01 | +4.52 | +1.87 |
| | (1.96) | (0.13) | (12.55) | (16.27) |

partial derivatives of equ.(5)

| state | BUSC | URBA | LPER | LYPP |
|---|---|---|---|---|
| NONE | +0.77 | +0.03 | -0.84 | -0.71 |
| | (19.69) | (4.76) | (20.71) | (19.25) |
| USED | -0.34 | -0.02 | +0.29 | +0.13 |
| | (8.23) | (3.88) | (8.60) | (4.47) |
| NEW | -0.30 | -0.01 | +0.23 | +0.38 |
| | (7.82) | (1.43) | (7.23) | (13.67) |
| MORE | -0.12 | -0.00 | +0.32 | +0.20 |
| | (5.49) | (0.75) | (16.55) | (12.39) |

| | number of sample observations | number of outliers |
|---|---|---|
| NONE | 1009 | 7 |
| USED | 944 | 3 |
| NEW | 691 | 4 |
| MORE | 175 | 9 |

Thus the least frequent category yields the largest number of outliers, and the relative incidence of outliers in that class is ten times as large as elsewhere.

This result prompts a closer look at the adjustment of the model in terms of the predicted probabilities $p_{s(i)i}(\hat{\theta})$ of each of the four automobile ownership classes. All loglikelihoods as well as their differences (like (11)) consist of summations, and they can therefore be decomposed into separate components for each of the four classes. This has been done in Table 4. In the top half we show how we proceed from the primitive base-line model to the standard logit model, and from there to the further refinement of a logit model with all the outliers deleted. In the bottom half we characterize the adjustment of each model for each ownership class by the geometric mean predicted probability, or

$$\exp\left\{\frac{1}{n_s}\sum_{i\in A_s}\log p_{s(i)i}(\hat{\theta})\right\}.$$

In an ideal adjustment the predicted probabilities of the state that actually occurs should be somewhere close to 1; in the more modest requirement that the state actually occurring has the largest probability, it should at least be .25 (since we have four alternatives).

The values of Table 4 lag far behind these values, in particular for the small categorie MORE. Now in the primitive or base-line model it is immediately clear that rare states are badly predicted, since the predicted probabilities are equal to the sample frequencies. It is not so immediately apparent, however, that this bias persists when we adjust a slightly more sophisticated model. Yet a little reflection will show that this must be so, since a small improvement in the mean

predicted probability of a large class contributes more to the maximization of logL than a substantial improvement in the mean of a small class. The stakes are thus weighted in favour of the predominant categories.

- see here Table 4 -

The analysis of Table 4 is still in terms of the overall or average fit allbeit within separate categories. The argument just given does not necessarily carry over to individual observations, for a bad fit (a low probability) for a single observation detracts as much from logL in a small category as in a large category. Yet in fact it turns out that the entire distribution of the predicted probabilities shifts along with their geometric mean. We show these distributions for each of the four classes in Figure 1, for the best model we have, i.e. the logit model after removal of 23 outliers. Against the background of the ideal values of 'close to 1' or at least 'over .25' that we have quoted, the actual distributions are terrible, and the actual distribution for the class MORE is the worst of all. It is now plain why most outliers occur in that group. It is also painfully clear how badly an apparently satisfactory analysis performs when it comes to the prediction of individual behaviour, even within the sample used for estimation.

- see here Figure 1 -

Table 4. Decomposition of loglikelihoods by ownership classes

| | | | base-line model | Δ add 12 β's | original logit model | deleting 23 outliers # | Δ obs | Δ fit | logit model without 23 outliers |
|---|---|---|---|---|---|---|---|---|---|
| loglikelihood | | | -3527.4 | +573.2 | -2954.1 | | +106.7 | +5.9 | -2841.5 |
| degrees of freedom | | | +2816 | -12 | +2804 | -23 | | | +2781 |
| s | $n_s$ | $f_s$ | $n_s \log f_s$ | | | | | | |
| 1 | 1009 | .358 | -1036.5 | +254.4 | -782.1 | -7 | +31.9 | +4.6 | -745.6 |
| 2 | 944 | .335 | -1032.4 | + 93.4 | -939.0 | -3 | + 1.6 | +4.1 | -923.3 |
| 3 | 691 | .245 | - 971.9 | + 95.8 | -876.1 | -4 | +22.2 | +0.4 | -853.5 |
| 4 | 175 | .062 | - 486.6 | +129.7 | -356.9 | -9 | +41.0 | -3.2 | -319.1 |
| | | | | | geometric mean predicted probabilities | | | | |
| 1 | | | .358 | | .461 | | | | .475 |
| 2 | | | .335 | | .370 | | | | .375 |
| 3 | | | .245 | | .281 | | | | .289 |
| 4 | | | .062 | | .130 | | | | .146 |

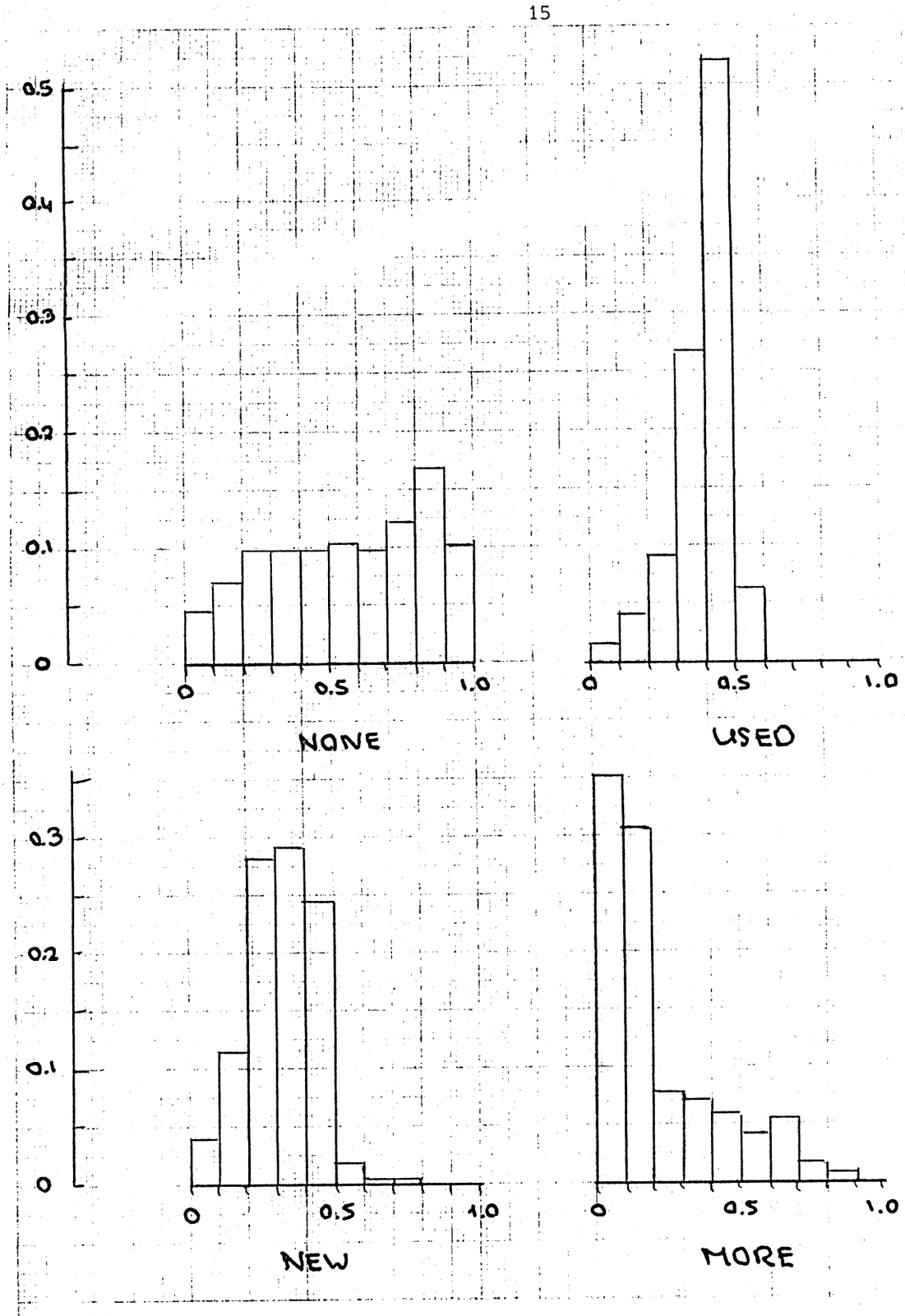


Figure 1. Relative frequency distributions of predicted probabilities in four ownership classes