

Working Paper nº 57

**A Score test for Non-nested Hypotheses
with Applications to Discrete Data Models**

J. M. C. Santos Silva

Working Paper nº 57

ISSN: 0872-895X

Depósito Legal nº: 90631/95

J. M. C. Santos Silva
Instituto Superior de Economia e Gestão
Universidade Técnica de Lisboa

Dezembro 1996

Trabalho apresentado no âmbito do seminário de Estatística e Econometria, ano lectivo
1996/97 - coordenado pelo Professor Doutor Bento Murteira.

A Score Test for Non-nested Hypotheses with Applications to Discrete Data Models

J. M. C. Santos Silva*

ISEG/Universidade Técnica de Lisboa

First draft: February 26, 1996

This version: November 11, 1996

Abstract

This paper suggests that a convenient score test against non-nested alternatives can be constructed from the linear combination of the likelihood functions of the competing models. It is shown that this procedure is essentially a test for the correct specification of the conditional distribution of the variable of interest. As in models for discrete data it is often necessary to fully specify the conditional distribution of the variate of interest, the test proposed here is particularly attractive in this context. The usefulness of the proposed tests is illustrated with applications to discrete choice and count data models.

JEL classification code: C52, Model evaluation and testing.

*Address for correspondence:

ISEG, R. Miguel Lupi 20, 1200 Lisboa, Portugal. Tel/Fax: 351 1 3951885. E-mail: jmcss@iseg.utl.pt.

1. INTRODUCTION

Tests against non-nested alternatives were pioneered by Cox (1961, 1962) and have become quite popular in econometrics [see for example the surveys by MacKinnon (1983) and Gouriéroux and Monfort (1994)].

A convenient way of constructing tests against non-nested alternatives uses the principle of artificial nesting suggested by Cox (1961) and developed by Atkinson (1970). These authors suggested that a test for departures from the null in the direction of a non-nested alternative can be obtained testing the null against a more general model that artificially combines the two competing models. The standard way of obtaining this artificial model [see Cox (1961) and Atkinson (1970)] is to use an exponential combination of the non-nested alternatives. Using this approach, the individual contribution to the likelihood of the compound model that artificially nests both alternatives is:

$$L_e(y|x, \alpha, \beta, \gamma) = \frac{L_1(y|x, \beta)^{(1-\alpha)} L_2(y|x, \gamma)^\alpha}{\int L_1(z|x, \beta)^{(1-\alpha)} L_2(z|x, \gamma)^\alpha dz}, \quad (1)$$

where $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$ denote the individual contributions to the likelihood functions of the competing models and α is the mixing parameter.

This way of combining the likelihoods of non-nested models has gained widespread popularity due to the fact that if both $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$ are normal densities, $L_e(y|x, \lambda, \beta, \gamma)$ is also normal [see Atkinson (1970) and Pesaran (1982)]. This is obviously a very convenient result if the researcher is exclusively interested in Gaussian models. However, if other kinds of models are considered, like the ones studied by Pesaran and Pesaran (1993), Orme (1995) and Weeks (1996), different ways of artificially nesting the competing alternatives may also be of interest.

In this paper, the artificial nesting approach is used to construct a test for the adequacy of distributional assumptions in models for discrete data. As in models of this type it is often important to correctly specify the conditional distribution of the variate of interest, the approach suggested here is likely to be useful in many applications. The approach followed here is slightly different from the one of Cox

(1961, 1962) and Atkinson (1970) in that a linear combination of the likelihoods of the non-nested alternatives is used to construct the artificial alternative. This approach leads to a very simple test and it is argued that, at least in the case of discrete data models, this alternative is intuitively appealing.

The plan of the paper is as follows. Section 2 discusses artificial nesting of non-nested models using linear mixtures. In section 3, the score test for non-nested hypotheses based on different kinds of artificial nesting is obtained. The use of the proposed test in the context of some specific models is discussed in section 4. The application of the proposed test is illustrated in sections 5 and 6 using discrete choice and count data models. Finally, section 7 concludes the paper.

2. LINEAR NESTING

Given two competing non-nested models, it is possible to investigate if the alternative provides information not given by the null testing this model against a more general specification that artificially combines both non-nested alternatives.

Although most tests based on the artificial nesting principle use an exponential combination of the competing alternatives, other compound models are possible. In particular, Atkinson (1970) mentions the possibility of using a linear combination of $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$, instead of (1). However, the author never explores the use of this model and argues that it "does not adequately represent a more general distribution to which the component distributions may be regarded as tentative approximations" [Atkinson (1970), page 352]. The linear mixing was also used by Quandt (1974) to construct a likelihood ratio to compare regression models under the normality hypothesis, but it does not seem to have been explicitly used in the more recent literature.

Using the linear nesting, the individual contribution to the likelihood of the compound model is given by

$$L_c(y|x, \alpha, \beta, \gamma) = (1 - \alpha) L_1(y|x, \beta) + \alpha L_2(y|x, \gamma), \quad (2)$$

where α is a parameter between 0 and 1. Because this convex combination of likelihood functions is still a proper likelihood, this way of artificially nesting the competing alternatives leads to a very simple specification test. This way of nesting the competing models has a number of other important properties that make it attractive.

Model (2) can be interpreted as describing a population that is a mixture of two sub-populations with densities $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$. This interpretation is particularly attractive when working with micro data since individual heterogeneity may result in the existence of sub-populations with different characteristics. Therefore, a test of the null against (2) is intuitively appealing as it can be viewed as a test to check if the null adequately describes all the population.

Clearly, the moments with respect to the origin of $L_c(y|x, \alpha, \beta, \gamma)$ are linear combinations of those of $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$. In particular,

$$E_c(y) = (1 - \alpha) E_1(y) + \alpha E_2(y), \quad (3)$$

where $E_c(y)$, $E_1(y)$ and $E_2(y)$ denote the conditional expectation of y under, respectively, $L_c(y|x, \alpha, \beta, \gamma)$, $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$. The form of (3) suggests that the P test for non-nested hypotheses introduced by Davidson and MacKinnon (1981) can be interpreted as test of the null against $L_c(y|x, \alpha, \beta, \gamma)$, when only the conditional mean is of interest. As in models for discrete data the whole conditional distribution is often of interest, there is no reason to focus only on the conditional mean. However, for the regression models considered by Davidson and MacKinnon (1981) the P test is the adequate test to use.

It is also worth noting that $L_c(y|x, \alpha, \beta, \gamma)$ always lies between $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$, a property not shared by (1), as it is pointed out by Loynes in the discussion of Atkinson's paper. In particular, for a point where $L_1(y|x, \beta) = L_2(y|x, \gamma)$, (1) is not in general equal to its components whereas (2) is.

The use of models which, like (2), are finite mixtures often rises some complications. Specifically, it is well known that the likelihood function of finite mixture models can be unbounded and that sometimes their parameters cannot be identified [see, for

example, the monographs by Everit and Hand (1981) and Titterton, Smith and Makov (1985)]. Examples of these problems in econometric contexts are found in switching regression models of the type studied by Quandt and Ramsey (1978) and in the mass point models pioneered by Heckman and Singer (1984). However, for the purpose that (2) is used here, these problems can largely be ignored. In fact, if y is a discrete random variable the individual contributions to the likelihood function are always bounded as they lay between 0 and 1. As for the identification problem, as long as $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$ are non-nested and depend on a set of regressors, all the parameters of $L_c(y|x, \alpha, \beta, \gamma)$ can generally be identified.

3. THE SCORE TEST FOR NON-NESTED HYPOTHESES

Although (1) and (2) are probably the two most interesting ways of artificially nesting $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$, they are just two special cases in a continuum of possibilities. In fact, under adequate regularity conditions and for any given value of ρ , the competing alternatives can be nested in a model of the form

$$L_\rho(y|x, \rho, \alpha, \beta, \gamma) = \frac{[(1 - \alpha) L_1(y|x, \beta)^\rho + \alpha L_2(y|x, \gamma)^\rho]^{1/\rho}}{\int [(1 - \alpha) L_1(z|x, \beta)^\rho + \alpha L_2(z|x, \gamma)^\rho]^{1/\rho} dz}. \quad (4)$$

It is easily recognized that $\lim_{\rho \rightarrow 0} L_\rho(y|x, \rho, \alpha, \beta, \gamma) = L_e(y|x, \alpha, \beta, \gamma)$ and that for $\rho = 1$ and $0 \leq \alpha \leq 1$, (4) equals (2). Therefore, both (1) and (2) are special cases of (4).

Without loss of generality, let $L_1(y|x, \beta)$ be the null hypothesis. For a given value of ρ , (4) can be used to check the adequacy of $L_1(y|x, \beta)$ by testing $H_0 : \alpha = 0$. Although, in principle this hypothesis can be tested using any of the three classical statistical tests, there are reasons to prefer the use of the score test. First, using a score test it is not necessary to estimate $L_\rho(y|x, \rho, \alpha, \beta, \gamma)$ which is certainly cumbersome and most of the times will add little to the understanding of the problem. Second, it is often assumed that $0 \leq \alpha \leq 1$. Because the values of interest for α lay on the boundary of the parameter space, standard likelihood ratio and Wald tests do not have the usual chi-square distribution. However, the score test maintains its usual

properties. Although the restriction that the mixing parameter is between 0 and 1 is not strictly necessary in (4), it is very convenient specially for the case of the linear combination, that is, when $\rho = 1$.

Using S_α^ρ to denote the individual contribution to the element of the score corresponding to α evaluated under the null, it is possible to write

$$S_\alpha^\rho = \left. \frac{\partial \ln L_\rho(y|x, \rho, \alpha, \beta, \gamma)}{\partial \alpha} \right|_{\alpha=0} = \frac{\left(\frac{L_2(y|x, \gamma)}{L_1(y|x, \beta)}\right)^\rho - 1}{\rho} - \int \frac{\left(\frac{L_2(z|x, \gamma)}{L_1(z|x, \beta)}\right)^\rho - 1}{\rho} L_1(z|x, \beta) dz. \quad (5)$$

It is clear that, under the null, S_α^ρ has zero expectation. However, if $L_1(z|x, \beta)$ is misspecified in some way, the expectation of S_α^ρ will generally be different from zero. Therefore, this kind of test may be powerful even against alternatives other than $L_2(y|x, \gamma)$. Nevertheless, it is possible to find exceptional cases where a test based on (5) will have no power at all. For example, if y is a binary variate and x is composed of the intercept and a dummy variable with only two categories, then, evaluated at their maximum likelihood estimates, the contributions to the likelihood function will be the same, whatever the binary model that is specified. In this case (5) will be identically zero and a simple score test based on it will have no power.

For its characteristics, a test based on (5) is particularly adequate in situations where the full conditional distribution of the variate of interest has to be correctly specified, as in discrete choice or count data models. However, when consistent estimators of the parameters of interest only require the correct specification of low order moments, this kind of test may lead to the rejection of perfectly satisfactory models.

When the exponential combination of the likelihoods of the competing models is used to construct the mixture, that is when ρ passes to zero, the individual contribution to this element of the score can be written as

$$S_\alpha^0 = \ln \left(\frac{L_2(y|x, \gamma)}{L_1(y|x, \beta)} \right) - \int \ln \left(\frac{L_2(z|x, \gamma)}{L_1(z|x, \beta)} \right) L_1(z|x, \beta) dz, \quad (6)$$

which is the centred likelihood ratio on which the Cox (1961, 1962) test is based. On the other hand, if the competing models are artificially nested using the convex combination of their likelihoods, the individual contribution to the element of the

score corresponding to α is

$$S_\alpha^1 = \frac{L_2(y|x, \gamma)}{L_1(y|x, \beta)} - 1. \quad (7)$$

The advantage of using (7) instead of (6) as a basis to construct a test to assess the correct specification of $L_1(y|x, \beta)$ is that the latter is generally much simpler to evaluate. In fact, because a linear convex combination of $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$ is still a likelihood function, the integral on the right hand side of (5) is identically zero when $0 < \alpha < 1$ and $\rho = 1$, and only in this case. Therefore, from all the combinations of the competing models that are possible using (4), the linear convex combination is the only one for which the computation of the score does not require the evaluation of an expectation. Since the evaluation of this expectation has been one of the major impediments to the widespread use of the Cox test [see, for example, Pesaran and Pesaran (1993) and Weeks (1996)], the simplification achieved by using (7) instead of (6) can be quite important.

The problem of using a score test to test $L_1(y|x, \beta)$ against (4) is that, under the null, the parameters of the alternative are not identified. This problem has been noted by Pesaran (1981) for the case in which exponential nesting is used.

It seems natural to evaluate γ at $\hat{\gamma}$, the estimates obtained maximizing $L_2(y|x, \gamma)$, as these provide the best description of the data under the alternative. This is the approach adopted by Cox (1961, 1962). However, valid score tests can be obtained evaluating γ at any arbitrary point that is asymptotically non-stochastic. An appealing alternative is to evaluate γ at $\bar{\gamma}$, a consistent estimate of its probability limit when the null is the true model [see Atkinson (1970), Fisher and McAleer (1981) and Kent(1986)]. This approach is particularly attractive in the construction of score tests since all parameters are evaluated under the null.

Since these procedures are asymptotically equivalent, the choice between them is very much a question of practical convenience and performance in finite samples. Simulation studies investigating this question in the case of exponential mixing suggest that an important determinant of the performance of the test statistics is the difference in the number of parameters under the null and alternative [see Davidson and

MacKinnon (1982) and Godfrey and Pesaran (1983)]. In fact, these studies indicate that unless the null and alternative have the same number of parameters, evaluating γ at its maximum likelihood estimates leads to tests that tend to over-reject the null in finite samples. On the other hand, if γ is evaluated at consistent estimates of their probability limits when the null is true, the test statistics have much better behaviour under the null but their power can drop very sharply. The behaviour in finite samples of the tests constructed using the linear mixing is investigated in section 5 using a very small simulation study.

In order to base a test on (5) it is necessary to obtain the asymptotic covariance matrix of $\frac{1}{\sqrt{N}} \sum S_\theta^\rho$, where S_θ^ρ denotes the individual contribution to the score of $L_\rho(y|x, \rho, \alpha, \beta, \gamma)$ with respect to $\theta' = \{\alpha, \beta'\}$ and the summations are over all N observations. Treating ρ and γ as constants, and using standard results in likelihood theory, it is easy to show that this matrix is given by:

$$\mathcal{V} = \frac{1}{N} \sum E_1 [-\mathcal{H}] = \frac{1}{N} \sum E_1 \begin{bmatrix} [S_\alpha^\rho]^2 & S_\alpha^\rho \left[\frac{\partial \log L_1}{\partial \beta} \right]' \\ \frac{\partial \log L_1}{\partial \beta} S_\alpha^\rho & -\frac{\partial^2 \log L_1}{\partial \beta \partial \beta'} \end{bmatrix},$$

where expectations are taken under the null, β is evaluated at the maximum likelihood estimates obtained under the null and the arguments of the likelihood functions are omitted to simplify the notation. In practice, an estimator of \mathcal{V} is needed. Such an estimator can be obtained as $\hat{\mathcal{V}} = \frac{-1}{N} \sum \mathcal{H}$, where the summations are again performed over all the observations.

Using (5) and $\hat{\mathcal{V}}$, the score test statistic for $H_0 : \alpha = 0$ can be expressed as

$$T_\rho = \frac{[\sum S_\alpha^\rho]^2}{\sum [S_\alpha^\rho]^2 - \sum \left[S_\alpha^\rho \left(\frac{\partial \log L_1}{\partial \beta} \right)' \right] \left[-\sum \frac{\partial^2 \log L_1}{\partial \beta \partial \beta'} \right]^{-1} \sum \left[\frac{\partial \log L_1}{\partial \beta} S_\alpha^\rho \right]},$$

where β is evaluated at the maximum likelihood estimates obtained under the null and γ is evaluated at an arbitrary point that is asymptotically non-stochastic. Assuming suitable regularity conditions [see for example White (1982)], under the null T_ρ is asymptotically distributed as a $\chi_{(1)}^2$ variate.

As for its behaviour under the alternative, given the null and alternative hypothesis, the power of a test based on T_ρ will depend both on the true conditional distribution

of y and on the chosen value of ρ . Although *a priori*, it is not possible to know which value of ρ will lead to the test with better performance, it can be expected that under certain circumstances the choice of ρ will have little impact on the performance of the test. Consider the case in which L_1 is close to L_2 . Expanding S_α^ρ around $\frac{L_1}{L_2} = 1$ leads to

$$S_\alpha^\rho = S_\alpha^1 + o(S_\alpha^1).$$

Of course, the quality of this approximation will depend on the magnitude of both ρ and S_α^1 . However, when L_1 is sufficiently close to L_2 , this approximation will be quite accurate for a wide range of values of ρ . For instance, for $\rho = 0$, if L_1 differs from L_2 by less than 5%, the error in the approximation is inferior to 0.0013. Therefore, in many cases the choice of ρ can be based on computational convenience, rather than on power considerations.

It is obvious that for $\rho = 0$, T_ρ is the Cox (1961, 1962) test statistic. For $\rho = 1$ the following statistic is obtained

$$T_1 = \frac{\left[\sum \frac{L_2}{L_1} - 1 \right]^2}{\sum \left[\frac{L_2}{L_1} - 1 \right]^2 - \sum \left[\left(\frac{L_2}{L_1} - 1 \right) \left(\frac{\partial \log L_1}{\partial \beta} \right)' \right] \left[- \sum \frac{\partial^2 \log L_1}{\partial \beta \partial \beta'} \right]^{-1} \sum \left[\frac{\partial \log L_1}{\partial \beta} \left(\frac{L_2}{L_1} - 1 \right) \right]}.$$

The matrix $-\sum \frac{\partial^2 \log L_1}{\partial \beta \partial \beta'}$ is just an estimator for the covariance matrix of the maximum likelihood estimator of β . The other quantities needed to evaluate T_1 are also easy to compute: $\frac{\partial \log L_1}{\partial \beta}$ is the individual contribution to the score of β evaluated at $\hat{\beta}$, and L_1 and L_2 are the individual contributions to the likelihood function of the two non-nested alternatives. Therefore, a test based on T_1 can be much easier to perform than the Cox test or any other test based on $L_\rho(y|x, \rho, \alpha, \beta, \gamma)$.

It is worth noting that a one sided test for $H_0 : \alpha = 0$ against $H_A : \alpha > 0$ can be based on the signed square root of T_ρ . However, it was pointed out above that this kind of tests may have power even against alternatives that are not in the direction of $L_2(y|x, \gamma)$. Since in this paper it is assumed that the purpose of these tests is to check the adequacy of $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$, and not just discriminate between them, the test based on T_ρ is preferred [see Fisher and McAleer (1979)].

4. TWO SPECIAL CASES

4.1. Models with excess zeros

Although (2) may be just a convenient alternative against which $L_1(y|x, \beta)$ and $L_2(y|x, \gamma)$ can be tested, there are cases in which it is of interest in itself. In fact, some models for discrete data can be viewed as special cases of (2).

In the analysis of discrete data it is often found that the process governing the occurrence of a given realization of the dependent variable (usually the zeros) is different from the mechanism generating the other realizations. Models of this kind were studied, among others, by Gaudry and Dagenais (1979), Mullahy (1986) and Lambert (1992). In these cases a population for which y is a random variable with a logit or Poisson distribution, respectively, is contaminated by individuals for which y is constant.

Consider the case in which the null is $L_1(y|x, \beta)$ but there is an excessive number of zeros. The likelihood function for this model can be written as

$$(1 - \alpha) L_1(y|x, \beta) + \alpha I(y = 0),$$

where $I(\cdot)$ is an indicator function. It is clear that this likelihood is a finite mixture of the form of (2). Therefore, the score test of the null against a zero inflated model can be interpreted as a test against non-nested alternatives of the kind introduced before. Examples of this are the tests of the Poisson model against the ZIP [Lambert (1992)] and of the logit against the Dogit [Gaudry and Dagenais (1979) and Tse (1987)].

In this model the individual contribution to the element of the score corresponding to α is given by

$$\frac{\partial \log L_c}{\partial \alpha} = \frac{L_2}{L_1} - 1 = \left[\frac{I(y = 0)}{P_1(y = 0|x)} - 1 \right],$$

where $P_1(y = 0|x)$ is $P(y = 0|x)$ under $L_1(y|x, \beta)$. This expression shows that, in this case, the test for $H_0 : \alpha = 0$ is essentially a test for the correct specification of $P_1(y = 0|x)$. Even if the true model is of the hurdle type as described in Mullahy (1986), this test may provide useful indications about the need to construct a separate model

for the occurrence of zeros. Mullahy (1990) proposed a related test in the context of count data models. Naturally, in this particular case, the non-nested alternative against which the null is tested is not a credible model for the data. Therefore, this is more a diagnostic test than a test against a non-nested alternative.

4.2. Discrete choice models

The consistency of the maximum likelihood estimator of a discrete choice model depends on the correct specification of the entire conditional distribution of the variate of interest. Moreover, different assumptions about the conditional distribution of the dependent variable lead to very different models that are not nested within each other. Examples of this in the case of binary data are the well known logit, probit and Gumbel (or complementary log-log) models. This is an ideal context in which to use the score test described in the previous section.

The individual contribution to the likelihood of a binary model can be written as

$$[P(y = 1|x)]^y [1 - P(y = 1|x)]^{1-y}.$$

Let P_1 and P_2 denote the specification of $P(y = 1|x)$ under L_1 and L_2 , respectively. Then, the linear combination of the two models leads to

$$L_c = (1 - \alpha) L_1 + \alpha L_2 = (1 - \alpha) [P_1]^y [1 - P_1]^{1-y} + \alpha [P_2]^y [1 - P_2]^{1-y},$$

which, because y is a binary indicator, can be written as

$$L_c = [(1 - \alpha) P_1 + \alpha P_2]^y [1 - (1 - \alpha) P_1 - \alpha P_2]^{1-y}. \quad (8)$$

In this model, the individual contribution to the element of the score vector corresponding to α , evaluated at $\alpha = 0$, can be written as

$$\frac{\partial \log L_c}{\partial \alpha} = \frac{L_2}{L_1} - 1 = \frac{(P_2 - P_1)}{P_1(1 - P_1)} [y - P_1]. \quad (9)$$

Consider now the important case in which L_1 and L_2 depend on x only through an index that is a linear combination of the regressors and parameters. In this case,

under the null $H_0 : \alpha = 0$, the individual contribution to the elements of the score corresponding to the parameters in the index can be expressed as

$$\frac{xP_1'}{P_1(1-P_1)}[y - P_1],$$

where P_1' denotes the derivative of P_1 with respect to the index. From this result, it is easy to recognize that (9) is the score for the parameter associated with a variable defined as $(P_2 - P_1)/P_1'$. Therefore, in this case, the score test for $H_0 : \alpha = 0$ can be interpreted as a score test against a more general model in which the index includes $(P_2 - P_1)/P_1'$ as an additional regressor. Therefore, using the results of Davidson and MacKinnon (1984) on testing for omitted variables in binary models, the relevant test statistic can be easily computed.

In particular, if the model depends on the regressors only through the index and there are no other parameters in the model, the test statistic is just the explained sum of squares of the regression of $\frac{y-P_1}{\sqrt{P_1(1-P_1)}}$ on $\frac{xP_1'}{\sqrt{P_1(1-P_1)}}$ and $\frac{(P_2-P_1)}{\sqrt{P_1(1-P_1)}}$.

This test statistic has been proposed by Davidson and MacKinnon (1993) as a variant of the P test for the case of binary choice models. Although the authors claim that the artificial model against which the null is tested is not "actually a binary response model" [Davidson and MacKinnon (1993), page 528], (8) makes it clear that the model that artificially nests L_1 and L_2 is still a valid binary choice model.

The close relationship between the P test and T_1 in this case is not surprising. In fact, in discrete choice models the conditional distribution of y is defined by its conditional expectation, and it was pointed out in section 2 that the P test can be viewed as a test for $H_0 : \alpha = 0$ in L_c , when attention is focused on the conditional expectation.

Consider now the case in which y has $J > 2$ categories. In this case the individual contribution to the likelihood function can be written as

$$\prod_{j=1}^J P(y_j = 1|x)^{y_j},$$

where y_j is a binary indicator defined as $y_j = 1(y = j)$. Let $P_1(j)$ and $P_2(j)$ denote the specification of $P(y_j = 1|x)$ under L_1 and L_2 , respectively. Then

$$L_c = (1 - \alpha) \prod_{j=1}^J P_1(j)^{y_j} + \alpha \prod_{j=1}^J P_2(j)^{y_j} = \prod_{j=1}^J [P_1(j) + \alpha (P_2(j) - P_1(j))]^{y_j},$$

and

$$\frac{\partial \log L_c}{\partial \alpha} = \frac{L_2}{L_1} - 1 = \sum_{j=1}^J \frac{P_2(j) - P_1(j)}{P_1(j)} [y_j - P_1(j)]. \quad (10)$$

If the model under the null is, for example, a multinomial logit, the form of (10) shows that this test can again be interpreted as the test for an omitted variable. In particular, the test can be viewed as the score test for the omission of $\frac{P_2(j) - P_1(j)}{P_1(j)}$. If the null is an ordered model, the test is equally easy to perform, although it cannot be interpreted as a test for omitted variables.

5. AN ILLUSTRATION USING CAR OWNERSHIP STATUS MODELS

Cramer (1991) uses a sample of 2820 observations collected by the Dutch Central Bureau of Statistics as part of the 1980 household budget survey, to illustrate several aspects of binary and multinomial logit models. Here, the same data is used to illustrate the application of the proposed tests for non-nested hypotheses. For this exercise, the variables used are car ownership status (no car, one used car, one new car and more than one car), log of household income and a categorical variable indicating the degree of urbanization.

Using this data, two binary models for the probability of having at least one car were estimated, a logit and a Gumbel or complementary log-log [see McCullagh and Nelder (1989)]. The results obtained are presented in table 1, together with the statistics to test one model against the other. These statistics were obtained evaluating γ at $\hat{\gamma}$ and $\bar{\gamma}$, and were computed using the artificial regression outlined above.

The results obtained indicate that none of the estimated models can be considered adequate, as they are both clearly rejected either using $T_1(\hat{\gamma})$ or $T_1(\bar{\gamma})$. This result is not at all surprising since it would be difficult for such a simple model to adequately describe the probability of an individual owning a car.

Table 1: Estimation results for binary models

	Logit	Gumbel
Intercept	-20.0398 (1.16374)	-11.1813 (0.633455)
Income	2.01991 (0.111574)	1.09257 (0.059621)
Urbanization	-0.089062 (0.024536)	-0.047304 (0.014567)
Sample size	2820	2820
Log-likelihood	-1617.80	-1640.49
$T_1(\hat{\gamma})$	36.29354	109.73876
$T_1(\bar{\gamma})$	42.50517	91.27342

Asymptotic standard deviations in parentheses.

This data has been used by Cramer (1991), Cramer and Ridder (1991) and Windmeijer (1994) to illustrate several features of the multinomial logit. However, the way the dependent variable is defined suggests that an ordered model might be adequate. To investigate this possibility, a four choices multinomial logit and an ordered probit were estimated and tested against each other using the test procedures discussed above. Table 2 reports the results obtained.

From the analysis of the estimation results, there is little to choose between these models. Both the multinomial logit and the ordered probit lead to estimates of the parameters with reasonable signs and the value of the log-likelihood at the maximum is very similar for the two models.

However, the test statistics presented at the bottom of table 2 suggest that, in this particular case, the ordered probit is preferable. In fact, although the test based $T_1(\hat{\gamma})$ clearly rejects both models, the test based on $T_1(\bar{\gamma})$ leads to the rejection of the multinomial logit but not of the ordered probit (at the usual 5% level). It is interesting to notice that when the null is the multinomial logit both test statistics

are very similar. However, when the null is the ordered probit there is a substantial difference between the two statistics.

Table 2: Estimation results for multiple choice models

	Multinomial logit			Ordered probit
	One used car	One car	Multiple cars	
Intercept	-16.5142 (1.28321)	-23.5703 (1.47742)	-46.9889 (2.73217)	-11.9029 (0.549419)
Income	1.63207 (0.12293)	2.25321 (0.14068)	4.31803 (0.25379)	1.19044 (0.05232)
Urbanization	-0.11453 (0.02734)	-0.05630 (0.03036)	-0.07422 (0.05198)	-0.02788 (0.01214)
Threshold 1				0.96938 (0.02729)
Threshold 2				2.12518 (0.04510)
Sample size		2820		2820
Log-likelihood		-3232.02		-3243.45
$T_1(\hat{\gamma})$		26.42814		19.48519
$T_1(\bar{\gamma})$		25.98236		3.62413

Asymptotic standard deviations in parentheses

5.1. A simulation study

To investigate the behaviour of both $T_1(\hat{\gamma})$ and $T_1(\bar{\gamma})$, a simple simulation study was performed. The models used in this experiment mimic as closely as possible the binary models estimated before. In particular, data was generated from both logit and Gumbel models using as explanatory variables the same regressors used to estimate the models and setting the value of the parameters of the true model equal to the estimates reported in table 1.

The choice of null and alternative hypotheses to compare was motivated by the

desire to study the behaviour of $T_1(\hat{\gamma})$ and $T_1(\bar{\gamma})$ when the null has the same, less or more parameters than the alternative. Thus, besides the logit and Gumbel specifications, the simulations included over-parameterized variants of the basic models obtained by adding to the true specification the square of household income and the product between household income and urbanization. The basic models were also tested against a model in which the probability that the individual has no cars is 1. In this case, because the alternative model has no parameters, the identification problem does not exist and the test statistic is uniquely defined. With this design, 2000 replications were used to compute the rejection frequencies at the nominal 5% level of the tests based on $T_1(\hat{\gamma})$ and $T_1(\bar{\gamma})$. The results obtained are reported in table 3.

The results in table 3 confirm that when the alternatives being tested have the same number of parameters $T_1(\hat{\gamma})$ and $T_1(\bar{\gamma})$ have a very similar behaviour, both under the null and under the alternative. Although both statistics show some tendency to under-reject when the null is the Gumbel model, the tests have reasonable size and good power. Given that, in this case, the two statistics lead to almost indistinguishable tests, the test based on $\hat{\gamma}$ is probably preferable as it is computationally much simpler.

Turning now to the case in which the alternative has no parameters, it is found that the results depend much on the nature of the misspecification. In fact, taking the logit as the null hypothesis, it was found that the test has good size and considerable power when the true model is Gumbel. However, when the Gumbel is the null, the test severely under-rejects and has an estimated power below its nominal size when the true model is the logit (although much larger than the estimated size). These findings are in agreement with results of Tse (1987) and Fry and Harris (1996), which indicate that tests against this alternative can severely under-reject the null and have low power against specifications different from the alternative considered by the test. Notice that in the experiments performed here the model against which the null is being tested is never the true model. Therefore it is not surprising to find that the test has good power in the direction of some alternatives and no power in other directions.

Table 3: Simulation Results

Rejection frequencies at the 5% level

Hypotheses		Test	True model	
Null	Alternative	statistic	Logit	Gumbel
Logit	Gumbel	$T_1(\hat{\gamma})$	0.0460	0.6115
		$T_1(\bar{\gamma})$	0.0450	0.6035
	$P(y = 0) = 1$	T_1	0.0455	0.5500
	O-P Gumbel	$T_1(\hat{\gamma})$	0.1265	0.6920
		$T_1(\bar{\gamma})$	0.0455	0.0730
	O-P Logit	Gumbel	$T_1(\hat{\gamma})$	0.0305
$T_1(\bar{\gamma})$			0.0315	0.0945
Gumbel	Logit	$T_1(\hat{\gamma})$	0.5505	0.0265
		$T_1(\bar{\gamma})$	0.4755	0.0275
	$P(y = 0) = 1$	T_1	0.0325	0.0010
	O-P Logit	$T_1(\hat{\gamma})$	0.5985	0.1110
		$T_1(\bar{\gamma})$	0.2545	0.0400
	O-P Gumbel	Logit	$T_1(\hat{\gamma})$	0.0370
$T_1(\bar{\gamma})$			0.0365	0.0085

Finally, the experiments involving the over-parameterized models confirm that this kind of tests is somewhat unreliable when the models being compared do not have the same number of parameters. In fact, when the null has less parameters than the alternative, $T_1(\hat{\gamma})$ clearly over-rejects the null. On the other hand, $T_1(\bar{\gamma})$ is much better behaved under the null but it has rather low power. These findings are in line with the results of simulation studies of Davidson and MacKinnon (1982) and Godfrey and Pesaran (1983) for the case of exponential mixing. For the cases in which the null has more parameters than the alternative, the two tests have very similar performances. In this case, the tests are clearly undersized and have extremely low power, specially when the null is the Gumbel model. These findings confirm the results obtained when the alternative has no parameters.

Although much more research is needed on this subject, the tentative conclusion of these experiments is that when the models being compared have the same number of parameters, the test based on $T_1(\hat{\gamma})$ has good properties both under the null and the alternative. Therefore, it may be quite useful in several kinds of applications. When the number of parameters of the models being considered is different, these tests should be used with great care. In these cases the tests either have very low power or they reject the null far too often. In particular, when the null has less parameters than the alternative, only the test based on $T_1(\hat{\gamma})$ has good size, but its low power and cumbersome computation reduce its attractiveness. On the other hand, when the null has more parameters than the alternative, the tests can be severely undersized but they might be useful to detect the presence of some types of misspecification.

6. AN ILLUSTRATION USING COUNT DATA MODELS OF THE DEMAND FOR HEALTH CARE

In count data models, the focus of interest is often restricted to the conditional expectation of the count variate. In this case, the consistency of the estimator for the parameters of interest depends only on the correct specification of the conditional mean. Therefore, the choice between competing non-nested specifications of the first conditional moment should be done using the P test of Davidson and MacKinnon (1981), rather than applying the test proposed in section 3.

However, if the sample is truncated or if it is necessary to compute the conditional probabilities of some occurrence, the complete specification of the conditional distribution is needed. Because a wide variety of count data models may have the same specification of the conditional expectation¹, the P test is often unusable. In these cases, the score test introduced in section 3 can be used to check the adequacy of the

¹Besides the traditional Poisson specification, see for example the NEGBIN1 and NEGBIN2 of Cameron and Trivedi (1986), the NEGBIN_k of Winkelmann and Zimmermann (1991) and the generalized Poisson and restricted generalized Poisson regression models of Consul and Famoye (1992) and Famoye (1993).

distribution underlying a given model.

In a recent paper, Pohlmeier and Ulrich (1995) estimate several models for the number of visits to a general practitioner or to a specialist. Although the authors clearly favour hurdle specifications in both cases, here their preliminary estimates of the NEGBIN1 models are used as an illustration of the proposed tests for non-nested hypotheses.

NEGBIN1 models are usually adopted when there is evidence that the conditional variance of the variate is a linear function of the conditional mean. However, the generalized Poisson regression model suggested by Consul and Famoye (1992) also has this characteristic. Therefore, it is interesting to test these models against each other. Also, because the results obtained by Pohlmeier and Ulrich (1995) clearly indicate that the occurrence of zeros needs to be modelled separately, it is interesting to test the NEGBIN1 specification against a zero inflated model of the type described in section 4.

The results of the estimation of the NEGBIN1 and generalized Poisson regressions for the number of visits to general practitioners and specialists are not reported here due to their extension. Table 4 presents the computed test statistics. Since both models have the same number of parameters, all the statistics are computed evaluating γ at $\hat{\gamma}$.

These results suggest that the NEGBIN1 model is not adequate to describe neither the visits to general practitioners nor to specialists. In fact, in both cases the NEGBIN1 model is rejected when tested against the generalized Poisson regression. However, the results obtained do not indicate that the inadequacy of the NEGBIN1 results from an excess of zeros in the sample. On the other hand, the generalized Poisson regression is also clearly rejected in the case of the model for visits to general practitioners but it is not rejected as a model for visits to specialists.

These results support the conclusion of Pohlmeier and Ulrich (1995) that simple NEGBIN1 models are not appropriate in this context and the hurdle specifications suggested by the authors may be the correct solution to this problem.

Table 4: Tests of demand for health care models

Tests	General	
	practitioners	Specialists
NEGBIN1 <i>vs.</i> Zero inflated model	0.73682	1.18232
NEGBIN1 <i>vs.</i> Generalized Poisson	42.0060	28.7199
Generalized Poisson <i>vs.</i> NEGBIN1	30.3349	0.02752
Log-likelihood		
NEGBIN1	-7241.85	-6329.00
Generalized Poisson	-7192.62	-6308.00
Sample size	5096	5096

7. SUMMARY AND CONCLUSIONS

In contradistinction with what happens with standard linear regression, the analysis of discrete data often requires the full specification of the conditional distribution of the variate of interest. Because usually there is little *a priori* information about the shape of such distribution, researchers often have to choose between several seemingly adequate models. Moreover, it is important to test if the preferred specification is indeed adequate. It is well known that these goals can be attained using tests for non-nested alternatives.

In this paper it is suggested that a convenient score test for non-nested alternatives can be constructed from the linear combination of the likelihood functions of the competing models. It is shown that this procedure is essentially a test for the correct specification of the conditional distribution of the variate of interest. Therefore, the test is particularly adequate when the correct specification of the conditional distribution is required. On the other hand, when the inferences to be made do not require such strong parametric assumptions, the test is not adequate as it may lead to the rejection of satisfactory models.

The main problem with the proposed test stems from the fact that under the null

the parameters of the alternative are not identified. Since this problem is shared by other procedures in which the principle of artificial nesting is used, the solutions previously suggested in the literature were adopted here. This point clearly merits further investigation since the results of a small simulation experiment reported in section 5 indicate that the finite sample performance of the tests depends critically on the treatment of this problem.

Finally, the potential usefulness of the proposed tests was illustrated with applications to discrete choice and count data models.

ACKNOWLEDGEMENTS

I am grateful to J. S. Cramer and to the Foundation for Economic Research, University of Amsterdam, for allowing me to work with the Dutch Budget Survey data used in section 5. The data used in section 6 was kindly made available by Winfried Pohlmeier and are from the public use version of the German Socio-economic Panel Study. I thank the Deutsches Institut für Wirtschaftsforschung for authorizing its use. I am most grateful to Andrew Chesher for helpful discussions and for his suggestions on the material in section 3. I am also indebted to Richard Blundell, Costas Meghir, Artur Lopes, Peter Moffatt, Frank Windmeijer and participants at seminars at the University of East Anglia and University College London for helpful comments. This work was partially carried out at Department of Economics of the University College London with the financial support of the Human Capital and Mobility program of the European Commission, which is gratefully acknowledged. The usual disclaimer applies.

REFERENCES

- Atkinson, A. C. (1970). "A Method for Discriminating Between Models". *Journal of the Royal Statistical Society, Series B*, 32, 323-353.
- Cameron A. C. and Trivedi P. K. (1986). "Econometric models based on count data: comparisons and applications of some estimators and tests". *Journal of Applied Econometrics* 1 : 29-53.
- Consul, P. C. and Famoye, F. (1992). "Generalized Poisson Regression Model". *Communications in Statistics - Theory and Methods*, 21(1), 89-109.
- Cox, D. R. (1961). "Tests of Separate Families of Hypotheses", in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, 105-123, University of California Press, Berkeley.
- Cox, D. R. (1962) "Further results on Tests of Separate Families of Hypotheses". *Journal of the Royal Statistical Society, Series B*, 24, 406-424.
- Cramer, J. S. (1991). *The Logit Model*, London, Edward Arnold.
- Cramer, J. S. and Ridder, G. (1991) "Pooling States in the Multinomial Logit Model". *Journal of Econometrics*, 47 267-272.
- Davidson, R. and MacKinnon, J. G. (1981). "Several Tests of Model Specification in the Presence of Alternative Hypotheses". *Econometrica*, 49, 781-793.
- Davidson, R. and MacKinnon, J. G. (1982). "Some Non-Nested Hypotheses tests and the Relations Among Them". *Review of Economics Studies*, 49, 551-565.
- Davidson, R. and MacKinnon, J. G. (1984). "Convenient Specification Tests for Logit and Probit Models". *Journal of Econometrics*, 25, 241-262.

- Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford, Oxford University Press.
- Everit, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*, London, Chapman and Hall.
- Famoye, F. (1993). "Restricted Generalized Poisson Regression Model". *Communications in Statistics - Theory and Methods*, 22(5), 1335-1354.
- Fisher, G. R. and McAleer, M. (1979). "On the Interpretation of the Cox Test in Econometrics". *Economics Letters*, 4, 145-150.
- Fisher, G. R. and McAleer, M. (1981). "Alternative Procedures and Associated Tests of Significance for Non-nested Hypotheses". *Journal of Econometrics*, 16, 103-119.
- Fry, T. R. L. and Harris, M. N. (1996). "A Monte Carlo Study of Tests for the Independence of Irrelevant Alternatives Property", *Transportation Research, Series B*, 30, 19-30.
- Gaudry, M. J. I. and Dagenais, M. G. (1979). "The Dogit Model". *Transportation Research*, 13B, 105-112.
- Godfrey, L. and Pesaran, M. H. (1983). "Tests of Non-Nested Regression Models: Small Sample Adjustments and Monte Carlo Evidence". *Journal of Econometrics*, 212, 133-154.
- Gourieroux, C. and Monfort, A. (1994). "Testing Non-Nested Hypotheses", in Engle, R. F. and McFadden (Eds.), *Handbook of Econometrics*, Vol. IV, North-Holland, Amsterdam.

- Heckman, J. and Singer, B. (1984). "A Method for Minimising the Impact of Distributional Assumptions in Econometric Models for Duration Data ". *Econometrica*, 52, 271-320.
- Kent, J. T. (1986). "The underlying Structure of Nonnested Hypothesis tests". *Biometrika*, 73, 333-343.
- Lambert, D. (1992). "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing". *Technometrics*, 34, 1-14.
- MacKinnon, J. G. (1983). "Model Specification Tests Against Non-Nested Alternatives". *Econometric Reviews*, 2, 85-110.
- McCullagh P. and Nelder J. A. (1989). *Generalized Linear Models*, Second Edition, Chapman and Hall, London.
- Mullahy, J. (1986). "Specification and Testing in Some Modified Count Data Models". *Journal of Econometrics*, 33, 341-365.
- Mullahy, J. (1990). "Detecting Heterogeneity and Overdispersion in Poisson Models via Goodness of Fit", Working Paper, Dept. of Economics, Trinity College.
- Orme, C. (1995). "Non-Nested Tests for Discrete Choice Models". University of Manchester, mimeo.
- Pesaran, M. H. (1981). "Pitfalls of Testing Non-Nested Hypotheses by the Lagrange Multiplier Method". *Journal of Econometrics*, 17, 323-331.
- Pesaran, M. H. (1982). "On the Comprehensive Method of Testing Non-Nested Regression Models". *Journal of Econometrics*, 18, 263-274.

- Pesaran, M. H. and Pesaran, B. (1993). "A Simulation Approach to the Problem of Computing Cox's Statistic for Testing Non-Nested Models". *Journal of Econometrics*, 57, 377-392.
- Pohlmeier, W. and Ulrich, V. (1995). "An Econometric Model of the Two-Part Decision Process in the Demand for Health". *The Journal of Human Resources*, 30, 339-361.
- Quandt, R. E. (1974). "A comparison of Methods for Testing Non-Nested Hypotheses". *Review of Economics and Statistics*, 56, 251-255.
- Quandt, R. E. and Ramsey, J. B. (1978). "Estimating Mixtures of Normal Distributions and Switching Regressions". *Journal of the American Statistical Association*, 73, 730-738.
- Titterton, D. M.; Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Chichester, John Wiley & Sons.
- Tse, Y. K. (1987). "A Diagnostic Test for the Multinomial Logit Model". *Journal of Business & Economics Statistics*, 16, 283-286.
- Weeks, M. (1996). "Testing Binomial and Multinomial Choice Models Using Cox's Non-Nested Test". Paper presented at the ESEM'96, Istanbul.
- White, H. (1982). "Regularity Conditions for Cox's Test of Non-Nested Hypothesis". *Journal of Econometrics*, 19, 301-318.
- Windmeijer, F. A. G. (1994). "A goodness-of-fit test in the multinomial logit model based on weighted squared residuals". *Statistica Neerlandica* 48, 271-283.
- Winkelmann, R. and Zimmermann, K. F. (1991). "A New Approach for Modelling Economic Count Data". *Economics Letters*, 37, 139-143.