Margarita Genius
University of Crete

Elisabetta Strazzera
University of Cagliari and CRENoS
e-mail: strazzera@unica.it

EVALUATION OF LIKELIHOOD BASED TESTS FOR NON-NESTED DICHOTOMOUS CHOICE CONTINGENT VALUATION MODELS

Abstract

Distributional assumptions are crucial in the estimation of the value of public projects assessed by means of contingent valuations analyses, and it would seem obvious that tests for model specification should play an important part in the statistical analysis. It can be observed, though, that when the competing hypotheses are non nested, the choice of the model is often based on heuristic grounds, or, at most, on deterministic selection model criteria such as Akaike s (1973). In this paper we study two alternative, probabilistic, approaches to checking model specification, that, like Akaike s, are based on the Kullback-Leibler Information Criterion (KLIC): the model selection testing proposed by Vuong (1989) and the non nested model test proposed by Cox, in the simulated approach of Pesaran and Pesaran (1993). The three approaches are confronted by comparing their performance in selecting among different contingent valuation models applied to simulated data. Our preliminary results seem to warrant the use of Vuong s test, complemented in same cases by the application of the Cox test.

November 2000

1. Introduction

Survey data for contingent valuation analyses are often obtained through a dichotomous choice questioning framework: individuals are asked if they would be willing to pay some specified amount to insure access to some public good, and the answer may be Yes or No. In single bound models the elicitation procedure stops at this stage; while in multiple bound models further payment questions follow. Individual responses are then analyzed by means of statistical models to produce an estimate of the value that the public places on the good.

While non parametric or semi-parametric approaches are becoming more popular among contingent valuation practitioners, it is often necessary, for inference or prediction purposes, to uncover a functional relationship between the demand for the public good and individual socioeconomic characteristics. Since the dependent variable is discrete, estimates of the relevant parameters are generally obtained through a maximum likelihood procedure, and the value of the mean, or median, willingness to pay is calculated as a function of the estimated parameters. It is well known that maximum likelihood estimates are consistent if the model specification is correct, but that this does not hold in general for misspecified models: the risk of producing biased estimates of the benefits stemming from the public good is quite serious, and this may diminish the reliability of the analysis for public choice purposes.

Since distributional assumptions are so crucial in the estimation results, it would seem obvious that tests for model specification should play an important part in the statistical analysis of discrete data. It can be observed, though, that while in contingent valuation studies the application of tests for nested models is standard practice, the analysis is often less accurate when the competing hypotheses are non nested.

The analysis of non nested models has followed two distinct approaches in literature: model selection criteria, and hypothesis

testing (cfr. Gourieroux and Monfort (1995)). In the model selection approach, each competing model is evaluated by means of a numerical criterion: for a given sample observation, the procedure consists of selecting the model that optimizes the chosen criterion. A typical example in linear regression is the (adjusted) R2 criterion, while in maximum likelihood estimation a commonly used criterion is the above cited Akaike information criterion, or one of its variants.

The problem of the model selection approach is that it produces a deterministic outcome, defined by the ranking of the values of the criterion, and it does not take into account the probabilistic nature of that result. Vuong (1989) points out that differences in the criterion values may not be statistically significant: yet the deterministic model selection approach would consider a model superior to another one, while in fact they may be considered as statistically equivalent. He then sets the information criterion in a testing framework, where the null hypothesis is that the two competing models are equally close to the true model.

The hypothesis testing approach takes a step further, extending the classical testing procedures to the case of non nested hypotheses: examples are the generalized Wald test, the generalized score test, and the Cox test, which is a generalized likelihood ratio test; or, in a different line, the tests based on artificial nesting: the Davidson-MacKinnon (1981) test, the Atkinson test and the Quandt test belong to this category (cfr. Gourieroux and Monfort, cit.).

Non nested competing models are generally assessed by means of selection criteria, such as Akaike s (1973), while we are not aware of any testing approach in contingent valuation studies; and it might be added that such applications are very few in discrete data modeling in general. The extra computational difficulties that the testing approach entails, may explain why this path has been so neglected. But there may be also a more theoretically founded justification for the choice of model selection criteria over hypothesis testing to test economic

theories: as pointed out by Granger et al. (1995), the choice of the null hypothesis and the significance level is arbitrary, and this is even more so when testing is applied to non nested hypothesis. In their view, when the choice of the particular model is data dependent it is better to use well-thought-out model selection procedures rather then formal hypothesis testing.

The aim of this paper is to shed some light on the matter, analyzing the performance of either approach in selecting among different contingent valuation models applied to simulated data. In particular, we compare three methods that are based on the Kullback-Leibler Information Criterion (KLIC):

- the Akaike information criterion;
- the Vuong test;
- the Cox test, in the simulated approach of Pesaran and Pesaran (1993).

The structure of the paper is the following: section 1 gives a brief background about the KLIC and explains the 3 procedures above, section 2 describes the experimental setting of our simulation, section 3 reports the results of our experiments and finally section 4 contains our conclusions.

2. Methods

In order to describe the different statistics or criteria we introduce some notation and terminology.

Consider a sequence (Y_i, X_i) i=1,2 .. of i.i.d. random vectors. The modeler is interested in the conditional probability distribution of the vector Y_i given X_i . Define the true conditional density as:

$$\ell_0(y \mid x) = \prod_{i=1}^{n} j_0(y_i \mid x_i),$$

which is unknown. To evaluate its proximity to a specified parametric model, that we define as:

$$\ell(y \mid x;\theta) = \prod_{i=1}^{n} f(y_i \mid x_i;\theta), \theta \in \Theta,$$

we make use of the notion of Kullback-Leibler Information Criterion (KLIC):

$$K_n(\ell(y \mid x;\theta)/\ell_0(y \mid x)) = \frac{1}{n} E_0\left(\log \frac{\ell_0(Y \mid x)}{\ell(Y \mid x;\theta)} \right).$$

We will be interested in comparing pairs of competing parametric families of conditional densities of Y_i given X_i given by

$$H_f : \left\{ f(y_i \mid x_i; \beta), \beta \in \mathrm{B} \subset \Re^F \right\},$$
$$H_g : \left\{ g(y_i \mid x_i; \gamma), \gamma \in \Gamma \subset \Re^G \right\},$$

where the models H_f and H_g are strictly non-nested.

It can be shown (cfr. Gourieroux and Monfort, cit.) that the asymptotic Kullback-Leibler proximity between the true probability distribution and a given parametric model is approximated by

$$\tilde{K} = \frac{1}{n} \sum_{i=1}^{n} \log f_0(y_i \mid x_i) - \frac{1}{n} \sum_{i=1}^{n} \log f(y_i \mid x_i; \hat{\theta}_n).$$

Since f_0 is unknown $\tilde{K}$ cannot be used; it can be noticed, though, that when two models are compared, the first term of $\tilde{K}$ remains constant, so that minimization of the criterion only depends on the second term, i.e. on the maximum likelihood of the two competing models.

Denoting by $\hat{\beta}_n$ and $\hat{\gamma}_n$ the (quasi) maximum likelihood estimators of β and γ under H_f and H_g respectively, this amounts to calculating:

$$LR_n(\hat{b}_n,\hat{g}_n)=\sum_{i=1}^{n}\log f(y_i|x_i,\hat{b}_n)-\sum_{i=1}^{n}\log g(y_i|x_i,\hat{g}_n),$$

i.e. the likelihood ratio of the two models.

The drawback of using $\tilde{K}$ as such, is that it increases for more general models. In order to overcome this problem, Akaike (1973) proposed a correction of this criterion, that penalizes more complex models. The Akaike Information Criterion (AIC) penalizes the log-likelihood of each model by a quantity equal to the number of its parameters:

$$AIC=\left(\sum_{i=1}^{n}\log f(y_i\mid x_i;\hat{b}_n)-p\right),$$

where p is the number of parameters. The Akaike criterion for model selection simply consists in comparing the AIC values for the two models:

$$AICMS=\left(\sum_{i=1}^{n}\log f(y_i\mid x_i;\hat{b}_n)-p\right)-\left(\sum_{i=1}^{n}\log g(y_i|x_i,\hat{g}_n)-q\right)$$

If the value is positive the first model is chosen, otherwise the second will be deemed best. Obviously, if the two models are characterized by the same number of parameters p and q, the Akaike criterion for model selection reduces to LR_n.

A criticism to the use of model selection criteria such as Akaike s is that they are deterministic: the model that satisfies the given criterion is selected. However, some authors point out that this result is just the outcome of a random draw from the sample space, and as such should be treated in probabilistic terms.

This issue is addressed by Vuong (1989), whose approach sets the model selection criterion in a hypothesis testing framework. More specifically, it tests whether the models under consideration are equally close to the true model, where closeness is measured by the KLIC.

The null hypothesis is given by:

$$H_0 : E_0\left[\log\frac{f(y_i/x_i;\beta^*)}{g(y_i/x_i;\gamma^*)}\right]=0,$$ (both models are equivalent)

against

$$E_0\left[\log\frac{f(y_i/x_i;\beta^*)}{g(y_i/x_i;\gamma^*)}\right]\succ 0,$$ (H_f is better than H_g), or

$$E_0\left[\log\frac{f(y_i/x_i;\beta^*)}{g(y_i/x_i;\gamma^*)}\right]\prec 0,$$ (H_g is better than H_f),

where β^* and γ^* are the pseudo-true values of β and γ respectively. The tests statistics proposed by Vuong are the following:

-an unadjusted LR statistic given by

$$n^{-1/2}LR_n(\hat{\beta}_n,\hat{\gamma}_n)/\hat{w}_n ,$$

where

$$\hat{w}_n=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{f(y_i/x_i;\hat{\beta})}{g(y_i/x_i;\hat{\gamma})}\right]^2-\left[\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(y_i/x_i;\hat{\beta})}{g(y_i/x_i;\hat{\gamma})}\right]^2};$$

-an adjusted LR statistic given by

$$n^{-1/2}\,L\tilde{R}_n(\hat{\beta}_n,\hat{\gamma}_n)/\hat{w}_n$$ where

$L\tilde{R}_n(\hat{\beta}_n,\hat{\gamma}_n)\equiv LR_n(\hat{\beta}_n,\hat{\gamma}_n)-x_n$, and ξn is a correction factor that penalizes each model for model complexity. Different correction factors, as well as a slightly different version of the denominator term, give rise to different variants of the Vuong s statistics, that in any case, for non nested models, is asymptotically standard normal under H0.

While Vuong s approach is to test if the two models are statistically different, the Cox approach aims at testing if the true conditional probability distribution belongs to one of the competing models under examination. The null hypothesis may be that the true Data Generating Process (DGP) belongs to Hf ; but it also may be that the DGP belongs to Hg. Due to the special role of the null hypothesis in this context, it is not obvious which null hypothesis we should choose. Many (see Pesaran and Pesaran (1993), henceforth P&P; Weeks (2000)) advocate performing the non-nested test twice by reversing the role of the null and alternative hypothesis. This procedure could very well lead to a situation where both models are accepted or both are rejected.

Following P&P, the standardized Cox statistic is asymptotically normal and under the null Hf is given by

$$S_f(\hat{b}_n, \hat{g}_n) = \sqrt{n} T_f \big/ \hat{v}_f \ ,$$

where $\hat{v}_f$ is an estimate of the asymptotic variance, and

$$T_f = \frac{1}{n} LR_n(\hat{b}_n, \hat{g}_n) - \hat{E}_f\left(\frac{1}{n} LR_n(\hat{b}_n, \hat{g}_n)\right)$$

The expression $\hat{E}_f(\cdot)$ stands for the conditional expectations operator under Hf; it should be noted that $E_f\left(\frac{1}{n} LR_n(\hat{b}_n, \hat{g}_n)\right)$ is zero when we have nested models but does not vanish in the case of non-nested models. Due to the difficulties in computing this term (see Pesaran and Weeks (2000)), this test has not been widely applied outside of the linear regression model. The difficulty lies in computing

an estimate of $E_f\left(\frac{1}{n}LR_n(\hat{b}_n,\hat{g}_n)\right)$, because it entails finding an estimate of the pseudo true value g^*, i.e the value that maximizes $E_f(\log L_g(g))$.

In the case of discrete choice models, P&P have derived a simulation method to compute the above statistic which we can apply to the case of the single bound model. P&P simulate R independent samples of n indicators (dependent variable) assuming that F is the true distribution; then for each one of the R simulated samples they compute the maximum likelihood estimate of γ using the c.d.f G. Denoting by $\hat{g}^*_n(R)$ the average of the R estimates of γ, this is a consistent estimate of the pseudo true value. Finally we can estimate the expected value above as follows:

$$\hat{E}_f\left(\frac{1}{n}LR_n(\hat{b}_n,\hat{g}_n)\right)=\frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{l}(1-F_i(\hat{b}_n))\log\left(\frac{1-F_i(\hat{b}_n)}{1-G_i(\hat{g}^*_n(R))}\right)+\\ F_i(\hat{b}_n)\log\left(\frac{F_i(\hat{b}_n)}{G_i(\hat{g}^*_n(R))}\right)\end{array}\right\}.$$

Since there is no a priori reason why F should be the null hypothesis, P&P suggest to reverse the null and alternative hypothesis, i.e. testing G against F: therefore it will be necessary to find an estimate of the expression $E_g\left(\frac{1}{n}LR_n(\hat{b}_n,\hat{g}_n)\right)$ and this in turn will require finding an estimate of the value that maximizes the expected value of the log-likelihood using model F when G is the null model.

3. Experimental Design

The dichotomous choice elicitation method for contingent valuation produces a dichotomous type of response to payment questions that are differentiated among individuals. This particular setting allows different modeling options: the latent dependent variable can be modeled either as a dichotomous variable, as in the random utility model (RUM) framework used in the utility differential model by Hanemann (1984); or as a censored variable, which is the approach proposed by Cameron and James (1987) and Cameron (1988). The latter produces separate estimates for the coefficients and the scale parameter of the model, and allows for a more straightforward calculation for the mean or median value of the public good, and was therefore chosen for this application.

Depending on the assumptions on the individuals preferences, the wtp can be modeled as a linear or non linear function of the individual socioeconomic covariates. The econometric modeling involves further assumptions on the distribution of the error term, and its functional relationship with the deterministic part of the wtp model: the combination of the two components can possibly give rise to many modeling specifications, but in practice probit, logit, log-normal, log-logistic, weibull are the most commonly used. This choice may be due to the fact that they can easily be estimated with econometric modules available in popular statistical packages like Limdep, Stata, or Sas (cfr. Hanemann and Kanninen, 1999).

In our experiment, the true model is the following:

$$wtp_i = 27 + 1.5x_{i1} - 3x_{i2} + 0.5x_{i3} + e_i,$$

where x_1 and x_3 are continuous variables respectively ranging from 4 to 75 and from 0.5 to 1.5; while x_2 is a qualitative variable, taking values zero or one. The error term is distributed as a Normal with zero mean and standard deviation 15.

We hypothesize that the researcher assumes a model linear in the covariates, with an additive error term, obtaining the following econometric model: for each individual *i*,

$$Y_i = x_i'\delta + \varepsilon_i,$$

where x is the vector of regressors. In this model the latent variable Y_i is unobserved: the observed variable is the answer YES or NO to the question regarding whether or not the individual would be willing to pay a given amount t_i.

For a given sample of n independent observation, the generic log-likelihood function is:

$$LogL = \sum_{i=1}^{n} \{ I_i \log [1 - \Psi_i ((t_i - x_i'\delta)/\nu)] + (1 - I_i) \log [\Psi_i ((t_i - x_i'\delta)/\nu)] \}$$

$$= \sum_{i=1}^{n} \{ I_i \log [1 - \Psi_i (q)] + (1 - I_i) \log [\Psi_i (q)] \}$$

where Ψ represents generically one of the distributions hypothesized by the researcher; $q = (\delta, \nu)$, and Ii is a dummy variable assuming value one if the individual response to the bid question is positive, zero otherwise. Since bids are varied among individuals, δ and v can be estimated separately.

A further assumption is that the researcher (righteously) thinks that the deterministic part of the model is correctly specified, but is unsure about the distribution of the random term, and tries different specifications: Normal, Logistic, Extreme Value, that combined with the linear function for the deterministic part of the model give rise to the Probit, Logit, and Weibit[1] models. As mentioned earlier, the first two models are frequently applied by contingent valuation practitioners: the underlying distributions for these two models are

[1] Notice that this is a different specification from the non linear Weibull model frequently used in contingent valuation studies.

both symmetric, with fatter tails for the logistic. The weibit model is much less common; we choose it because it is an example of asymmetric distribution associated to a linear functional form for the deterministic part of the model, and this facilitates comparisons between models for our purposes.

Since the model checking methods under analysis are based on the KLIC, the models have to be compared in pairs. The Akaike criterion only requires maximization of the log-likelihood for each model, and then the calculation is straightforward; in Vuong s approach the calculation is slightly more involved, but still it only requires the computation of the maximum log-likelihood of each model. For the simulated Cox test the procedure is definitely more arduous, since it involves the computation of the quasi maximum likelihood estimates of each model assuming that the other model is the true DGP.

4. Results

We now examine the results of our simulation for the normal DGP and the three candidate models: probit, logit and weibull. As explained in the preceding section, the KLIC requires that the three models should be compared in pairs, so the tables report results for the pairs probit vs logit, probit vs weibit, and logit vs weibit.

It can be observed that parameter estimates do not differ very much across models. The parameter v is a scale parameter for the logit and the weibit, that should be multiplied respectively by $p/\sqrt{3}$ and $p/\sqrt{6}$ to obtain the estimated standard deviations for the two models. Moreover, since the mean of the extreme value distribution is not zero we have to add 0.5772v to the estimated constant by the weibit to compare it to the corresponding probit estimate. The major differences in the regressors coefficients estimates concern the parameter δ_4: the weibit model gives the poorest estimates, but it

should be noted that even the probit estimates of the above parameter are not close to the true value for samples as big as 1000.

Similar parameter estimates produce similar estimates for the mean wtp value, as it can be observed in tables 1.2, 2.2 and 3.2; but it can be observed that the asymmetry of the extreme value distribution produces a substantial downward bias in the estimated median. Since use of the median rather than the mean for asymmetric distributions is the the standard choice in contingent valuation studies, we will focus on the estimated value for this central tendency measure for the weibit model. We first rank the models in terms of closeness of their estimated mean and median wtp values and the true population value, measured by MSE; and then we see the ranking provided by each of the three methods under investigation, the Akaike information criteria, Cox and Vuong tests in the framework of the three experiments.

In experiment 1 the two models under consideration are the probit (H_g) and the logit (H_h). Table 1.3 shows some descriptive statistics for the Akaike criteria, the Vuong and the Cox tests. Since we are computing the Akaike criteria as the difference in the log-likelihoods it is natural for the variance to be increasing with sample size.

Table 1.4 shows the probability of rejecting the null for the three tests when the nominal size is five percent. In the case of the Vuong, the null hypothesis is that the two models are *equivalent* and therefore the probability of rejecting the null should approach 1 if the null is false, while it should approach 0.05 if the null is true. Since in this case the probit is correctly specified we would expect this probability to approach one, while it falls quite short of one even for n=1000. We are not aware of any simulation study where the behavior of the Vuong test has been investigated, so we cannot compare our results with other benchmarks.

The first Cox test the probit against the logit- should have probability rejections approaching the nominal size as the sample becomes larger. The rejection probability attains the 12% level for

n=1000. The second Cox test-logit versus probit- should have probability rejections that approach 1 as the sample size increases, but for n=1000 it only reaches the 53% level. Our results are comparable to Weeks (2000) in the sense that the probability of rejecting the null for the first Cox test decreases with sample size although it reaches lower levels in Weeks study while for the second Cox test the probability of rejecting the null reaches the level of 60% for a sample of 2000 in his study. It should be noted that Weeks is comparing alternative variants for the Cox statistic and does not use the simulated version of P&P for all sample sizes.

In table 1.5 we analyze the number of times one model would be chosen over the other according to the different criteria. The Akaike criteria chooses the right model over 70% of the time for all sample sizes, and since it is well known that probit and logit estimates do not differ that much, it is not surprising that it chooses the logit specification 20% of the time. Vuong s test finds that the logit and probit models are *equivalent* over 70% of the time, while the rest of the time it chooses the right model. Further research is called for to see what sample size is needed for the Vuong test to discriminate better between the two models. We might ask ourselves how close the probit and logit models are, and of course the answer depends on the criterion used to compare them. The KLIC is one of many alternative measures of closeness (cfr. Aznar Grasa (1989)). However, because the probit model is correctly specified, Vuong s statistic should be able to discriminate between the two for a large enough sample. As far as the Cox tests are concerned, we can see that for the sample sizes considered, around 40% of the time the probit is chosen while 40% of the time both models are accepted.

In experiment two, where we compare the probit to the weibit, the results are a little bit different. The rejection probability of the null reaches the 52% level for the Vuong statistic, while the first Cox test

reaches the size level of 11% and the second Cox test reaches a rejection probability of 98%, all for n=1000 (table 2.3).

Overall the three methods perform better in experiment 2, for n=1000 the Akaike criteria chooses the probit 98% of the time, Vuong s test chooses it 52% of the time, while Cox 88.5% (table 2.4). The density function used in the weibit likelihood function is not symmetric and its tail behavior is quite different from the normal. Still the Vuong test seems to lack power in this case as well.

In experiment 3, we compare the logit and the weibit, both of them being misspecified. The Akaike criteria will choose the logit over the weibit up to 92% of the time for n=1000, while the Vuong test seems to point out that both models are equivalent but as the sample size increases, the probability that the logit model is chosen increases as well. Once again bigger samples are needed to determine the behavior of the Vuong test. As for the Cox test, out of the four possible outcomes it chooses the logit 48.6 percent of the time, while it rejects both models 47% of the time for n=1000 (table 3.5).

5. Conclusions

From an operative point of view, it is important that the selection method is able to signal a possible misspecification when its consequencs are more serious. In our context this could be answered by comparing the differences between the parameter estimates and between the wtp estimates for the two models.

In experiment 1, the choice of logit instead of the correct specification probit does not carry any serious consequence, both in terms of the parameter estimates and the central measure estimate for wtp.

Experiments 2 and 3 show that if the true DGP is symmetric and an asymmetric distribution is fitted instead, the estimated median is seriously biased downwards. It can be observed from tables 2.5 and 3.5 that the Akaike method, the only method usually employed in

contingent valuation studies to select between non nested models, selects quite often (about 20% of the times for the small size, but in experiment 3 also the medium size sample) the wrong model. Application of the Vuong test is relatively easy, and it seems to provide a good guidance not to choose the wrong model (since the percentages of selection of the wrong model are negligible in all experiments for all sample sizes). In other words, if the Vuong test selects one model, the researcher can quite confidently rely on it. On the other hand, the Vuong test can be a problem because it very often accepts the null hypothesis of equivalence between the two models. In such a case, the researcher should probably try further specifications; if no model can be found that passes the Vuong test, then a Cox test should be applied to investigate if equivalence means that both models can be accepted, or both models are to be rejected. The Cox test has higher percentages than Vuong s of acceptance of the wrong model; this, and the computational complexity of the procedure, do not warrant a routinely application of the Cox test.

In conclusion, based on the results of this set of experiments, we do recommend use of the Vuong test, complemented by a Cox test in case of uncertainty; further experiments are under way to check the consistency of this results with other models and DGP.

References:

Akaike, H. (1973): Information theory and an extension of the maximum likelihood principle, in *Proceedings of the 2nd International Symposium on Information Theory*, ed. by N. Petrov, and F. Csadki, pp. 267-281. Akademiai Kiado, Budapest.

Aznar Grasa, A. (1989): *Econometric Model Selection: A New Approach*. Kluwer Academic Publishers, Spain.

Cameron, T. A. (1988): "A new paradigm for valuing non-market goods using referendum data: maximum likelihood estimation by censored logistic regression," *Journal of Environmental Economics and Management,* **15**, 355-79.

Cameron, T. A., and M.D. James (1987): "Efficient estimation methods for closed-ended contingent valuation surveys, *The Review of Economics and Statistics*, **69**, 269-76.

Davidson, R., and J.G. MacKinnon (1981): Several tests for model specification in the presence of alternative hypotheses, *Econometrica,* 49, 781-793.

Gourieroux, C., and A. Monfort (1995): *Statistics and Econometric Models,* Cambridge University Press.

Granger, C., M. L. King, and H. White (1995): Comments on Testing Economic Theories and the Use of Model Selection Criteria, *Journal of Econometrics,* 67, 173-187.

Hanemann W. M., and B. J. Kanninen (1999): "Statistical analysis of discrete response cv data", in Bateman, I. and K. Willis, *Valuing Environmental Preferences,* Oxford University Press.

Pesaran, M. H., and B. Pesaran (1993): "A simulation approach to the problem of computing Cox;s statistic for testing nonnested models," *Journal of Econometrics,* **57**, 377-392.

Pesaran, M. H., and M. Weeks (2000): Non-nested hypothesis testing: an overview , in *Companion to Theoretical Econometrics,* ed. Badi H. Baltagi, Basil-Blackwell, Oxford. (forthcoming).

Vuong, Q.H. (1989): Likelihood Ratio Tests for Model Selection and Non-Nested Hypothesis, *Econometrica,* **57(2)**, 307-333.

Weeks, M. (2000): Testing the Binomial and Multinomial Choice Models Using Cox s Non-Nested Test, in Mariano, R., Schuermann, T. and Weeks, M.J., *Simulation based inference in econometrics: methods and applications, Cambridge University Press, Cambridge.*

Appendix 1

Experiment 1. Probit vs logit

Table 1.1. *Parameter estimates[a] for normal DGP using H_f (normal) and H_g (logistic) across 300 replications[b].*

	Sample Size					
	300		600		1000	
Parameters	H_f	H_g	H_f	H_g	H_f	H_g
δ_1	26.488	26.745	26.450	26.688	26.609	26.853
	(9.056)	(9.154)	(5.747)	(5.661)	(4.647)	(4.674)
δ_2	1.508	1.502	1.501	1.497	1.503	1.497
	(0.118)	(0.118)	(0.083)	(0.084)	(0.067)	(0.068)
δ_3	-2.817	-2.803	-2.847	-2.872	-3.009	-2.965
	(4.264)	(4.216)	(2.907)	(2.935)	(2.120)	(2.126)
δ_4	0.531	0.497	0.796	0.783	0.762	0.722
	(6.260)	(6.411)	(4.439)	(4.390)	(3.407)	(3.450)
v^c	14.752	8.290	14.829	8.346	14.863	8.325
	(2.018)	(1.215)	(1.387)	(0.894)	(1.021)	(0.643)

Mean values and standard deviations (in parenthesis) over 300 replications.

The actual number of successful experiments was 284, 288 and 293 for the 300, 600 and 1000 sample size respectively.

The estimated scale parameter of the logit should be multiplied by $\pi/3^{1/2}$ for comparison with the corresponding probit estimate.

Table 1.2. *Mean wtp and MSE*

	H_f		H_g	
	Mean-Median (st.dev)	MSE SMSE	Mean (st.dev)	MSE SMSE
300	86.314	6.773	86.301	6.788
	(1.869)	*3.482*	(1.886)	*3.545*
600	84.351	1.359	84.374	1.374
	(1.158)	*1.343*	(1.167)	*1.361*
1000	85.455	1.892	85.458	1.895
	(0.991)	*0.979*	(0.990)	*0.977*

MSE with respect to the population true mean wtp: 84.5
MSE with respect to the sample true mean wtp: 86.292, 84.429, and 85.457 for the 300, 600 and 1000 sample size respectively.

Table 1.3. *Mean and standard deviation of Akaike criterion, Vuong and Cox statistics*

Methods	Sample Size		
	300	600	1000[a]
Akaike	0.290(0.668)	0.626(1.459)	0.812(1.457)
Vuong	0.883(1.308)	1.083(1.330)	1.096(1.357)
Cox: $H_0=H_f$	0.587(2.374)	0.774(3.282)	0.433(2.600)
Cox: $H_0=H_g$	-2.632(2.621)	-2.698(2.649)	-2.562(2.184)

One abnormal value in the Cox statistics has been discarded.

Table 1.4. *Vuong and Cox tests: Percentage of rejections of the null (0.05-level test)*

Statistics	Sample size		
	300	600	1000
Vuong	0.21	0.26	0.23
Cox: $H_0=H_f$	0.20	0.14	0.12
Cox: $H_0=H_g$	0.54	0.56	0.53

Table 1.5. *Conclusions drawn from each method (percentage)*

Method		Conclusion			
Akaike		H_f is better		H_g is better	
	300	0.711		0.288	
	600	0.781		0.218	
	1000	0.789		0.211	
Vuong		H_f is better	H_g is better	H_f and H_g equivalent	
	300	0.204	0.003	0.792	
	600	0.26	0	0.74	
	1000	0.228	0.003	0.769	
Cox		H_f accepted H_g rejected	H_f accepted H_g rejected	Both H_f and H_g accepted	Both H_f and H_g rejected
	300	0.373	0.024	0.429	0.172
	600	0.423	0.013	0.423	0.138
	1000	0.413	0.003	0.467	0.116

Experiment 2. Probit vs Weibit

Table 2.1. ***Parameter estimates[a] for normal DGP using H_f (normal) and H_g (extreme value) across 300 replications[b].***

	Sample Size					
	300		600		1000	
Parameters	H_f	H_g	H_f	H_g	H_f	H_g
δ_1	26.688	19.308	26.875	19.360	27.231	19.649
	(8.438)	(9.631)	(5.872)	(6.644)	(4.497)	(5.019)
δ_2	1.499	1.507	1.495	1.498	1.497	1.503
	(0.106)	(0.118)	(0.084)	(0.089)	(0.063)	(0.065)
δ_3	-3.357	-3.512	-2.903	-3.016	-2.888	-3.045
	(4.291)	(4.756)	(2.877)	(3.212)	(2.329)	(2.541)
δ_4	1.033	1.014	0.747	0.774	0.236	0.076
	(6.268)	(6.736)	(4.629)	(4.920)	(3.269)	(3.633)
v[c]	14.573	12.880	14.676	13.093	14.761	13.281
	(1.940)	(2.053)	(1.426)	(1.584)	(1.151)	(1.149)

a) Mean values and standard deviations (in parenthesis) over 300 replications.

b) The actual number of successful experiments was 292 for the 300, and 296 for the 600 and 1000 sample size.

c) The estimated scale parameter of the weibit should be multiplied by $\pi/6^{1/2}$ for comparison with the corresponding probit estimate and we should add the factor 0.5772v to the constant of the weibit.

Table 2.2. Mean and Median[a] estimated wtp and their MSE

	H_f		H_g			
	Mean-Median (st.dev)	MSE SMSE	Mean (st.dev)	MSE SMSE	Median (st.dev)	MSE SMSE
300	86.227	6.242	86.469	8.549	83.755	5.110
	(1.808)	*3.262*	(2.165)	*4.703*	(2.138)	*10.988*
600	84.425	1.616	84.542	2.000	81.783	9.324
	(1.271)	*1.611*	(1.416)	*2.011*	*2.138*	*8.942*
1000	85.404	1.668	85.477	1.891	82.679	4.244
	(0.924)	*0.853*	(0.969)	*0.937*	(0.964)	*8.645*

a) For the probit model the two values coincide.
b) MSE with respect to the population true mean wtp: 84.5
c) MSE with respect to the sample true mean wtp: 86.292, 84.429, and 85.457 for the 300, 600 and 1000 sample size respectively.

Table 2.3. Mean and standard deviation of Akaike criterion, Vuong and Cox statistics

Methods	Sample Size		
	300	600	1000
Akaike	1.685(2.700)	3.777(3.568)	6.993(3.787)
Vuong	1.022(1.339)	1.533(1.244)	2.065(1.026)
Cox: $H_0=H_f$	-0.705(2.052)	-0.586(1.601)	-0.349(1.212)
Cox: $H_0=H_g$	-3.159(2.144)	-4.082(1.641)	-5.106(1.424)

Table 2.4. Percentage of rejections of the null (0.05-level test).

Statistics	Sample size		
	300	600	1000
Vuong	0.25	0.35	0.52
Cox: H_0=H_f	0.29	0.20	0.11
Cox: H_0=H_g	0.71	0.92	0.98

Table 2.5. Conclusions drawn from each method (percentage)

Method		Conclusion			
Akaike		H_f is better		H_g is better	
	300	0.764		0.236	
	600	0.905		0.094	
	1000	0.983		0.017	
Vuong		H_f is better	H_g is better	H_f and H_g equivalent	
	300	0.250	0.000	0.750	
	600	0.344	0.006	0.648	
	1000	0.523	0.000	0.476	
Cox		H_f accepted H_g rejected	H_f accepted H_g rejected	Both H_f and H_g accepted	Both H_f and H_g rejected
	300	0.589	0.164	0.123	0.123
	600	0.790	0.074	0.006	0.128
	1000	0.885	0.013	0.003	0.097

Experiment 3. Logit vs Weibit

Table 3.1. Parameter estimates[a] for normal DGP using H_f (logistic) and H_g (extreme value) across 300 replications[b].

	Sample Size					
	300		600		1000	
Parameters	H_f	H_g	H_f	H_g	H_f	H_g
δ_1	26.903	19.276	26.800	19.160	27.383	19.649
	(8.476)	(9.615)	(6.417)	(7.037)	(4.583)	(5.265)
δ_2	1.512	1.527	1.494	1.497	1.490	1.506
	(0.122)	(0.136)	(0.085)	(0.094)	(0.061)	(0.068)
δ_3	-3.351	-3.638	-2.924	-2.755	-2.982	-3.251
	(4.137)	(4.255)	(2.953)	(3.009)	(2.279)	(2.538)
δ_4	0.347	0.185	0.999	0.908	0.538	0.238
	(6.075)	(6.364)	(4.531)	(4.816)	(3.443)	(3.892)
v[c]	8.264	13.085	8.273	13.137	8.276	13.293
	(1.020)	(2.052)	(0.942)	(1.508)	(0.660)	(1.124)

a) Mean values and standard deviations (in parenthesis) over 300 replications.

b) The actual number of successful experiments was 288, 290 and 294 for the 300, 600 and 1000 sample size respectively.

c) The estimated scale parameter of the logit and the weibit should be multiplied by $\pi/3^{1/2}$ and by $\pi/6^{1/2}$ respectively, for comparison with the corresponding probit estimate. We should add as well the factor 0.5772v to the constant of the weibit.

Table 3.2. Mean and Median[a] estimated wtp and their MSE

	H_f		H_g			
	Mean-Median (st.dev)	MSE SMSE	Mean (st.dev)	MSE SMSE	Median (st.dev)	MSE SMSE
300	86.284 (1.772)	6.315 *3.130*	86.459 (2.044)	8.001 *4.190*	83.702 (2.014)	4.678 *10.748*
600	84.548 (1.393)	1.936 *1.948*	84.548 (1.476)	2.199 *2.228*	81.900 (1.441)	8.828 *8.463*
1000	85.526 (0.940)	1.932 *0.885*	85.526 (0.978)	2.154 *0.973*	82.795 (0.988)	3.881 *8.060*

a) For the logit model the two values coincide.
b) MSE with respect to the population true mean wtp: 84.5
c) MSE with respect to the sample true mean wtp: 86.292, 84.429, and 85.457 for the 300, 600 and 1000 sample size respectively.

Table 3.3. Mean and standard deviation of Akaike criterion, Vuong and Cox statistics

Methods	Sample Size		
	300	600	1000
Akaike	1.462(2.606)	2.754(10.395)	5.919(4.703)
Vuong	0.663(1.223)	1.062(1.254)	1.481(1.023)
Cox: $H_0=H_f$	-2.520(2.422)	-2.368(2.686)	-2.235(1.897)
Cox: $H_0=H_g$	-2.122(1.722)	-2.846(1.954)	-3.629(1.179)

Table 3.4. Vuong and Cox tests: Percentage of rejections of the null (0.05-level test)

Statistics	Sample size		
	300	600	1000
Vuong	0.15	0.24	0.34
Cox: $H_0=H_f$	0.57	0.50	0.51
Cox: $H_0=H_g$	0.57	0.80	0.96

Table 3.5. Conclusions drawn from each method (percentage)

Method		Conclusion			
		H_f is better		H_g is better	
Akaike	300	0.753		0.247	
	600	0.813		0.186	
	1000	0.921		0.079	
		H_f is better	H_g is better	H_f and H_g equivalent	
Vuong	300	0.118	0.034	0.847	
	600	0.217	0.020	0.762	
	1000	0.333	0.003	0.663	
		H_f accepted H_g rejected	H_f accepted H_g rejected	Both H_f and H_g accepted	Both H_f and H_g rejected
Cox	300	0.392	0.395	0.038	0.173
	600	0.493	0.189	0.006	0.310
	1000	0.486	0.040	0.000	0.472

Contributi di Ricerca CRENoS

I Paper sono disponibili in: **http://www.crenos.unica.it**

00/12 *Elisabetta Strazzera, M. Genius*, Evaluation of Likelihood Based Tests for non-nested Dichotomous Choice Contingent Valuation Models

00/11 *Elisabetta Strazzera, R. Scarpa, G. Hutchinson, S. Chilton*, Analisys of Mixed Structure Data for Valuation of Forest Resources for Recreation

00/10 *Luca Deidda*, On the Real Effects of Financial Development

00/9 *Cristiano Antonelli, Roberto Marchionatti, Stefano Usai*, Productivity and External Knowledge: the Italian Case

00/8 *Maria Musumeci*, Innovazione tecnologica e beni culturali. uno studio sulla situazione della Sicilia

00/7 *Maria Musumeci*, Informazione e processi di apprendimento nello sviluppo locale

00/6 *Elisabetta Strazzera, Riccardo Scarpa, Pinuccia Calia, Guy Garrod, Ken Willis*, "Modelling Zero Bids in Contingent Valuation Surveys

00/5 *L. Robin Keller, Elisabetta Strazzera*, Examining Predictive Accuracy among Discounting Models

00/4 *Antonio Sassu, Sergio Lodde,* Saperi locali, innovazione tecnologica e sviluppo economico: indagine su un campione di imprese sarde

00/3 *Sergio Lodde,* Capitale umano e sviluppo economico. Cosa sappiamo in teoria e nei fatti?

00/2 *Raffaele Paci, Stefano Usai,* Externalities, Knowledge, Spillovers and the Spatial Distribution of Innovation

00/1 *Raffaele Paci*, Convergenza e divergenza tra le regioni europee. Implicazioni per lo sviluppo economico in Sardegna

99/17 *Paolo Piacentini, Giovanni Sulis*, Crescita virtuosa e crescita neodualistica nell ambito regionale: tendenze recenti per le aree europee in ritardo di sviluppo

99/16 *Sergio Lodde*, Nuova teoria della crescita e sviluppo locale. Alcune possibili connessioni

99/15 *Raffaele Paci, Stefano Usai,* The Role of Specialisation and Diversity Externalities in the Agglomeration of Innovative Activities

99/14 *Gianna Boero*, *Emanuela Marrocu*, Modelli non lineari per i tassi di cambio: un confronto previsivo

99/13 *Luca Deidda*, Interaction between Economic and Financial Development

99/12 *Gianna Boero*, *Costanza Torricelli*, The Information in the Term Structure: Further Results for Germany

99/11 *Sergio Lodde,* Education Growt:: Some Disaggregate Evidence from the Italian Regions

99/10 *Robin Naylor,* "Endogenous Determination of Trade Regime and Bargaining outcome"

99/9 *Raffaele Paci, Francesco Pigliaru,* "Technological Catch-Up and Regional Convergence in Europe"

99/8 *Raffaele Paci, Nicola Pusceddu,* "Lo stock di capitale fisso nelle regioni italiane. 1970 - 1994"

99/7 *Raffaele Paci,* "L'evoluzione del sistema economico della Sardegna negli anni novaanta"

99/6 *Alessandro Lanza, Francesco Pigliaru,* "Why Are Tourism Countries Small and Fast-Growing?"

99/5 *Pinuccia Calia, Elisabetta Strazzera,* "A Sample Selection Model for Protest Non-Response Votes in Contingent Valuation Analises"

99/4 *Adrianan Di Liberto, James Simons,* "Some economics Issues in Convergence Regression"

99/3 *Rosanna Carcangiu, Giovanni Sistu, Stefano Usai*, "Struttura socio-economica dei comuni della Sardegna. Suggerimenti da un'analisi cluster"

99/2 *Francesco Pigliaru*, "Detecting Technological Catch-Up in Economic Convergence"

99/1 *Marzio Galeotti, Alessandro Lanza*, "Desperately Seeking (Environmental) Kuznets"

98/7 *Elisabetta Strazzera,* "Option values and Flexibility Preference"

98/6 *Roberto Marchionatti, Stefano Usai,* "International Technological Spillovers and Economic Growth. The Italian Case"

98/5 *Sergio Lodde,* "Invidia e imprenditorialità. Alcune note sul ruolo delle emozioni nello sviluppo economico"

98/4 *Adriana Di Liberto, James Symons,* "Human Capital Stocks and the Development of Italian Regions: a Panel Approach"

98/3 *Raffaele Paci, Francesco Pigliaru*, "Growth and Sectoral Dynamics in the Italian Regions"

98/2 *Rossella Diana, Elisabetta Serra, Elisabetta Strazzera,* "Politiche non sostenibili per lo sviluppo sostenibile. Il caso del Parco del Gennargentu"

98/1 *Pinuccia Calia, Elisabetta Strazzera*, Bias and Efficiency of Single Vs. Double Bound Models for Contingent Valuation Studies: A Monte Carlo Analysis"

97/8 *RaffaelePaci, Stefano Usai,* Technological Enclaves and Industrial Districts. An Analysis of the Regional Distribution of Innovative Activity in Europe

97/7 *Marta Sanna*, "Spillover tecnologici nord-sud: una nota a Coe - Helpman - Hoffmaister"

97/6 *Sergio Lodde*, "Human Capital and Growth in the European Regions. Does Allocation Matter?"

97/5 *Raffaele Paci, Francesco Pigliaru*, Is Dualism still a Source of Convergence across European Regions?

97/4 *Gianna Boero, Costanza Torricelli*, The Expectations Hypothesis of the Term Structure: Evidence for Germany

97/3 *Raffaele Paci, Francesco Pigliaru*, European Regional Growth: Do Sectors Matter?

97/2 *Michael Pontrelli,* Un analisi econometrica del contenuto informativo della struttura a termine dei tassi di interesse tedeschi

97/1 *Raffaele Paci, Andrea Saba*, The empirics of Regional Economic Growth in Italy. 1951-1993

96/12 *Francesco Pigliaru*, Economia del turismo: note su crescita, qualità ambientale e sostenibilità

96/11 *Riccardo Contu*, Rapporti scientifico-contrattuali e adattamenti istituzionali nella dinamica impresa-accademia: persistenza delle New Biotechnology Firms nell industria biotecnologica USA degli anni 90"

96/10 *Elisabetta Schirru*, Modelli di determinazione del tasso di cambio: un analisi di cointegrazione

96/9 *Raffaele Paci*, More Similar and Less Equal. Economic Growth in the European Regions

96/8 *Daniela Sonedda*, Commercio internazionale e crescita economica nei casi della Corea del Sud e delle isole Filippine: un analisi di causalità

96/7 *Raffaele Paci, Francesco Pigliaru*, β-Convergence and/or Structural Change? Evidence from the Italian Regions

96/6 *Paolo Piacentini, Paolo Pini*, Domanda, produttività e dinamica occupazionale: un analisi per moltiplicatori

96/5 *Raffaele Paci, Riccardo Rovelli*, Do Trade and Technology reduce Asymmetries? Evidence from Manufacturing Industries in the EU

96/4 *Riccardo Marselli, Marco Vannini,* La criminalità nelle regioni italiane: il ruolo del sistema sanzionatorio, delle motivazioni economiche e del contesto sociale

96/3 *Anna Maria Pinna*, Sectoral Composition of Trade and Economic Growth: some New Robust Evidence

96/2 *Emanuela Marrocu*, A Cointegration Analysis of W.A. Lewis Trade Engine Theory

96/1 *Rinaldo Brau, Elisabetta Strazzera*, Studio di valutazione monetaria per il parco nazionale del Gennargentu. Indagine preliminare

95/5 *Raffaele Paci, Stefano Usai*, Innovative Effort, Technological Regimes and Market Structure

95/4 *Stefano Usai, Marco Vannini*, Financial Development and Economic Growth: Evidence from a panel of Italian Regions

95/3 *Sergio Lodde*, Allocation of Talent and Growth in the Italian Regions

95/2 *Rinaldo Brau*, Analisi econometrica della domanda turistica in Europa: implicazioni per lo sviluppo economico delle aree turistiche

95/1 *Antonio Sassu*, *Raffaele Paci, Stefano Usai*, Patenting and the Italian Technological System