Stefania Mignani

Notes on goodness of fit tests for the logistic regression model

Serie Ricerche 1996, n.1

Dipartimento di Scienze Statistiche "Paolo Fortunati"
Università degli studi di Bologna

Finito di stampare nel mese di Aprile 1996
presso le Officine Grafiche TECNOPRINT S.N.C.
Via del Legatore 3, Bologna

1. The logistic regression model

1.1 Introduction

The looking for models synthesizing the relationships between a set of variables has found an answer in the linear model that is a simple and, at the same time, a powerful tool to describe and interpret these relations. From the first developments by Gauss and Legendre for the study of continuous variables (with a clear reference to normal error distribution), the linear model has also been extended to discrete and qualitative variables. The generalized linear models include all the models based on a linear combination of explanatory variables to analyze data in the form of counts and in the form of proportions.

In general, the model is defined by the equation

$$Y = F(\beta' \mathbf{X}) + \varepsilon \tag{1.1}$$

where F, the link function, is a real-valued function which operates in the index $\beta' \mathbf{X}$ and constitutes the systematic part of the model; E is an error term. Assuming that $E(\varepsilon | \mathbf{X} = \mathbf{x}) = 0$, model (1.1) is usually represented as

$$E(Y | \mathbf{X} = \mathbf{x}) = F(\beta' \mathbf{X}) \tag{1.2}$$

This model can take several forms allowing to study a variety of situations: McCullagh and Nelder (1989) give a wide survey on generalized models showing several applications in biometrics; Amemya (1981) describes these models from an econometric point of view In the context of binary response models, if Y is coded as one or zero (success or failure), equation (1.2) can have the probit formulation $E(Y | \mathbf{X} = \mathbf{x}) = \Pr(Y = 1 | \mathbf{X} = \mathbf{x}) = \Phi(\beta' \mathbf{x})$ or the logit formulation:

$$\pi(\mathbf{x}) = \Pr(Y = 1 | \mathbf{X} = \mathbf{x}) = \frac{\exp(\beta' \mathbf{x})}{1 + \exp(\beta' \mathbf{x})} \tag{1.3}$$

that determines the logistic regression model, formalized by Cox (1970), but already used in an intuitive way in some biological researches (Berkson, 1944, 1951; Dyke and Patterson, 1952).

In (1.3) the logarithm of the conditional odds of success to failure is modeled as a linear function of the K explanatory variables $\mathbf{X} = (X_1,\ldots,X_k\ldots,X_K)$:

$$logit(\mathbf{x}) = ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = g(\mathbf{x}) = \beta'\mathbf{x} \tag{1.4}$$

The sign of the k-th parameter β_k determines whether an increase in the k-th variable X_k leads to corresponding decrease in $\pi(\mathbf{x})$. From this viewpoint the analogy with the linear regression model is immediate. On the other hand, however, the binary nature of the dependent variable leads to fundamentally different assumptions: in fact the conditional distribution of the outcome variable has a binomial distribution with probability given by the conditional mean $E(Y|\mathbf{X}=\mathbf{x}) = \pi(\mathbf{x})$, bounded between zero and one. hus the error term has a binomial distribution with variance equal to $\pi(\mathbf{x})(1-\pi(\mathbf{x}))$ depending on $\mathbf{x}$.

1.2. The model fitting

Let $\{y_i, x_i = x_{i1},\ldots,x_{iK}; i = 1,\ldots,n\}$ be a sample of n observations. If some of the explanatory variables are continuos there may be a specific covariate pattern (setting) $x_1,\ldots,x_K$ for each subject, otherwise, if only categorical variables are considered, repeated covariate settings may be possible.

An estimate of the unknown parameters can be obtained by applying the maximum likelihood method[1]; the log-likelihood is defined as[2]:

[1] We don't deal with the method of least square because in the case of dichtomot outcome it do not has the usual properties

[2] For sake of simplicity, π_i stands for $\pi(\mathbf{x}_i)$

$$l\{\beta; y\} = \sum_{i=1}^{n}\{y_i \ln\pi_i + (1-y_i)ln(1-\pi_i)\} \tag{1.5}$$
$$= \sum_{i=1}^{n}\{y_i\beta' x_i - ln(1+exp(\beta' x_i))\}$$

The maximum likelihood estimate $\hat{\beta}$ maximizes (1.5) and satisfies condition $\mathbf{x}'\mathbf{r} = \mathbf{0}$ where $\mathbf{r} = \mathbf{y} - \hat{\pi}$ and $\hat{\pi}_i = exp(\hat{\beta}'\mathbf{x}_i)/(1+exp(\hat{\beta}'\mathbf{x}_i))$, showing the linkage with the least squares obtained for the linear regression. The first order equations are however non linear in $\hat{\beta}$ and therefore iterative methods are required to solve them; using the Newton-Raphson method $\hat{\beta}$ is expressed at the $(t+1)$-th iteration as

$$\hat{\beta}(t+1) = \hat{\beta}(t) + (\mathbf{X}'\mathbf{V}(t)\mathbf{X})\mathbf{X}'\mathbf{r}(t), \quad t = 1,2,\ldots \tag{1.6}$$

where $V = diag\{\hat{\pi}_i(1-\hat{\pi}_i)\} = Cov(y)$ and the arguments of $\mathbf{V}$ and $\mathbf{r}$ refer to the values of these quantities calculated at $\hat{\beta}(t)$. It often takes only a few iterations to get satisfactory convergence.

1.3 Adequacy of an estimated model

Once a model has been chosen and fitted to the observed data, it is necessary to assess how effective the functional expression is in describing the outcome variable in terms of variability. This is the well-known problem of " goodness of fit", requiring to check whether the estimated (or predicted) values give an accurate representation of the observed values, *i.e.* if the observed differences can be attributed to sampling error or model misspecifications. The overall assessment of goodness of fit follows two logical step: the computation of a measure of fit and the application of an hypothesis testing.

Measures of discrepancy may be evaluated in different ways. One strategy is based on the comparison of the simplest model (or null model) containing only the intercept with the saturated model (or full model) containing as many parameters as the data points. Other solutions are obtained by evaluating the difference between estimated values and observed values.

In this paper the most popular methods of assessing the fit of an estimated logistic regression model are discussed with the assumption that the model contains all the significant variables.

2. Likelihood ratio statistics and R^2 type measures

In the ordinary linear regression model the residuals sum of squares plays a central role in assessing the goodness of fit. Several attempts have been done to define similar measures in the case of a qualitative dependent variable but none of these measures are is widely used.

Let l_K be the log-likelihood for the fitted model (current model) containing the intercept and the the K covariates and let l_s denote the (maximized) log-likelihood for the saturated model. A first measure is given by the likelihood ratio, called deviance

$$D = -2(l_K - l_s) = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{(1-\hat{\pi}_i)}{(1-y_i)}\right)\right] \qquad (2.1)$$

Other quantities have been proposed in analogy with the well known R^2 for the linear regression model. Let be l_0 the (maximized) log-likelihood for the null model. As the model complexity increases, the parameters space expands, so the value of the maximized likelihood increases. Thus the enequality $l_0 \le l_K \le l_s$, holds and the measure (improperly called R^2)

$$R^2 = \frac{l_K - l_0}{l_s - l_0} \qquad (2.2)$$

that lies in the range from 0 to . can be computed.

From equation (1.5) the null model gives $\hat{\pi_i} = \sum_i y_i / n$ so that

$$l_0 = n[\bar{y}(\log(\bar{y}) + (1-\bar{y})\log(1-\bar{y})]$$

The saturated model has a dummy variable for each subject, and so $\hat{\pi_i} = y_i$ for all i . Thus, $l_s = 0$ and the (2.2) becomes

$$R^2 = 1 - \frac{l_K}{l_0} \qquad (2.3)$$

as proposed by McFadden (1974).

In these equations R^2 is a different expression of the likelihood ratio, as Hosmer and Lemeshow pointed out (1989), so it is not a measure of goodness of fit. In fact this ratio compares fitted values under two models rather than comparing the observed values to those fitted by the current model.

An alternative solution belongs to a family of measures developed by Efron (1978) using an axiomatic approach. The proposed index

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{\pi_i})^2}{\sum_i (y_i - \bar{y})^2} \qquad (2.4)$$

measures the associatioii between the observed responses and their fitted values. It is a natural criterion corresponding to the standard one used for the linear regressioii[3].

According to this R^2 failures to incorporate the dependence of the variance of Y_i on π in the error structure. Amemya (1981) suggested a related measure -

[3]When the probability linear model is fitted by ordinary least squares this particular R^2 simplifies to the standard R^2 for regression modelling.

less simple to interpret- that weights square deviations by a weightcoefficient which is inversely proportional to the predicted variance.

Latila (1993) has recently proposed a modified version of R^2 that can be interpreted in terms of explained and unexplained variation of the underlying latent model.

Qualitative dependent variable models can be viewed as consisting of two parts: the first specifies a structural relationship between the regressors and an underlying variable y_i^*, while the second part specifies how the dependent variable is observed. In the logistic regression model we have:

$$y_i^* = \beta' \mathbf{x}_i + u_i \qquad (2.5)$$

where the cumulative distribution function of u_i is the logistic i.e.: $u_i \approx L.\left(0, \sigma^2 = \frac{\pi^2}{3}\right)$.

In practice y_i^* is unobservable, what we observe is a dummy variable y defined hy

$$y_i = 1 \ \text{if} \ y_i^* > 0 \qquad (2.6)$$
$$y_i = 0 \ \text{otherwise}$$

The proposed $R^2{}_P$ is based on the expression for the standard R^2 in terms of estimatcd coefficicnt and regressors sample covariance matrix:

$$R_2^P = \tilde{\beta}' \tilde{\Sigma} \tilde{\beta} / (\hat{\sigma}^2 + \tilde{\beta}' \tilde{\Sigma} \tilde{\beta}) \qquad (2.7)$$

where β is a consistent estimator of β. This index has the same asymptotic limit and the same interp,etation as the conventional R^2.

3. Chi-square type measures.

Different solutions have been proposed according to the kind of observations: if there is a limited number of different covariate patterns and replicate measurements for each of them, goodness of fit can be examined by the methods developed for categorical data. otherwise it is necessary to group sample units in some way.

3.1 Statistics for repeated observations

One of the most popular statistic is the Pearson chi square that can be directly calculated when repeated observations are avaliable. This statistic is based on residuals. In general. the raw residual is defined as the observed value minus the fitted value. In a logistic model the residuals are difficult to interpret, because they are differences between discrete and continuous quantities for which a normal distribution is usually not appropriate (Azzalini *et* al. 1989). Furthermore. ench of these residuals has a two-point distribution that depends on $\mathbf{x}$ through $\pi(\mathbf{x})$: that is, each residual has a specific distribuiion. Cox and Snell (1968) defined modified residuals which reduce the problem of discreteness but difficulties remain in samples with sparse data where covariate values are irregularly spread over a large number of points.

Let J be the number of distinct values of observed $\mathbf{X}$ (if some units have the same value of $\mathbf{X}$ then $J < n$) and let m_j ($j=1,\ldots,J$) denote the number of subjects wiih $\mathbf{x} = \mathbf{x}_j$: it follows that $\sum_i m_j = n$. Furthermore let y_j be the numher of units with positive response $Y = 1$ then, defined the estimnted value as

$$m_j \hat{\pi}_j(\mathbf{x}) = m_j (\exp[\hat{g}(\mathbf{x})] / \{1 + \exp[\hat{g}(\mathbf{x})]\}) \qquad (3.1)$$

Pearson's residuals are defined as follows

$$r(y_j,\hat{\pi}_j)=\frac{(y_j-m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}} \qquad (3.2)$$

The summary statistic based on these residuals is

$$X^2=\sum_{j=1}^{J} r(y_j,\hat{\pi}_j)^2 \qquad (3.3)$$

Another measure of the distance between the observed and the fitted values is the deviance residual defined as

$$d(y_j,\hat{\pi}_j)=\pm\left\{2\left[y_j \ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right)\right]+(m_j-y_j)\cdot\left(\frac{(m_j-y_j)}{m_j(1-\pi_j)}\right)\right\}^2 \qquad (3.4)$$

where the sign is the same as the sign of $(y_j-m_j\hat{\pi})$; the summary statistic based on these residuals is the deviance

$$D=\sum_{j=1}^{J} d(y_j,\hat{\pi}_j)^2 \qquad (3.5)$$

that corresponds to the quantity shown in equation (2.1).

The distribution of both statistics, under the null hypothesis that the fitted model is fully adequate is approximated by a chi-square with $J-K-1$ degrees of freedom. This statement derives from the fact that D is the likelihood ratio test statistic of a satured model with J parameters versus the model fitted with $k+1$ parameters. A similar theory provides the null distribution of X^2 but this measure is unstable for fitted values near zero or one. If however $J \approx n$ the p-values are incorrect because distributional results have been obtained under the condition that only n becomes large (the so-called n asymptotics). The problem thus remains of how grouping the subjects to meet the condition called m-asymptotics according to which, when fixing $J <$ nand letting n become large, each value of m_j tends to become large.

3.2 *Pooling observations according to estimated probabiliiies: the Hosmer - Lemeshow test statistics*

Hosmer and Lemeshow (1980. 1982) considered the chi-sqaure statistic determined in a $2 \times J$ table, where the rows correspond to the two values of the outcome variable $Y = 0,1$ and the columns correspond to the J values of the explanatory variables in the group, showing that the p-values are correct when the expected values are sufficiently "large" in each cell. This condition holds under m-asymptotics. In the $2\times J$ table the expected values will always be quite small since the number of columns increases as n increases: this gives rise to the problem of sparse data that must be treated with specific solutions.

Thus, goodness of fit statistics require to group observed and fitted values in such a way that m-asymptotics can be used. Hosmer and Lemeshow proposed a very simple solution: reduce the number of columns, *i.e.* collapse them on the basis of the estimated probabilities.The estimated probabilities are set out in increasing order and grouped in g intervals that identify the columns of a new $2\times g$ contingency table in which each cell of the first row gives the observed number of units with $Y=1$ and each cell of the second row contains the units with $Y = 0$.

The problem is then how to construct the intervals: two alternative solutions have been proposed. The first is based on percentiles of the estimated probability. according to which, having fixed a priori the number g of groups (very often it is assumed $g = 10$) and set out the probabilities $\hat{\pi}_i$ in increasing order, the cutpoints of the intervals are fixed so as to include the same number of units: in other words this involves asserting that the marginal row distribution is uniform. The first group contains then $/\,g$ units having the smallest estimated probnbilities while the last group includes the n/g units having the largest estiinnted probnbilities. Intuitively. if the model holds, then the probabilities $\hat{\pi}_i$ for those individuals that show $Y=1$ will be found in the upper percentiles *i.e.* in the last columns of the table. With this in mind. for

each interval the observed value is calculated as the sum of units with $Y = 1$ and the expected value is calculated as the sum of estimated probabilities for all units belonging to the interval. Table 1 shows this solution.

Table 1

Dependent variable	Probability intervals				
	1	2	...	g	Total
y=1	o_{11}	o_{12}	...	o_{1g}	n_1
y=0	o_{01}	o_{02}	...	o_{0g}	n_2
Total	n/g	n/g		n/g	n

The statistic comparing the observed values with expected values is defined as

$$\ddot{C}_g = \sum_{h=0}^{1}\sum_{l=1}^{g}\frac{(o_{hl} - e_{hl})^2}{e_{hl}} \qquad (3.6)$$

where

$$o_{1l} = \sum_{i\in T_l} y_i$$

$$o_{0l} = \sum_{i\in T_l}(1 - y_i)$$

$$e_{,,} = \sum_{i\in T_l}\hat{\pi}_i \qquad (3.7)$$

$$e_{,,} = \sum_{i\in T_l}(1-\hat{\pi}_i)$$

and T_l indicates the set of the n / g units of l -th percentile (or interval).

Under the null hypothesis, the distribution of $\hat{C}_g$ is well approximnted by the chi-square distribution with $g-2$ degrees of freedom (Hosmer and Lemeshow, 1980[4])

This solution ensures a reasonable number of units for each probability interval. but on the other hand. the true values of the estimated probabilities are ignored. Furthermore each interval dcpends on the particular sample observed and therefore makes the compnrison with other samples very difficult. As an alternative solution, Hosmer and Lemeshow suggested to form groups with preset intervals and then to calculate, as in the previous statistic, the observed values as the sum of the units of each interval and the expected values as the sum of the probabilities. Table 2 summarises this solution

Table2

Dependent variable	Probability intervals				
	[O-0.1)	[0.1-0.2)	...	[0.9-1.0]	Total
y=1	o'_{11}	o'_{12}	...	o'_{1g}	n_1
y=0	o,,,	o'_{02}	...	o'_{0g}	n_2
Total	n'_1	n'_2		n'_g	n

This scheme gives the statistic

[4] This approximation has heen shown by simulation, starting from the results offered by Moore and Spruill (1975) who dealt with the problem of how to extend the usual thery for the goodness of fit chi-sqare test when the parameters are estimated for ungrouped data and the frequencies in tlie 2xg table depend on tlie estunated parameters. *i.e.* the cells are variable and not fixed

$$\hat{H}_g = \sum_{h=0}^{1} \sum_{l=1}^{g} \frac{(o'_{hl} - e'_{hl})^2}{e'_{hl}} \qquad (3.8)$$

which, underthe null hypothesis, also has a distribution approximated by a χ^2 with $g-2$ degrees of freedom.

The validity of these two solutions is essentially proved in an empirical way by means a set of simulations. Hosmer and Lemeshow have stated that the statistic $\hat{H}$ is more powerful than $\hat{C}$ even if the latter appears to have a better approximation to χ^2_{g-2}, in particular when many of the estimated probabilities are low, e.g less than 0.2 (Hosmer, Lemeshow and Kler, 1988). In fact if the sample is small and the probabilities are concentrated in a few values, when the second method is used it leads to classes with a few units and therefore problems arise in the determination of the expected values and in the assessment of the results obtained applying the statistic $\hat{H}$.

If on the one hand, as pointed out by Demaris (1992), some results appear to emphasize the tendency of these statistics to confirm the model too often, on the other hand they offer, at least in terms of a first data control, the possibilty of assessing, on the basis of the tables 1 or 2 (where the expected values are reported), in which regions the model have not a good fit.

These grouping criteria offer a valid solution but some cautions are required. These procedures consider similar units that have very close probabilities of giving specific results, but these units might refer to $\mathbf{X}$ values which can be similar with respect to the relationshiè studied and can be very distant in the covariate space. It would therefore be advisable to assess the variability of the regressors inside each group to check the results obtained: it may even occur that units with very close $\mathbf{X}$ values have an estimated probability that falls within different intervals. Moreover, another very delicate problem concerns the choice of the number of intervals in which the estimated probabilities are divided: one of the most frequently solutions adopted, in particular in epidemiological studies, is that of assuming $g = 10$, but this choice can be justified only empirically. This problem makes it even more

imperative to establish some group identification criteria that are not linked to the specific situation, but that can be used in all situations, and thus makes the results less subjective.

The methods based on grouping are therefore completely insensitive to differences in logit within the pooled groups, in fact the observations are grouped in such a way that the local fluctations are canceled out in each pooled cell. Hence these statistic are not able to detect the deviations of the model: le Cassie and van Houwelingen (1991) have shown this bad behaviour through differrent simulations.

4. A solution based on partitioning the covariate space

The grouping strategy based on partitioning the covariate space nto distinct regions is a way to deal with these problems. The idea is that units with similar covariate pattern have the same probability of having $Y = 1$. By means of a clustering algorithm the covariate space is divided into W distinct regions to form a matrix $Z_{(n \times W)}$ where the element z_{iw} $(i = 1,\ldots,n;\ w = 1,\ldots,W)$ is equal to one or zero depending on whether the i-th unit belongs to the w-th region or not. Consider the model

$$logit(\pi) = \gamma'\mathbf{z} + \beta'\mathbf{x} \qquad (4.1)$$

The null hypothesis to test is $H_0: \gamma_1 = \gamma_2 = \ldots \gamma_W = 0$ or equivalently

$$H_0: logit(\pi) = \beta'\mathbf{x} \qquad (4.2)$$

This hypothesis can be tested directly (Fienberg and Gong, 1984) using a conditional likelihood ratio test of model (4.2) versus model (4.1). If n is sufficiently larger than W and if these groups are set before, then this statistic should be distributed as a χ^2 with $W-1$ degrees of freedom (Haberman, 1974).

Tsiatis (1980) suggested a similar solution based on a quadratic form of observed values minus expected vaiues in each group. First of all the log-likelihood is determined according to (4.1):

$$l = \sum_i \left[y_i(\beta' \mathbf{x}_i + \gamma' \mathbf{z}_i) - \ln\{1 + \exp(\beta' \mathbf{x}_i + \gamma' \mathbf{z}_i)\} \right] \tag{4.3}$$

where $\mathbf{z}_i$ is the vector. obtained from $\mathbf{Z}$, that refers to unit i. The partial derivatives of l with respect to γ_w are calculated at $\gamma = 0$ and $\beta = \hat{\beta}$

$$\frac{\sum_i^n y_i z_{iw} - \sum_i^n z_{iw} \exp(\hat{\beta}' \mathbf{x}_i)}{\{1 + \exp(\hat{\beta}' \mathbf{x}_i)\}} = o_w - e_w \tag{4.4}$$

and are equal to the difference between the observed vaiue and the expected one for the W-th region. Indicating with as the W-dimensional vector of the partial denvatives $\mathbf{P}' = (\partial l/\partial \gamma_1, \ldots, \partial l/\partial \gamma_W)$, the statistic which is used is the quadratic form:

$$Q = \mathbf{P}' \mathbf{G}^{-1} \mathbf{P} \tag{4.5}$$

where $\mathbf{G}$ is the covariance matrix obtained as

$$\mathbf{G} = \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}' \tag{4.6}$$

with

$$\begin{aligned} A_{ww'} &= -\partial^2 l/\partial \gamma_w \partial \gamma_{w'} \quad (w, w' = 1, \ldots, W) \\ B_{ww'} &= -\partial^2 l/\partial \gamma_w \partial \beta_{w'} \quad (w, = 1, \ldots, W; w' = 1, \ldots, k) \\ C_{ww'} &= -\partial^2 l/\partial \beta_w \partial \beta_{w'} \quad (w, w' = 1, \ldots, k) \end{aligned} \tag{4.7}$$

Under the null hypothesis, the Q statistic is asymptotically distributed as a chi-square with degrees of freedom equal to the rank of the matrix $\mathbf{G}$, that is shown to be equal to $W - 1$.

This test is independent of the number of covariates, it holds under very general regularity conditions, and can be easily interpreted as a "score test" (or Rao test) and therefore is asymptotically equivalent to the Waid test and the likelihood ratio test. The precision of the results is in part offset by the computntions necessary to evaluate the statistic (in particular as regards the determination of the genernlized inverse of $\mathbf{G}$)[5].

The choice of clustering prcdictor variables, instead of grouping the units by the estimated probabilities. is motivated (Landwehr et al., 1984) to the local analysis and the concept of near-neighbours; moreover observations with very similar estimnted probnbilities can dcrive from explanatory variables that are by no means close. Otherwise. not only are there many cluster techniques but none of these appears to offer reliable results when the number of regressors is quite large, unless the sample size is also large. A problem arises that when the number of predictors, K, is high, so each unit tends to be very far from the others in the $\mathbf{X}$ metric, thus the groups obtained may include units with very different values of $\mathbf{X}$ and therefore with very different true probabilities of having $Y = 1$ versus $Y = 0$ that is, for large vaiues of k the groups tend towards randomization with respect to $pr(Y = 1|\mathbf{x})$ (Rubin, 1984). This is a similar problem, although posed in a different way, to the one emphasized in the Hosmer and Lemeshow solution based on the grouping of estimated probabilities.

Whatever solution is adoptcd, there is always a component of subjectivity linked to the number of groups to be considered; this means that criteria cannot be generalized. and each time it is necessary to decide this number on the basis of various considerations. This choice is strongly conditioned by sample characteristics: aggregation on the basis of the explanatory variables cannot

[5] Several variants of score test (or Lagrange inultiplier statistics) have been proposed to reduce some coinputational difficulties and to avoid problems in small size samples. For a general review see Davidsoii and MacKinnon (1983)

exclude a careful exarnination of the variability exhibited within the sample. If the sample is small with respect to the number of regressors involved. the formation of hornogeneous groups is partially cornprornised: no aggregation criterion can guarantee the separation of the sample variability from that involved in the characteristics; it is also very difficult to identify sufficiently nurnerous groups of homogeneous units for the observed characteristics. in particular if some of these are continuous. Aggregation as a function of the estirnated probabilities is also critical if the majority of the estimates are concentrated around a few values: there is a risk of identifying a few groups, or of an excessive disaggregation. enhancing the variability between the subjects.

5. An outline on other approaches

5.1 *Graphical* soluriotz

Landwehr, Pregibon and Shoemaker (1984) prefer to set up a graphic solution rather than an analytical test. still in the same spirit of grouping the observations. According to what has already been developed for the linear regression rnodel (Daniel and Wood, 1980), they suggest to partition the residual deviance into a pure-error component and a lack-of-fit component : if the model has a good fit, the latter component will be srnall. otherwise a high value may be interpreted as a systernatic behavior that the logistic rnodel fails to account for.

If there are repeated values of $\mathbf{X}$ (*i.e.* $m_j > 0$) then the pure-error component is quickly obtained; otherwise. it is necessary to group the units and perform an approxirnate factorization of the deviance. The procedure requires to partition the n units into W clusters with n_w units in each.

On the basis of rnodel(4.1) it is possible to calculate the contribution of each observation to local deviance $d(\hat{\pi}_{iw}, y_{iw})$, where y_{iw} and $\hat{\pi}_{iw}$ are respectively the observed values and the estirnated probabilities for the i-th unit of w-region; then we can calculated the sum of these devìances as $D_w = \sum d(\hat{\pi}_{iw} y_{iw})$.

Reordering the W groups so that $O_i \le O_2\ 1 .. O_w$ where O is a measure of group inhomogeneity, e.g. if a hierarchical algorithm has been used, it is the

height (or distance) in the tree at which the group is formed. Then the estimates of the error cornponent are determined

$$\overline{D}(t) = \sum_{w=1}^{t} D_w \Big/ \sum_{W=1}^{t} (n_w - 1) \qquad (5.1)$$

where $\overline{D}(t)$, t = 1,...W represents the local rnean deviance for the tightest groups.

Finally a plot of $\overline{D}(t)$ versus its degree of freedom is made. Superimposing on this plot the line of the global mean deviance, and observing its position relative to the set of points: one can conclude that there is lack of fit when this line, that shows the variability of the data about the fitted rnodel, is systernatically above the points corresponding to the local mean deviances.

Landwehr, Pregibon and Shoernaker prefer the gmphic solution as they think that it provides more information on the data and their fit to the model: on the other hand, Jennings (1986) has demonstrated with simple examples that the local mean deviance overestimates the global one for π close to 0.5 and underestirnates it for π close to 1 or 0. Thus the order in which the groups are formed has a strong effect on the gmphical results. Moreover, this graphical approach has the problem of how to define precisely when to accept and when to reject the null hypothesis.

5.2 *A general approach: the Brown* test

Brown (1982) proposed a solution where the logistic model is ernbedded in a larger family of parametric models in which the logistic, as a special case, depends upon two additional parameters.

The general familiy of models is defined as (Prentice. 1976)

$$p(\mathbf{x}) = \int_{o}^{G(\mathbf{x})} z^{a-1}(1-z)^{b-1} dz \,/\, B(a,b) \qquad (5.2)$$

where $a, b > 0$, $B(a,b)$ is the beta function and

$$G(\mathbf{x}) = \exp(\beta' \mathbf{x})/(1+\exp(\beta' \mathbf{x})) \tag{5.3}$$

The logistic model corresponds to the parameter values $a = b = 1$.

The assumption of model (5.3) allows a statistical test of the adequancy of fit of the specific logistic model relative to the general parametric model. The null hypothesis is $a = b = 1$ and the statistical procedure is based on the asymptotic distribution of the score statistic for the parameters in the general model.

The log-likelihood of observed data is given by:

$$l = \sum_{i}^{n} (y_i \ln(p(\mathbf{x}_i) + (1 - y_i) \ln(1 - p(\mathbf{x}_i)) \tag{5.4}$$

The score statistic of the parameters are defined, as we have just seen, to be the partial derivates of the log-likelihood with respect to each parameter:

$$\frac{\partial l}{\partial \beta_k} = \sum x_{ik} (y_i - p(\mathbf{x}_i)) \ , \ k = 0, \ldots, K$$

$$\frac{\partial l}{\partial a} = \sum \ (y_i - p(\mathbf{x}_i))(1 + \ln(p(\mathbf{x}_i)) \ / (1 - p(\mathbf{x}_i)) \tag{5.5}$$

$$\frac{\partial l}{\partial b} = \sum \ (y_i - p(\mathbf{x}_i))(1 + \ln(1 - p(\mathbf{x}_i)) \ / p(\mathbf{x}_i)$$

The score statistic are asymptotically jointly normally distributed and the test for adequacy of logistic model can be based on the distribution of $(\partial l / \partial a, \partial l / \partial b)$ under the null hypothesis.

Let be $\mathbf{q} = (q_1, q_2) = (\partial l / \partial a, \partial l / \partial b)$ and $\mathbf{C}$ the estimated covariance matrix of $\mathbf{q}$, the statistic test is

$$T = \mathbf{q}' \mathbf{C}^{-1} \mathbf{q} \tag{5.6}$$

This statistic is asymptotically distributed as a chi square with 2 degrees of fredoom. A large value for this test would indicate that the two additional parametrs in the extended model may be different from one and that the extended model fits better than the logistic one.

The test statistic is relatively easy to compute and does not suffer from the practical limitations of the other procedure, even for continuos covariates and samples as small as 50, the null distribution of the statistic is adequately described by a chi square variate. This procedure constitutes a method for examining the logistic model as providing a reasonable description of the data relative to another model in a general class and as such has power against certain patternd of deviations between the observed and fitted data. On the contrary, this solution has poor power against other patterns with deviations which appear to be randomly distributed throughout the range of response rates. In this situations, the Hosmer and Lemeshow tests should perform better.

5.3 Non parametric approach

Nonparametric regression can be used to assess the relationship between a response and a set of explanatory variables. The idea is to check the validity of the systematic part of the model by comparing a non parametric estimate of the regression curve with a parametric one. In the context of the logistic regression this approach represents a good solution because of the difficulties that arise when applying standard residual-based model checking techniques.

Kernel methods have been firstly used firstly by Copas (1983) to examine graphically the fit in one dimensional problems. A kernel estimation $\tilde{g}(\mathbf{x})$ is calculated and $logit(\tilde{g}(\mathbf{x}))$ is plotted against $\mathbf{x}$: if the model fits well , a straight line at 45° through the origin is obtained.

Azzalini, Bowman, and Hardle (1989) have generalized this approach: they compared the function $\hat{\pi}(\mathbf{x})$ with a kemel estimate $\tilde{\pi}(\mathbf{x})$ by defining a pseudo-likclihood ratio statistic.

$$\sum_i \left[y_i \ln\left\{\frac{\tilde{\pi}(\mathbf{x}_i)}{\hat{\pi}(\mathbf{x}_i)}\right\} + (1-y_i)\ln\left\{\frac{1-\tilde{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)}\right\}\right] \tag{5.7}$$

The significance of the observed value of this statistic is estimated by simulating its distribution under the null hypothesis. The nature of any difference between $\hat{\pi}(\mathbf{x})$ and $\tilde{\pi}(\mathbf{x})$ is assessed by using simulations to construct bands for the non parametric curve under the assumptions that the logistic model is correct.

To build pointwise simulation bands a complete set of simulated responses is derived $\{y_1^*,\ldots,y_n^*\}$ from the fitted model. that is y_i^* has a binomial distribution with probabilty $\hat{\pi}(\mathbf{x}_i)$, and a new non parametric estimation $\tilde{\pi}^*$ is produced using the same smoothing parameter employed forthe original data. This operation is repeated a large number of times, say N. For a given ε. empirical upper and lower $\frac{1}{2}\varepsilon$ percentage points of the $\tilde{\pi}^*$ -s at each design define the simulation bands.

However, a problem occurring in this solutions is the bias in the non parametric estimate. le Cassie and van Houwelingen (1991) proposed a test statistic based on a kemel estimate of the standardized residuals that solves many of the problems with the methods above menlioned.

le Cassie and van Houwelingen (1991) considered a smoothing function of the standardizcd residuals obtained by the kemel estimate of Watson (1964)

$$\tilde{r}(\mathbf{x}) = \frac{\sum_j r(\mathbf{x}_j)\mathbf{H}((\mathbf{x}-\mathbf{x}_j)/h_n)}{\sum_j \mathbf{H}((\mathbf{x}-\mathbf{x}_j)/h_n)} \tag{5.8}$$

$\mathbf{H}(\mathbf{z})$ is the multiplicative kemel function defined for $\mathbf{z}=(z_1,\ldots,z_K)$ as

$$\mathbf{H}(\mathbf{z}) = \prod_{d=1}^{K} H(z_d) \tag{5.9}$$

where H is a one-dimensioiial non negative symmetric bounded kemel function, zero outside a closed interval and normalized according to $\int H(z)dz = 1$ and $\int H(z)^2 dz < \infty$. The parameter h, is the bandwidth which controls the amount of smoothing and $\mathbf{z}_j = \mathbf{x} - \mathbf{x}_j$; thus the smoothed residual is a weighted average of the residuals in the neighborhood of $\mathbf{x}$, where the handwidth determines the size of the region over which the residuals are averaged and the kemel determines the weighting.

It is not difficult to show ihat for each $\mathbf{x}$ the mean of $\tilde{r}(\mathbf{x})$ conditional on the observed values for the covariates X is equal to zero and the variance is

$$\mathrm{var}(\tilde{r}(\mathbf{x})) = \frac{\sum_j \mathbf{H}\left[(\mathbf{x}-\mathbf{x}_j / h_n)\right]^2}{\left\{\sum_j \mathbf{H}\left[(\mathbf{x}-\mathbf{x}_j)/h_n\right]\right\}^2} \tag{5.10}$$

The test statistic T is defined as

$$T = n^{-1}\sum_i \tilde{r}(\mathbf{x}_i)v(\mathbf{x}_i) \tag{5.11}$$

where

$$v(\mathbf{x}_i) = \frac{\left\{\sum_j \mathbf{H}\left[(\mathbf{x}_i - \mathbf{x}_j)/h_n\right]\right\}^2}{\sum_j \mathbf{H}\left[(\mathbf{x}_i - \mathbf{x}_j/h_n)\right]^2} \qquad (5.12)$$

In this statistic each observation gives the same contribution under the null hypothesis.

le Cessie aiid van Houwclingen gave an explicit expression for the mean and the variance of the test statistic [6]; they also studied the asymptotic properties. The distribution of T under the null hypothesis was derived and the cutoff points of T were compared with the cutoff points of the normal distribution obtaining quite good results because the distribution of T has heavier tails. The comparison with a scaled chi-square gave better approximations for the smaller significance levels. Thus with this solution it can be formally specified when the null hypothesis should be accepted and when rejected, since the cutoff points are well approximated under the null hypothesis by its asymptotic normal distribution and even better by the scaled chi-square distribution.

A crucial point is the choice of the bandwidth. The performance of the statistic was determined for different values: if it is chosen too small the statistic has no power and if it is chosen too big all local deviations are smoothed away. The author suggested a bandwidth such that each region over which the residuals are averaged contains approximately $\sqrt{n}$ observations.

In summary, this method has some advantages: it detects deviations of the model in all directions and it doesn't require partitioning the data. However, further researches is needed especially in the case of categorical regressors to improve the statistic test.

[6] When $g(x)$ is known le Cessie and van Houwelingten showed that the exact results are quite simple; in the case of estiinated functions the behaviour of the statistic is asymptotically the same but the effect of estiination is negligibile for finite samples

5.4 An exact *approach*

All the parametric tests described in the previous sections use the corresponding chi-square approximations for the statistics. But when ihe sample size is small and the data are sparse the accuracy of the asymptotic approximations is questionable. Under these circumstances the use of exact inferential procedures would seem to be a better solution.

Exact methods for the logistic model, aiid in general for categorical data, have received particular attention in recent years (see Agresti, 1992 for an complete review). One of the most interesting methods developed to check model fit follows a conditional approach, in which one obtains sampling distribution not dependent on unknown parameters by conditioning on their sufficient statistics (McCullagh, 1985, 1986: Bedrick and Hill, 1990, 1992).

Under modcl (1.4), the vector of sufficieiit statistics for $\boldsymbol{\beta}$ is given by $\mathbf{S} = \mathbf{XY}$ aiid assuming the maximum likelihood estimate as $\hat{\boldsymbol{\pi}}$ the probability distribution of the data $pr(\mathbf{Y}, \boldsymbol{\beta})$, indexed by $\boldsymbol{\beta}$, can be factonzed in the marginal distribution of the stntistic $\mathbf{S}$ and the conditional distribution of the observations given by $\mathbf{S}$:

$$\mathrm{pr}(\mathbf{Y};\boldsymbol{\beta}) = \mathrm{pr}(\mathbf{Y}|\mathbf{S})pr(\mathbf{S};\boldsymbol{\beta}) \qquad (5.13)$$

Following Fisher (1950). inferences about $\boldsymbol{\beta}$ are based just on $pr(\mathbf{Y};\boldsymbol{\beta})$ whereas model checks must also be based on $pr(\mathbf{Y}|\mathbf{S})$.

A model check caii be implemented by specifying a statistic T that quantifies the discrepancy between observed and fitted data, nnd computing a significance level for T on the basis of the conditional distribution. This distribution is

$$pr(\mathbf{Y} = \mathbf{y}|\mathbf{S} = \mathbf{s}) = T_{obs}\prod_{j=1}^{J}\binom{m_j}{y_j} \qquad (5.14)$$

where

$$T_{obs}^{-1} = \sum_{(a_1,\ldots,a_J)' \in A_{obs}} \prod_{j=1}^{J} \binom{m_j}{y_j} \qquad (5.15)$$

and $A_{obs} = \{\mathbf{a} = (a_1,\ldots,a_J)', a_j \text{ integer: } 0 \le a_j \le m_j\}$ with $\mathbf{X'a} = \mathbf{s}$.

A_{obs} is the set of response vectors that give the same value of the ufficient statistic as the observed data[7]

The p-value for testing the model is the conditional probability that the goodness of fit statistic is at least as large as observed, i.e

$$p(T) = pr(T > T_{obs} | \mathbf{S} = \mathbf{s};\ H,) \qquad (5.16)$$

The criterion is the one typical of significance tests: no alternative hypothesis is requested. but the result is provisional and requires further tests.

As supported by mnny researchers, the conditioning by a sufficient statistic remains one of the more reasonable solutions, especially when the sample size is small or the date are sparse. Nevertheless further studies are needed to make exact methods more widely applicable. This a useful topic for future research.

REFERENCES

A. Agresti (1992), *A survey of exact inference for contingency tables*, "Statistical Science". 7, pp.131-153

T. Amemiya (1981). *Qualitative response models: a survey.* "Journal of economic literature",19, pp.1483-1536

A. Azzalini, A.W. Bowman, W. Hardle (1989), *On use of non parametric regression for model checking*,"Biometrika", 76, pp.1-11

E.J. Bedrick, J.R. Hill (1990). *Outlier tests for logistic regression: a conditional approach* "Biometrika". 77, pp. 815-827

E.J. Bedrick, J.R. Hill (1992). *Contribution to the paper by Agresti,* "Statistical Science",7, pp.153-57

[7] THirji et al. (1987) developed an algoritlun to generate the reference set. Bedrick and Hill (1992) suggested an alternative method that can be applied to a range of problems.

J. Berkson (1944), *Application of the logistic function to bio-assay*, "Journal of the American Statistical Association", 39, pp.357-65

J. Berkson (1951), *Wliy I prefer logits to probits*, "Biometrics",7,pp.327-339

C.C Brown (1982), *On a goodness-of-fit test for the logistic model based on score statistics*, "Communications in Statitistics- Theory and Methods", 11, pp. 1087-1105

S. le Cassie, J.C. Houwelingen (1991). *A goodness of fit test for binary regression models, based on smothing method*, "Biometrics",47, pp.1267-1282

J.B Copas (1983), *Plotting p against x*, "Applied Statistics", 32, pp.25-31

D.R Cox. E.J. Snell (1968), *A general definition of residual (with discussion)*, "Jouriial of the Royal Statistical Society", B, 30, pp.248-75

D.R. Cox (1970). The analysis of binary data. Methuen, London

R. Davidson, J.G. MacKinnon (1983), *Small sample properties of alternative forms of the Lagrange multiplier test*, "Economics Letters", 12, pp.269-275

C.Daniel, F.S. Wood (1980). *Fitting equations to data*, John Wiley & sons, New York, 2-nd ed.

A. DeMaris (1992), *Logit modelling. Practical application*, Sage university paper. New York

G.V. Dyke, H.D. Patterson (1952), *Analysis of factorial arrangements when the data ai·e proportions*, "Biometrics", 8, pp.1-12

B. Efron (1978). *Regression and ANOVA with zero-one data: measures of residual variation*, "Jouriial of America Statistical Association", 73, pp.113-121

S.E. Ficnberg. G.D. Gong (1984), *Contribution to the discussion of a paper by J.M. Landwehr, D. Pregibon, A.C. Shoemaker*, "Jouriial of the American Statistical Association", 79, pp.77-82

R.A. Fishcr (1950). *The significance of deviations from expectation in a Poisoon series*, "Biometrics", 6, pp.17-24

S.J. Haberman (1974), *The analysis of frequency data*, University of Chicago Press, Chicago

K. F. Hirji, C.R. Mehta, N. R. Patel (1987), *Computing distribution for exact logistir regression*, "Jouriial of the American Statistical Association", 82, pp.1110-17

D.W. Hosmer, S. Lemeshow (1980). *Goodness of fit tests for the multiple logistic regression model*, "Communications in Statistics, Theory and methods", 10, pp.1043-1069

D.W. Hosmer, S. Lenicshow (1982). The use of goodness of fit statistics in the development of logistic regression models, "American Journal of Epidcmiology", 115, pp.92-106

D.W. Hosmer, S. Lenicshow. J. Klar (1988), *Goodness of fit testing for multiple logistic regression analysis when estimated probabilities nre small*, "Biometrical Journal",30, pp.911-924

D.W. Hosmer, S. Leineshow (1989), *Applied logistic regression.* John Wiley & Sons, New York

J.M. Lnndwehr, D. Pregibon, A.C. Shoemaker (1984). *Graphical methods for assessing logistic regression models*, "Journnl of American Statistical Association", *385, pp.61-83*

T. Laitila (1993), *A pseudo* R^2 *measure for limited and qualitative dependent variable models*, Journal of Econometrics"£, 56, pp.341-356

P. McCullagh, J. A. Nelder (1989), *Generalized linear models*, Chapman and Hall. New York, 2nd cd.

P. McCullagh (1986), *The conditional distribution of goodness of fit statistics for discrete data*, "Journal of the American Statistical Associntion". 81, pp.104-107

D. S. Moore, M.C. Spruill (1975), *Unified large-sample theory of general chi square statistics for test of fit*, "Annals of Mathematical Statistics", 4. pp.147-156

C. R. Rao (1973), *Linear statistical inference and its applications*, (2nd ed.) John Wiley & Sons, New York

Lnndwehr, D. Pregibon, A.C. Shoemaker, "Journal of the Americnn Statistical Associntion", 79, pp.79-80

A.A. Tsintis (1980). A note on a goodness of fit test for the logistic regression model. "Biometrika", 67, pp.250-251

G.S. Watson (1964), *Smooth regression analysis*, "Sankhya", series A. 26, pp.359-372