mathematics

MDPI

Article

Selection Criteria for Overlapping Binary Models—A Simulation Study

Teresa Aparicio and Inmaculada Villanúa *

Department of Economic Analysis, University of Zaragoza, Gran Vía, 2, 50005 Zaragoza, Spain; aparicio@unizar.es

* Correspondence: villanua@unizar.es; Tel.: +34-976-762213

Abstract: This paper deals with the problem of choosing the optimum criterion for selecting the best model out of a set of overlapping binary models. The criteria we studied were the well-known AIC and SBIC, and a third one called C_2. Special attention was paid to the setting where neither of the competing models was correctly specified. This situation has not been studied very much but it is the most common case in empirical works. The theoretical study we carried out allowed us to conclude that, in general terms, all criteria perform well. A Monte Carlo exercise corroborated those results.

Keywords: binary choice models; overlapping models; Kullback–Leibler distance; discrepancy; information criterion; correctly specified models

check for updates

Citation: Aparicio, T.; Villanúa, I. Selection Criteria for Overlapping Binary Models—A Simulation Study. *Mathematics* **2022**, *10*, 478. https://doi.org/10.3390/math10030478

Academic Editor: Lev Klebanov

Received: 28 December 2021
Accepted: 29 January 2022
Published: 2 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

This work focused on the analysis of model selection criteria within the framework of binary choice models (BCM), where the endogenous variable is binary (Y_i), representing the choice of the decision-maker (i) between two options which are quantified by the values 1 and 0. These models are usually expressed as $p_i = F(x_i'\beta)$, F being the cumulative distribution function (c.d.f.), x_i the regressors vector and p_i the probability that $Y_i = 1$. The c.d.f. can be normal or logistic, leading to a probit model or a logit model, respectively. Although the common analysis procedure for these models is to apply the maximum likelihood estimation (MLE) method, they can also be implemented from a Bayesian framework using Gibbs Sampling Markov Chain Monte Carlo (MCMC) methods [1,2]. Nevertheless, in this work, we considered the conventional context; thus, the MLE procedure was used.

This paper compares several models in order to select the "best" of them. In a general context, and following [3], the compared models can be nested, overlapping and non-nested models. In the specific framework of BCM, two binary models are nested if they possess the same c.d.f. (both probit or both logit) and the regressors of one of the models are included in the other one. Two binary models are overlapping if both possess the same c.d.f. (both probit or both logit) with some common explanatory variables and other specific variables. Finally, the compared models are non-nested if they possess only specific regressors. Moreover, the models are also non-nested when they possess different c.d.f. (probit versus logit), even if there are some common regressors. Many works define nested and non-nested models, and describe the way of working in every situation [4–6], while the overlapping models have been the least analyzed. In this paper, we compared overlapping models, and we found that they were equivalent or that one of them was better than the other (non-equivalent).

Although the hypothesis testing procedures (HTP) are widely used to discriminate between models, we can only use them to choose between pairs of models. In comparison, the selection criteria allowed us to select the best model from quite a large set. This is an important advantage in empirical econometric works. The latter approach allows researchers to express their objectives in the form of a loss function, or by using the

discrepancy concept. As [7] established, the discrepancy concept is a particular case of loss function. For non-linear regression models (our framework), the procedures developed by [3,8,9] belongs to the first category (HTP). The second category involves the well-known AIC [10] and SBIC [11], where the discrepancy was obtained from the Kullback–Leibler distance. Additionally, the use of the mean square error (MSE) of prediction as a discrepancy enabled us to derive another criterion, denoted as C_2 (see [12]).

Many works have studied the behaviour of selection procedures in linear regression models. However, this subject has been less analysed in a non-linear regression context, and the nested framework is nearly always assumed [13–15]. The performance of some selection procedures has also been studied in phylogenetics, where partitioned models were used [16–19]. Specific references for discrete choice models are [12] for nested models, and [20] for non-nested models.

In this paper, the competing models we selected from were overlapping models. The purpose was to investigate the discriminatory power of certain model selection criteria assuming two situations: (i) at least one of the models was correctly specified; (ii) neither of the models was correctly specified. According to [21], a well-specified model can include irrelevant variables together with the set of regressors of the data generating process (DGP). In our opinion, situation (ii) is the most interesting in practice but the least studied in the literature. Given that, in this case, no model was well-specified, we could not consider consistency as the condition that makes a given selection criterion adequate. The requirement we proposed is that the criterion selects the closest model to the DGP.

The article is organised as follows. In Section 2, we establish the general context and the methodology. Section 3 is dedicated to study the theoretical behaviour of the criteria. Section 4 presents and discusses the results from a Monte Carlo experiment. Conclusions are presented in Section 5.

2. Materials and Methods

Consider the following DGP:

$$M0 : p_i = F\left(x_i'\beta^0\right) = F(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \qquad (i = 1, \ldots, N) \tag{1}$$

and a pair of overlapping models which, in general terms, are defined as follows:

$$\begin{array}{l} M1 : p_i = F\left(a_i'\gamma\right) \\ M2 : p_i = F\left(b_i'\delta\right) \end{array} \qquad (i = 1, \ldots, N) \tag{2}$$

where $F(\cdot)$ is the cumulative distribution function (c.d.f), which can be normal or logistic, leading to the probit model or the logit model, respectively. The two competing models have the same c.d.f.; a_i' and b_i' are the $1 \times k_1$ and $1 \times k_2$ explanatory variables vectors of $M1$ and $M2$, for the i-observation; γ and δ are the corresponding parameter vectors. Given the definition of overlapping models, $a' \not\subset b'$ and $b' \not\subset a'$ are satisfied, and both vectors have some common variables.

In order to describe the relationship of each of the competing models with the true model (DGP) we used the Kullback–Leibler distance (KLIC) to the DGP:

$$\text{KLIC}\ (M0,\ M\text{j}) = E_0 \ln \left[\frac{f_0}{f_j}\right] \qquad (j = 1, 2) \tag{3}$$

being f_0 the density function of the DGP and f_j corresponding to model Mj.

From (3), we can write:

$$\text{KLIC}\ (M0,\ M1) - \text{KLIC}(M0, M2) = E_0\ [\ln\ f\ (y \mid b,\ \delta^*\ )\ ] - E_0\ [\ \ln\ f\ (y \mid a,\ \gamma^*)\ ] = E_0[\ell_2^* - \ell_1^*] \tag{4}$$

where γ^* and δ^* are the corresponding pseudo-true parameter vectors (see [22]).

It is well-known that if this statistic (expression (4)) is positive, then $M2$ is the preferred model, $M1$ being preferred if (4) is negative. If it is null, the two models are equivalents.

Given the DGP of (1), and following [21], any model which is correctly specified can be written as:

$$p_i = F\left(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \sum_{j=3}^{k_j} \gamma_j d_{ji}\right)$$

where d_j are additional regressors, including the particular case, where d_j do not exist.

It is worth noting that, for each competing model, the maximum likelihood estimation of the parameter vectors satisfies:

$$\begin{array}{c} \hat{\gamma} \xrightarrow{p} \gamma^* \\ \hat{\delta} \xrightarrow{p} \delta^* \end{array} \tag{5}$$

We can distinguish two cases: the case where at least one of the competing models was correctly specified, and the case where neither of them was correct.

2.1. Case 1: At Least One of the Competing Models Is Correctly Specified

Then, the situations we considered are:

- Case 1.1: Both models were well-specified (or both models included the DGP):

$$\begin{array}{l} M1 : p_i = F\left(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \sum\limits_{j=3}^{k_1} \gamma_j d_{ji}\right) \\ M2 : p_i = F(\delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 z_i) \end{array} \tag{6}$$

  with $z \neq d_j \; \forall j$.

- Case 1.2: Only one of them was well-specified (or only one of them included the DGP):

$$\begin{array}{l} M1 : p_i = F\left(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \sum\limits_{j=3}^{k_1} \gamma_j d_{ji}\right) \\ M2 : p_i = F(\delta_0 + \delta_1 x_{1i} + \delta_2 z_i) \end{array} \tag{7}$$

Let β^{0+} be the parameter vector extended with elements equal to zero in the places corresponding to the variables that are not included in the DGP, that is, $\left(\beta^{0+}\right)' = \left(\beta^0 | 0\right)'$. From the convergence result (5), and according to [23], in case 1.1, the equality $\gamma^* = \delta^* = \beta^{0+}$ held, implying that both models are equivalent. However, in case 1.2 $\gamma^* = \beta^{0+}$ but $\delta^* \neq \beta^{0+}$, $M1$ being better than $M2$.

2.2. Case 2: Neither of the Models Is Correctly Specified (or Neither of the Models Includes the DGP)

In this situation the compared models are:

$$\begin{array}{l} M1 : p_i = F(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 w_i) \\ M2 : p_i = F(\delta_0 + \delta_1 x_{2i} + \delta_2 w_i) \end{array} \tag{8}$$

We again used the convergence result (5) to conclude that, in this case, $\gamma^* \neq \beta^{0+}$ and $\delta^* \neq \beta^{0+}$, it being possible that the competing models are equivalent or not. Specifically, according to [3], there are two possible situations:

- Case 2.1: $f(y_i; \gamma^*, a_i) = f(y_i; \delta^*, b_i)$ that is, the density functions of y_i in $M1$ and $M2$, evaluated at the corresponding pseudo-true parameter vectors, were observationally identical. It implies that $M1$ and $M2$ are equivalent specifications.
- Case 2.2: $f(y_i; \gamma^*, a_i) \neq f(y_i; \delta^*, b_i)$. In this situation the models can be:
  - (a) Equivalent, which means that $E_0[\ell_1^*] = E_0[\ell_2^*]$.
  - (b) Non-equivalent, or $E_0[\ell_1^*] \neq E_0[\ell_2^*]$.

Now, we present the selection criteria, whose behaviour was the aim of our paper. Specifically, we are discussing the well-known information criteria (IC) of Akaike (AIC)

and Schwarz (SBIC), and another criterion we call C_2. To obtain them, we adopted the discrepancy concept (see [7]). As we can see in [12], "A discrepancy measures the lack of fit between the proposed model and the DGP, in the aspect which the researcher considers the most relevant". Then, the discrepancy for model $M1$ could be written as $\Delta(F_1, F_0)$, and we wished to minimize the "overall discrepancy", expressed as $\Delta\left(F_{\hat{\gamma}}, F_0\right)$, or equivalently $\Delta(\hat{\gamma})$, with $F_{\hat{\gamma}}$ the estimated model M1 (that is, $\hat{p}_i = F(a'_i\hat{\gamma})$). The estimation of the expected overall discrepancy, $\hat{E}_0\Delta(\hat{\gamma})$, constitutes the selection criterion. Details about this procedure can be found in [7] and [12]. We assume two discrepancies, called Δ_1 and Δ_2. The first one is the Kullback–Leibler distance. For a model Mj, it is expressed as $\Delta_1\left(F_j, F_0\right) = KLIC(Mj, M0)$, and leads to the information criteria (IC) AIC and SBIC:

$$IC(\mathrm{Mj}) = -\frac{\hat{\ell}_j}{N} + \frac{K_N(\mathrm{Mj})}{N} \tag{9}$$

where $\hat{\ell}_j$ $(j = 1, 2)$ denotes the log-likelihood of model Mj, evaluated at the corresponding vector of estimates, k_j is the number of parameters of Mj, $K_N(\mathrm{Mj})$ is the correction factor (k_j for AIC and $\frac{k_j \log N}{2}$ for SBIC).

The second discrepancy is the mean square error (MSE) of prediction. For model $M1$, this discrepancy is $\Delta_2(F_1, F_0) = E_0(Y_{N+1} - F(a'_{N+1}\gamma))^2$, with "$N + 1$" indicating an out-sample observation. For any Mj model, and following the previously mentioned procedure, the expression of the criterion is:

$$C_2(Mj) = \frac{SSD_j}{N}\left(1 + \frac{2k_j}{N}\right) \tag{10}$$

and $SSD_j = \sum_{i=1}^{N}\left(Y_i - \hat{F}_{ji}\right)^2$, the squared sum of the differences between the binary variable and the estimated probability with model Mj. The proof of (10) is developed in [24].

The model having the criterion with the lowest value was chosen, so different criteria could have led to a different choice. Nevertheless, we were interested in analysing if the criteria worked well, that is, if the selection was correct, in the sense we define in the following section.

3. Theoretical Results

In this section, we study the theoretical behaviour of the criteria, in order to prove if they perform well. All proofs of the results we present in this section can be seen, in detail, in [24].

We carry out an asymptotic analysis, which needs a set of initial assumptions, the results and definitions that we state below.

Assumption 1. *The x'_i, a'_i and b'_i regressor vectors of the models specified in (1) and (2) are non stochastic. The variables of these vectors have sample means and variances with finite limits.*

Lemma 1. *Let be y_i a variable, which is not i.i.d., but heterogeneous (non-identical means and non-identical variances). Then:*

$$N^{-1}\sum_{i=1}^{N} a(y_i, \widetilde{\theta}) \xrightarrow{p} E\left[\frac{1}{N}\sum_{i=1}^{N} a(y_i, \theta_0)\right] \tag{11}$$

Proof. The proof of (11) is based on the law of large numbers for heterogeneous variables, together with a lemma of [25].

The law of large numbers for heterogeneous variables is expressed in the following terms [26]: "Let the sequence $\{y_i - \mu_i\}$ be independent with $E(y_i - \mu_i) = 0$. If $E|y_i - \mu_i|^{1+\delta} \leq B < \infty \ \forall i$ with $\delta > 0$, then $\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_i) \xrightarrow{p} 0$".

The lemma of [25] (p. 2156) is expressed as follows: "If z_i is i.i.d., a (z, θ) is continuous at θ_0 with probability one, and there is a neighbourhood Γ of θ_0 such that $E[\sup_{\theta\in\Gamma}\|a(z,\theta)\|] < \infty$, then for any $\widetilde{\theta} \xrightarrow{p} \theta_0$, $N^{-1}\sum_{i=1}^{N} a(z_i, \widetilde{\theta}) \xrightarrow{p} E[a(z, \theta_0)]$". This lemma, together with the law of large numbers, allows us to write (11). □

Definition 1. *M1 and M2 are equivalent models if*

$$E_0\left[\log\frac{f(y_i; a_i, \gamma^*)}{f(y_i; b_i, \delta^*)}\right] = 0 \tag{12}$$

which leads to:

$$\left[F_{0i}\log\frac{F_{1i}}{F_{2i}} + (1 - F_{0i})\log\frac{1 - F_{1i}}{1 - F_{2i}}\right] = 0 \tag{13}$$

where $F_{0i} = F(x_i'\beta)$, $F_{1i} = F(a_i'\gamma^*)$ *and* $F_{2i} = F(b_i'\delta^*)$.

Definition 2. *M1 is closer to the DGP than M2 if:*

$$E_0\left[\log\frac{f(y_i; a_i, \gamma^*)}{f(y_i; b_i, \delta^*)}\right] > 0 \tag{14}$$

which leads to:

$$\left[F_{0i}\log\frac{F_{1i}}{F_{2i}} + (1 - F_{0i})\log\frac{1 - F_{1i}}{1 - F_{2i}}\right] > 0 \tag{15}$$

Definition 3. *Let* $R(\cdot)$ *be a model selection criterion. It is said that* $R(\cdot)$ *is adequate when:*

(i) If M1 and M2 are equivalent, plim [R(M1)] = plim [R(M2)].
(ii) If M1 is closer than M2 to the DGP, then plim [R(M1)] < plim [R(M2)].

Now, for every case we enuntiated in the previous section, we must prove whether the definition of "adequate criterion" is satisfied.

Result 1. *The IC criteria behave well in all settings.*

Proof. The basic tool for achieving this result is the comparison of Definitions 1 and 2 with Definition 3. In this sense, Definitions 1 and 2 establish the condition that must be met when the compared models are equivalent or non-equivalent, respectively. On the other hand, Definition 3 tells us the requirements for determining if a specific criterion is adequate in each context (of equivalence or not).

Expression (9) can be written as:

$$IC(Mj) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log\hat{F}_{ji} + (1 - y_i)\log\left(1 - \hat{F}_{ji}\right)\right] + \frac{K_N(Mj)}{N} \quad j = 1, 2 \tag{16}$$

with $\hat{F}_{1i} = F(a_i'\hat{\gamma})$ and $\hat{F}_{2i} = F(b_i'\hat{\delta})$.

Using Lemma 1 and the convergences given in (5) for the first term, we obtain:

$$\frac{\hat{\ell}_j}{N} \xrightarrow{p} \frac{1}{N}\sum_{i=1}^{N} E_0\left[y_i\log F_{ji} + (1 - y_i)\log\left(1 - F_{ji}\right)\right] \quad j = 1, 2 \tag{17}$$

The correction factor $\frac{K_N(Mj)}{N}$ converges to zero for every model.

When the competing models are equivalent (cases 1.1, 2.1 and 2.2.(a)), the IC criteria will be adequate if equality of Definition 3 (i) holds, which, using (17), leads to the following expression:

$$\frac{1}{N}\sum_{i=1}^{N}\left[F_{0i}\log\frac{F_{1i}}{F_{2i}}+(1-F_{0i})\log\frac{(1-F_{1i})}{(1-F_{2i})}\right]=0 \tag{18}$$

This result is always satisfied, given Definition 1, so we can say that the IC criteria performed well.

When the models are non-equivalent, and assuming $M1$ is always better than $M2$ (cases 1.2 and 2.2.(b)), the IC criteria will be adequate if equality of Definition 3 (ii) holds, which, using (17), leads to:

$$\frac{1}{N}\sum_{i=1}^{N}\left[F_{0i}\log\frac{F_{1i}}{F_{2i}}+(1-F_{0i})\log\frac{(1-F_{1i})}{(1-F_{2i})}\right]\ >\ 0 \tag{19}$$

this result being identical to Definition 2. Thus, the IC criteria performed well. □

It should be noted that $\frac{K_N(Mj)}{N}$ converges to zero faster for the model with a lower k_j. Additionally, expression (19) shows that AIC and SBIC are asymptotically identical, which is not strictly true when $k_1 \neq k_2$. In this situation, for a given pair of competing models, the difference between both criteria is due to the different rate of convergence to zero between $AIC(M1)-AIC(M2)$ and $SBIC(M1)-SBIC(M2)$. This difference is caused by the corrector factor.

Specifically, we can write:

$$\frac{SBIC(M1)-SBIC(M2)}{AIC(M1)-AIC(M2)}=\frac{O\left(\frac{\log N}{N}\right)}{O\left(\frac{1}{N}\right)}=O(\log N) \tag{20}$$

which means that, when N increases, the distance between the convergence rate of the numerator and the denominator of expression (20) becomes larger. This implies that SBIC will tend toward one of the models more than AIC. Which model? It is evident that, if $k_1 > k_2$, the tendency will be toward M2, given that both AIC and SBIC selects the model with a lower value of the criterion. It is important to remark that $plim(IC(M1)-IC(M2))=0$ is not contradictory with a higher tendency to the model that is more parsimonious, given that both models are equivalent.

Result 2. *The C_2 criterion is adequate, except in a specific situation.*

Proof. Expression (10) can be written as:

$$C_2(Mj)=\frac{\sum_{i=1}^{N}\left(y_i-\hat{F}_{ji}\right)^2}{N}\left(1+\frac{2k_j}{N}\right)\quad j=1,\ 2 \tag{21}$$

Applying Lemma 1 together with convergences (5) we obtain:

$$\frac{SSD_j}{N}=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{F}_{ji}\right)^2\xrightarrow{p}\frac{\sum_{i=1}^{N}E_0\left(y_i-F_{ji}\right)^2}{N}\quad j=1,\ 2 \tag{22}$$

Additionally, the term $\left(1+\frac{2k_j}{N}\right)$ (j = 1, 2) converges to 1 when $N\to\infty$.

When the compared models are equivalent and well-specified (case 1.1), the probability limit (22) is the same for both models. It implies that Definition 3 i) is satisfied, in other words, the C_2 criterion performed well. Note that the convergence rate is different between

$M1$ and $M2$ when $k_1 \neq k_2$. The term $\frac{2k_j}{N}$ is $O\left(\frac{1}{N}\right)$ and converges to zero faster for models with a lower k_j.

When the compared models are non-equivalent and only $M1$ is well-specified (case 1.2), the probability limit (22) is different for each model:

$$\frac{SSD_1}{N} \xrightarrow{p} \frac{\sum_{i=1}^{N} F_{0i}(1-F_{0i})}{N} = h_1 \tag{23}$$

$$\frac{SSD_2}{N} \xrightarrow{p} \frac{\sum_{i=1}^{N} F_{0i}(1-F_{0i})}{N} + \frac{\sum_{i=1}^{N} (F_{0i}-F_{2i})^2}{N} = h_1 + h_2 \tag{24}$$

being h_1 and h_2 positive terms. It is straightforward to see that Definition 3 (ii) is satisfied, so the C_2 criterion performed adequately.

If neither of the competing models is correctly specified (case 2), the probability limits of (22) for each model can be written as:

$$\frac{SSD_1}{N} \xrightarrow{p} \frac{\sum_{i=1}^{N} F_{0i}\,(1-F_{0i})}{N} + \frac{\sum_{i=1}^{N} [F_{0i}-F_{1i}]^2}{N} = h_1 + h_3 \tag{25}$$

$$\frac{SSD_2}{N} \xrightarrow{p} \frac{\sum_{i=1}^{N} F_{0i}\,(1-F_{0i})}{N} + \frac{\sum_{i=1}^{N} [F_{0i}-F_{2i}]^2}{N} = h_1 + h_2 \tag{26}$$

with h_i (i = 1, 2, 3) being positive constants.

Now, the final conclusions depend on the relationship between the density functions. Then, in Case 2.1, where the density functions were observationally identical (equivalent models), $F_{1i} = F_{2i}$ is satisfied. It implies that $h_3 = h_2$, so Definition 3 is verified, and the C_2 criterion behaved well.

In Case 2.2. (a), with non-observationally identical density functions and equivalent models, the only possibility for achieving $h_3 = h_2$ is that, on average, $[F_{0i} - F_{1i}] = -[F_{0i} - F_{2i}]$, or, equivalently, $2F_{0i} = F_{1i} + F_{2i}$. Therefore, there can be empirical works where the criterion C_2 did not behave well. The Monte Carlo experiment will allow us a more specific analysis of the behaviour of the criterion.

Finally, in Case 2.2.(b), where the competing models were non-equivalent, we assumed that M1 was better than M2. In order to study the power of the criterion, we applied a strategy similar to that used in the IC criteria. That is, we related Definitions 2 and 3 (ii). Definition 2 can be written as:

$$\left(\frac{F_{1i}}{F_{2i}}\right)^{F_{0i}} > \left(\frac{1-F_{2i}}{1-F_{1i}}\right)^{1-F_{0i}} \tag{27}$$

We wanted to find the combinations of F_{0i}, F_{1i} and F_{2i} that satisfy (27). The results we obtained are summarized in Table 1.

Table 1. Combinations of F_0, F_1 and F_2 which leads to $M1$ be better than $M2$.

Value of F_0	Condition Satisfied When $M1$ Is Better than $M2$
$F_{0i} = 0$	$F_{1i} < F_{2i}$ and $F_{1i} \neq 0$
$F_{0i} \in (0, 0.5]$	$F_{1i} < F_{2i}$ and $F_{1i} + F_{2i} > 1$ $F_{1i} > F_{2i}$ and $F_{1i} + F_{2i} < 1$ and $F_{0i} \cong 0.5$ $F_{1i} < F_{2i}$ and $F_{1i} + F_{2i} < 1$ and $F_{0i} \approx 0$
$F_{0i} \in (0.5, 1)$	$F_{1i} > F_{2i}$ and $F_{1i} + F_{2i} < 1$ $F_{1i} > F_{2i}$ and $F_{1i} + F_{2i} > 1$ and $F_{0i} \approx 1$ $F_{1i} < F_{2i}$ and $F_{1i} + F_{2i} > 1$ and $F_{0i} \approx 0.5$
$F_{0i} = 1$	$F_{1i} > F_{2i}$ and $F_{1i} \neq 1$

Definition 3 (ii) establishes that the C_2 criterion behaves well if $h_3 < h_2$, that is to say:

$$\frac{\sum_{i=1}^{N} [(F_{1i} - F_{2i})(F_{1i} + F_{2i} - 2F_{0i})]}{N} < 0 \tag{28}$$

For every combination presented in Table 1, we get the previous result, so the C_2 criterion is adequate. □

4. Simulation Study and Discussion

The objective of the Monte Carlo experiments is twofold: confirm the theoretical results and assess the performance of all criteria with finite samples.

The generation of the binary variable y_i is based on the latent linear model that underlies any binary model:

$$y_i^* = x_i'\beta + u_i \tag{29}$$

where y_i^* is a latent (unobservable) variable which generates y_i through:

$$y_i = \begin{cases} 1 & if \quad y_i^* > 0 \\ 0 & if \quad y_i^* \leq 0 \end{cases} \tag{30}$$

Under the assumption established in Section 3, and following the procedure of [27], we obtained the values of y_i. We considered different sets of parameter values and different kinds of explanatory variables (continuous and dummy) and the standard normal distribution function was chosen for the error term, implying exclusive focus on probit models. Two sample sizes N = 200 and 2000 were used; we carried out 500 replications for each experiment. Additionally, the intercept was fixed at a value of -2, in order to avoid a non-balanced number of ones in the sample of y_i, which would lead to problems when estimating and interpreting results.

In each of the 500 replications, we estimated M1 and M2, and calculated the value of the IC and C_2 criteria in each replication. The corresponding tables for every experiment show the number of times that each criterion selected M1. Note that we only present tables for N = 2000 and comment the differences from N = 200 if such differences exist. In all cases, the DGP is $p_i = F(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$.

4.1. Montecarlo Exercise When Both Models Are Correctly Specified (Case 1.1)

We consider the following well-specified models $M1$ and $M2$:

$$M1 \equiv p_i = F(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 w_i + \gamma_4 s_i)$$
$$M2 \equiv p_i = F(\delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 z_i)$$

Firstly, we assumed $\gamma_4 = 0$, so we chose between models with the same number of parameters and, afterwards, we assumed $\gamma_4 \neq 0$. These settings are called A and B, respectively; the results are presented in Table 2.

Table 2. Behaviour of the selection criteria in Case1.1 N = 2000.

	DGP				Specified Models				
	β'	x_1	x_2	w	z	s	**AIC**	**SBIC**	C_2
Setting A	(−2,3,1)	U	U	U	N(3,1)		273	273	275
	(−2,1,3)	U	U	U	N(3,1)		236	236	240
	(−2,1,1)	U	U	U	N(3,1)		234	234	244
	(−2,3,1)	χ^2	χ^2	U	N(3,1)		240	240	252
	(−2,1,3)	χ^2	χ^2	U	N(3,1)		273	273	261
	(−2,1,1)	χ^2	χ^2	U	N(3,1)		236	236	238
	(−2,3,1)	U	χ^2	U	N(3,1)		247	247	245
	(−2,1,3)	U	χ^2	U	N(3,1)		231	231	245
	(−2,1,1)	U	χ^2	U	N(3,1)		243	243	260
	(−2,3,1)	U	*dummy*	U	N(3,1)		259	259	269
	(−2,1,3)	U	*dummy*	U	N(3,1)		231	231	245
	(−2,1,1)	U	*dummy*	U	N(3,1)		242	242	261
	(−2,3,1)	χ^2	*dummy*	U	N(3,1)		255	255	251
	(−2,1,3)	χ^2	*dummy*	U	N(3,1)		248	248	257
	(−2,1,1)	χ^2	*dummy*	U	N(3,1)		251	251	234
Setting B	(−2,3,1)	U	U	U	N(3,1)	χ^2	144	10	179
	(−2,1,3)	U	U	U	N(3,1)	χ^2	133	10	151
	(−2,1,1)	U	U	U	N(3,1)	χ^2	126	2	144
	(−2,3,1)	χ^2	χ^2	U	N(3,1)	χ^2	128	12	215
	(−2,1,3)	χ^2	χ^2	U	N(3,1)	χ^2	140	10	230
	(−2,1,1)	χ^2	χ^2	U	N(3,1)	χ^2	130	11	185
	(−2,3,1)	U	χ^2	U	N(3,1)	χ^2	128	9	165
	(−2,1,3)	U	χ^2	U	N(3,1)	χ^2	106	2	192
	(−2,1,1)	U	χ^2	U	N(3,1)	χ^2	120	5	161
	(−2,3,1)	U	*dummy*	U	N(3,1)	χ^2	144	12	177
	(−2,1,3)	U	*dummy*	U	N(3,1)	χ^2	145	7	153
	(−2,1,1)	U	*dummy*	U	N(3,1)	χ^2	141	7	162
	(−2,3,1)	χ^2	*dummy*	U	N(3,1)	χ^2	147	10	200
	(−2,1,3)	χ^2	*dummy*	U	N(3,1)	χ^2	136	8	185
	(−2,1,1)	χ^2	*dummy*	U	N(3,1)	χ^2	143	12	185

For setting A, we can see that the number presented in each cell was around 250 (50% of 500 times), which is the correct behaviour for equivalent models. However, if we consider setting B, where $M1$ had more irrelevant regressors than $M2$, we could see that the criteria tended to select the model with fewer parameters, that is, it selected the more parsimonious model ($M2$), and that this tendency grew with the sample size. This behaviour was also correct given that both specifications were equivalent. Specifically, the most parsimonious criterion is SBIC, so the Monte Carlo exercise corroborated this theoretical aspect of the previous section. Additionally, we observed that neither the kind of variables nor the set of values of the DGP parameter vector seemed to affect the behaviour of the criteria.

4.2. Montecarlo Exercise When Only One of the Models Is Correctly Specified (Case 1.2)

In these experiments, $M1$ and $M2$ are expressed as follows:

$$M1 \equiv p_i = F(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 w_i)$$
$$M2 \equiv p_i = F(\delta_0 + \delta_1 x_{1i} + \delta_2 z_i)$$

The theoretical results were corroborated, so all the criteria tended to select M1 for whatever kind of explanatory variables. The corresponding table has been omitted, given that the value in all cells was 500.

However, for N = 200 the results were not so evident, although they tended towards adequate behaviour. Specifically, differences were found when the variables x_1 and x_2 were both uniform and the weight of x_2 was not greater than that of x_1; this difference was more evident for the SBIC criterion.

4.3. Montecarlo Experiment When Neither of the Models Is Correctly Specified (Case 2)

We needed to analyse each of the situations defined in Case 2 of Section 2 separately. The two compared models are:

$$M1 \equiv p_i = F(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 w_i)$$
$$M2 \equiv p_i = F(\delta_0 + \delta_1 x_{2i} + \delta_2 w_i)$$

The implementation of the experiments for 2.1 and 2.2.(a) required using the relationship between the true parameter vector (β^0) and each of the pseudo-true parameter vectors (γ^* and δ^*). In case 2.1, we should have been able to obtain the value of β^0 from them, satisfying the equality of density functions. In other terms, if we had $\gamma^* = m_1(\beta^0)$ and $\delta^* = m_2(\beta^0)$, we were interested in obtaining β^0 that makes $f(Y_i; m_1(\beta^0), a_i) = f(Y_i; m_2(\beta^0), b_i)$. The same idea could be used in 2.2.(a) in order to make $E(\ell_1^*)$ and $E(\ell_2^*)$ equal, that is, $E[\ell_1(m_1(\beta^0))] = E[\ell_2(m_2(\beta^0))]$. Nevertheless, the non-linear equation system that we needed to solve possessed insurmountable problems. Given that we could not obtain the exact relationship, we needed to approximate the equalities of densities and likelihoods. To this end, we generated several DGPs modifying the value of the parameter vector β^0 and the kind of explanatory variables. Again, the intercept was fixed at a value of -2, while β_1 and β_2 took values in the range of $(-2, 2)$ counting by 0.5 s; additionally, they took values 3, 4, 5 and 7. As a result of this strategy, we generated 132 different DGPs. Each of the outlined DGPs lead to a specific relationship between models $M1$ and $M2$: equivalent models (with identical or non-identical densities) or non-equivalent models.

In order to classify the 132 experiments into the two categories, we used the following indicators:

$$absel = \frac{1}{N}\sum_{i=1}^{N}|del_i| \tag{31}$$

$$\text{AM1} \equiv \text{times (of the N observations) that } del_i > 0 \tag{32}$$

$$absdifden = \frac{1}{N}\sum_{i=1}^{N}|difden_i| \tag{33}$$

with $del_i = E(\ell_{1i}^*) - E(\ell_{2i}^*)$ and $difden_i = f(y_i; \gamma^*, x_1, x_2) - f(y_i; \delta^*, x_1, x_2)$.

Firstly, we classified the experiments into "containing equivalent models", or "containing nonequivalent models". Secondly, in the first group, we distinguished identical from non-identical densities. Finally, we classified the non-equivalent models depending on their closeness to the DGP. The following three stages were carried out:

Step 1. The two requirements for considering the models as equivalents are:
- (R.1) A value of *absel* close to zero.
- (R.2) A value of AM1 close to $N/2$.

Taking into account *absel*, two models will tend to be equivalent if, for most of the observations, $del_i \approx 0$, which should lead to $absel \approx 0$. Could we have used only this measure to affirm that the models were equivalent? The answer is no, because we could find $absel \approx 0$ but with most of the observations satisfying $E(\ell_{1i}^*) > E(\ell_{2i}^*)$, which means that M1 was closer to the DGP. Using $AM1$ instead of *absel*, two models will tend to be equivalent if $AM1 \approx N/2$. Could we have used only *AM1* to classify the models? Again,

the answer is no, because it could happen that $AM1 \approx N/2$ but with a large value of *absel*, due to large values of $|del_i|$. Then, we need the two requirements (R.1) and (R.2).

Step 2. To consider that two equivalent models have identical densities, a value of *difden* close to zero is required.

Step 3. If model $M1$ is better than $M2$, $N/2 < AM1 < N$ must be satisfied, while $M2$ will be better if $0 < AM1 < N/2$.

Each experiment was numbered from 1 to 132, and was classified in type A, B or C, as we can see in Table 3.

Table 3. Number of every experiment in Case 2, and types of x in the DGP (A,B,C) [1].

Number	DGP	Number	DGP	Number	DGP	Number	DGP
A,B,C	(β_1,β_2)	**A,B,C**	(β_1,β_2)	**A,B,C**	(β_1,β_2)	**A,B,C**	(β_1,β_2)
1,45,89	(1,−2)	12,56,100	(1,7)	23,67,111	(3,5)	34,78,122	(7,1)
2,46,90	(1,−1.5)	13,57,101	(3,−2)	24,68,112	(3,7)	35,79,123	(−2,3)
3,47,91	(1,−1)	14,58,102	3,−1.5)	25,69,113	(−2,1)	36,80,124	(−1.5,3)
4,48,92	(1,−0.5)	15,59,103	(3,−1)	26,70,114	(−1.5,1)	37,81,125	(−1,3)
5,49,93	(1,0.5)	16,60,104	(3,−0.5)	27,71,115	(−1,1)	38,82,126	(−0.5,3)
6,50,94	(1,1)	17,61,105	(3,0.5)	28,72,116	(−0.5,1)	39,83,127	(0.5,3)
7,51,95	(1,1.5)	18,62,106	(3,1)	29,73,117	(0.5,1)	40,84,128	(1.5,3)
8,52,96	(1,2)	19,63,107	(3,1.5)	30,74,118	(1.5,1)	41,85,129	(2,3)
9,53,97	(1,3)	20,64,108	(3,2)	31,75,119	(2,1)	42,86,130	(4,3)
10,54,98	(1,4)	21,65,109	(3,3)	32,76,120	(4,1)	43,87,131	(5,3)
11,55,99	(1,5)	22,66,110	(3,4)	33,77,121	(5,1)	44,88,132	(7,3)

[1] In type A both variables are U(0,1), in B $x_1 \sim$ U(0,1) and $x_2 \sim \chi^2_1$, and C has both variables χ^2_1.

Those experiments with an extreme percentage of zeros in the sample of the binary variable were omitted. Table 4 presents the experiments with the lowest value of *absel*, the values of $AM1$ closer to 1000, and the values of *absdifden* near zero.

Table 4. Experiments sequenced by each of the three indicators (*absel*, *AM1* and *absdifden*).

Experiments with the Lowest Value of *absel*			**Experiments with *AM*1 around 1000**			**Experiments with the Lowest Value of *absdifden***	
Exp.	***absel***	***AM*1**	**Exp.**	***AM*1**	***absel***	**Exp.**	***absdifden***
3	0.00716535	1001	57	1011	0.17489579	27	0.0224
27	0.00716886	1003	27	1003	0.00716886	3	0.0227
28	0.00731291	505	3	1001	0.00716535	28	0.0257
4	0.00737039	1498	6	1000	0.01854382	4	0.0261
5	0.01279404	1514	63	995	0.25329468	48	0.0325
29	0.0127966	510	21	991	0.13463303	47	0.0341
48	0.01515962	1035	109	980	0.44042798	46	0.0349
						45	0.0354
						29	0.0510

We concluded that experiments 27, 3 and 6 included equivalent models, 27 and 3 having identical densities. The rest of the experiments corresponded to non-equivalent models, and we needed to classify them according to their closeness to the DGP. Given that AM1 was the adequate indicator, Table 5 shows all the experiments sequenced from the highest to the lowest value of this measure.

Table 5. Behaviour of the selection criteria in Case 2. $N = 2000$ [1].

Exp.	*AM*1	IC	C2	Exp.	*AM*1	IC	C2	Exp.	*AM*1	IC	C2	Exp.	AM1	IC	C2
104	1868	500	500	**20**	1393	500	500	**129**	805	0	0	**84**	315	0	0
34	1861	500	500	**76**	1366	500	500	**47**	786	2	22	**37**	315	0	0
17	1846	500	500	**119**	1365	500	500	**65**	764	0	0	**50**	261	0	0
16	1834	500	500	**30**	1359	499	499	**112**	719	0	0	**123**	245	0	0
33	1827	500	500	**13**	1342	500	500	**128**	708	0	0	**124**	232	0	0
92	1825	500	500	**107**	1321	500	500	**96**	705	0	0	**51**	226	0	0
103	1809	500	500	**88**	1292	500	500	**22**	701	0	0	**10**	224	0	0
91	1807	500	500	**42**	1288	500	500	**66**	686	0	0	**114**	209	0	0
90	1794	500	500	**118**	1268	500	500	**7**	685	2	2	**125**	208	0	0
102	1779	500	500	**59**	1252	500	500	**46**	668	0	5	**52**	204	0	0
89	1779	500	500	**132**	1219	500	500	**35**	662	0	0	**11**	195	0	0
101	1753	500	500	**108**	1188	500	500	**41**	636	0	0	**113**	194	0	0
32	1751	500	500	**131**	1156	500	500	**67**	601	0	0	**69**	190	0	0
105	1691	500	500	**62**	1154	67	244	**45**	589	0	2	**115**	180	0	0
18	1691	500	500	**94**	1112	441	451	**117**	583	0	0	**53**	170	0	0
15	1675	500	500	**58**	1112	352	494	**97**	576	0	0	**39**	170	0	0
78	1641	500	500	**130**	1093	500	500	**23**	537	0	0	**38**	170	0	0
122	1606	500	500	**87**	1067	447	496	**68**	518	0	0	**116**	167	0	0
44	1582	500	500	**48**	1035	72	195	**98**	513	0	0	**54**	162	0	0
121	1556	500	500	**57**	1011	41	322	**29**	510	1	1	**70**	162	0	0
19	1540	500	500	**27**	1003	245	240	**28**	505	10	13	**126**	161	0	0
120	1538	500	500	**3**	1001	250	245	**8**	504	0	0	**79**	157	0	0
31	1524	500	500	**6**	1000	249	252	**36**	494	0	0	**12**	137	0	0
77	1516	500	500	**63**	995	0	0	**74**	491	0	0	**71**	133	0	0
5	1514	499	498	**21**	991	240	223	**99**	486	0	0	**56**	131	0	0
93	1510	500	500	**109**	980	357	324	**40**	475	0	0	**55**	129	0	0
14	1504	500	500	**86**	932	0	2	**85**	444	0	0	**80**	128	0	0
4	1498	484	484	**110**	891	0	0	**100**	414	0	0	**73**	104	0	0
60	1492	500	500	**64**	877	0	0	**49**	390	0	0	**81**	97	0	0
106	1474	500	500	**95**	835	0	0	**24**	385	0	0	**72**	84	0	0
43	1438	500	500	**111**	813	0	0	**127**	329	0	0	**83**	80	0	0
61	1429	500	500	**75**	813	0	0	**9**	325	0	0	**82**	57	0	0

[1] Experiments 1, 2, 25 and 26 are omitted, due to extreme percentage of ones/zeros in the samples.

The variable w is always generated as N(3,1) and IC groups AIC and SBIC together, because the results of both criteria were identical.

We find the experiments with equivalent models at the middle of this table. Above them (the upper part of the table), we see the experiments where M1 was better and, below them (the lower part), those where $M2$ was the best model. The results showed that, in the upper end of the table, the values in columns IC and C2 tended towards 500 and, in the lower end, tended towards zero, corroborating the theoretical conclusions.

The experiments that contained equivalent models with identical densities (27 and 3) corroborated the theoretical results. On the other hand, in experiment 6 (equivalent models with non-identical densities), both IC and C_2 performed well. Nevertheless, the theoretical

results for C_2 concluded that this criterion was adequate only in some situations. We can affirm that experiment 6 belonged to one of these situations, characterized by uniform distribution of the DGP variables, and similar (but not oversized) weights for both variables. We think that these characteristics could cause C_2 to behave well.

Finally, we observed a non-adequate behaviour of the criteria in some experiments. In the upper part of the table, experiments 62, 48 and 57 showed values far less than 500 and large differences between the values of the criteria columns. In the lower part of the table, we find that experiments 21 and 109 have IC and C2 column values far from 0. Could this atypical behaviour be due to anomalous observations? Taking into account that del_i is the main element which underlies the indicators used to classify the experiments, we studied whether extreme values of $|del_i|$ were always associated with the same selection (same sign of del_i). We found that this happened in all the experiments, except in 21. Eliminating these extreme values, the behaviour of the criteria became adequate, as we show in Table 6.

Table 6. Atypical experiments in Case 2. Behaviour of the selection criteria.

Experiment	(β_1,β_2)	Type of x	N	*AM*1	absdel	IC	C2
62	(3,1)	B	1800	1144	0,1517	500	500
48	(1,−0.5)	B	1750	1035	0,0097	377	499
57	(3,−2)	B	1800	1003	0,12	499	500
109	(3,3)	C	1925	909	0,4	0	0

As a final comment, we observed a greatly reduced number of experiments with equivalent models. We understood that this was logical, because the experiments of Case 2 corresponded to pairs of models where the DGP was not nested in M1 or M2. Given that model M1 contained the variable x_1 and model M2 contained x_2 (x_1 and x_2 being the only DGP variables), it was very difficult to find cases where both the M1 and M2 models were equivalent.

When we re-executed the analysis for a sample size of 200, the results were similar in general terms, although the tendency toward correct behaviour of the criteria was slower. Nevertheless, we could affirm that the three criteria performed quite well for finite sample sizes.

5. Conclusions

Within the framework of overlapping binary models, we have studied the power of model selection criteria: the well-known information criteria AIC and SBIC, and the C_2 criterion, based on the mean square error of prediction.

As we previously mentioned, two binary models are overlapping if both have the same functional form (both probit or both logit), with some common explanatory variables and some specific variables. In this article, we distinguished two cases: (i) at least one of the competing models is well-specified, and (ii) neither of them is correctly specified. This last case is an important aspect of our work because it is not commonly considered in empirical works.

From a theoretical point of view, we have classified the competing models as equivalent or non-equivalent. Once this classification had been carried out, the task was to define the requirement that a given criterion must satisfy to be considered as adequate. Specifically, if two models are equivalent, the probability limits of a given criterion must be the same in both models. However, if one of them is better, its corresponding limit must be lower than that of the other model. The theoretical analysis carried out has confirmed that all the criteria performed well in every situation. Only C_2 did not, sometimes, behave well in a specific alternative.

These theoretical results have been corroborated by a Monte Carlo experiment. The most complicated situation to simulate was, as we expected, when neither of the two models were well-specified. This situation can lead to three possibilities: equivalent models

with identical densities, equivalent models with non-identical densities, and non-equivalent models.

In order to develop this part of the Monte Carlo exercise, we had to generate 132 different DGPs, leading to 132 different experiments. Each of the experiments corresponded to one of the three theoretical relationships mentioned above. To establish the specific relationship, we have defined three indicators:

(a) The average of the absolute differences between the expected log-likelihoods (at the pseudo-trues) of both models. We have denoted it as *absel*.
(b) The number of observations in the sample where the expected log-likelihood in model M1 is larger than in *M*2. Note that we have assumed that *M*1 is the closest to the DGP. This indicator is called *AM*1.
(c) The average of the absolute differences between the density functions (at the pseudo-trues) of both models. We have denoted it as *absdifden*.

The general conclusion is that the three criteria behaved well for overlapping binary models: when neither of the two competing models was well-specified, the criteria tended to choose the best of them, that is, the closest to the DGP. In the most commonly studied case, where at least one of the competing models was correct, our conclusion was that the criteria also performed well, as we expected. Furthermore, when both models were correct, the criteria tended to choose the most parsimonious model.

It is important to note that these criteria are also used when we compared an extensive set of models, correctly specified or misspecified. The criteria AIC, SBIC and C_2 allow us to order them, being in the first places those correctly specified, which will be equivalent to each other. Among them, the first one (the selected model) will be the most parsimonious if we use the SBIC criterion. The misspecified models will be at the bottom of the ranking.

This paper has been focused on the restricted framework of overlapping binary models. In order to complete this analysis, a future work should study the behaviour of AIC, SBIC and C_2 in the non-nested framework. Moreover, the wider context of multinomial dependent variables could be the aim of future research. Given that the MLE procedure was also applied to estimate these models, the formal expression of the IC would be quite straight, while C_2 would require a deeper analysis.

Author Contributions: Conceptualization, T.A. and I.V.; methodology, T.A. and I.V.; software, I.V.; formal analysis, T.A. and I.V.; numerical simulation, I.V.; writing—original draft preparation, T.A. and I.V.; writing—review and editing, I.V.; supervision, I.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by DGA Reference Group S40_20R and Agencia Estatal de Investigación Reference PID2019-106822RB-I00.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data were generated in the simulation study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aljarallah, R.; Kharroubi, S.A. Use of Bayesian Markov Chain Monte Carlo Methods to Model Kuwait Medical Genetic Center Data: An Application to Down Syndrome and Mental Retardation. *Mathematics* **2021**, *9*, 248. [CrossRef]
2. Li, Z.; Wang, E.; Su, J.; Yu, Y. Using MCMC Probit Model to Value Coastal Beach Quality Improvement. *J. Environ. Prot.* **2011**, *2*, 109–114. [CrossRef]
3. Vuong, Q.H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* **1989**, *57*, 307–333. [CrossRef]
4. Lewis, F.; Butler, A.; Gilbert, L. A Unified Approach to Model Selection Using the Likelihood Ratio Test. *Methods Ecol. Evol.* **2011**, *2*, 155–162. [CrossRef]
5. Hendry, D.F. *Econometric Modelling*; Department of Economics, University of Oslo: Oslo, Norway, 2000.
6. Hong, H.; Preston, B. *Nonnested Model Selection Criteria*; Department of Economics, Stanford University: Stanford, CA, USA, 2006; pp. 1–33.

7. Linhart, H.; Zucchini, W. *Model Selection*; John Wiley and Sons: New York, NY, USA, 1986.
8. Pesaran, M.H.; Pesaran, B. A Simulation Approach to the Problem of Computing Cox's Statistics for Testing Nonnested Models. *J. Econom.* **1993**, *57*, 377–392. [CrossRef]
9. Santos Silva, J.M.C. A Score Test for Non-Nested Hypotheses with Applications to Discrete Data Models. *J. Appl. Econom.* **2001**, *16*, 577–592. [CrossRef]
10. Akaike, H. Information Theory and an Extension of the Likelihood Ratio Principle. In Proceedings of the Second International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, 2–8 September 1971; Petrov, B.N., Csáki, F., Eds.; Akademiai Kiado: Budapest, Hungary, 1973; pp. 267–281.
11. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
12. Aparicio, T.; Villanúa, I. Some Selection Criteria for Nested Binary Choice Models: A Comparative Study. *Comput. Stat.* **2007**, *22*, 635–660. [CrossRef]
13. Kim, H.; Cavanaugh, J.E. Model Selection Criteria Based on Kullback Information Measures for Nonlinear Regression. *J. Stat. Plan. Inference* **2005**, *134*, 332–349. [CrossRef]
14. Van Der Hoeven, N. The Probability to Select the Correct Model Using Likelihood-Ratio Based Criteria in Choosing Between Two Nested Models of Which the More Extended One Is True. *J. Stat. Plan. Inference* **2005**, *135*, 477–486. [CrossRef]
15. Lalou, P.; Chalikias, M.; Skordoulis, M.; Papadopoulos, P.; Fatouros, S. A Probabilistic Evaluation of Sales Expansion. In Proceedings of the 5th International Symposium and 27th National Conference on Operation Research, Egaleo, Greece, 9–11 June 2016; pp. 109–113, ISBN 978-618-80361-6-1.
16. Seo, T.K.; Thorne, J.L. Information Criteria for Comparing Partition Schemes. *Syst. Biol.* **2018**, *67*, 616–632. [CrossRef] [PubMed]
17. Jhwueng, D.C.; Huzurbazar, S.; O'Meara, B.C.; Liu, L. Investigating the Performance of AIC in Selecting Phylogenetic Models. *Stat. Appl. Genet. Mol. Biol.* **2014**, *13*, 459–475. [CrossRef] [PubMed]
18. Susko, E.; Roger, A.J. On the Use of Information Criteria for Model Selection in Phylogenetics. *Mol. Biol. Evol.* **2020**, *37*, 549–562. [CrossRef] [PubMed]
19. Dziak, J.J.; Coffman, D.L.; Lanza, S.T.; Li, R.; Jermiin, L.S. Sensitivity and Specificity of Information Criteria. *Brief. Bioinform.* **2020**, *21*, 553–565. [CrossRef] [PubMed]
20. Monfardini, C. An Illustration of Cox's Non-Nested Testing Procedure for Logit and Probit Models. *Comput. Stat. Data Anal.* **2003**, *42*, 425–444. [CrossRef]
21. Bierens, H.J. *Topics in Advances Econometrics*; Cambridge University Press: Cambridge, UK, 1994.
22. White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **1982**, *50*, 1–25. [CrossRef]
23. Kohn, R. Consistent Estimation of Minimal Subset Dimension. *Econometrica* **1983**, *51*, 367–376. [CrossRef]
24. Aparicio, T.; Villanúa, I. *Selection Criteria for Overlapping Binary Models*; Documentos de Trabajo; Facultad de Economía y Empresa, Universidad de Zaragoza: Zaragoza, Spain, 2012; pp. 1–54.
25. Newey, W.K.; McFadden, D. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*; Elsevier: Amsterdam, The Netherlands, 1994; Volume 4, pp. 2111–2245.
26. Davidson, J. *Econometric Theory*; Blackwell: Oxford, UK, 2000.
27. Gourieroux, C.; Monfort, A.; Renault, E.; Trognon, A. Simulated Residuals. *J. Econom.* **1987**, *34*, 201–252. [CrossRef]