Fondazione Eni Enrico Mattei

Model Selection and Tests for Non Nested Contingent Valuation Models: An Assessment of Methods

Margarita Genius* and Elisabetta Strazzera**

NOTA DI LAVORO 34.2001

JUNE 2001

ETA – Economic Theory and Applications

*Department of Economics, University of Crete
**DRES and CRENoS, University of Cagliari

Fondazione Eni Enrico Mattei
Corso Magenta, 63, 20123 Milano, tel. +39/02/52036934 – fax +39/02/52036946
E-mail: letter@feem.it
C.F. 97080600154

MODEL SELECTION AND TESTS FOR NON NESTED CONTINGENT VALUATION MODELS: AN ASSESSMENT OF METHODS

Margarita Genius
Department of Economics, University of Crete
Elisabetta Strazzera
DRES and CRENoS, University of Cagliari

Address for correspondence:
Elisabetta Strazzera,
DRES and CRENoS,
University of Cagliari,
Via Fra Ignazio 78,
I-09123, Cagliari, Italy
tel. +39 070 675 3763; fax +39 070 675 3760

e-mail: STRAZZERA@UNICA.IT

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. In contingent valuation studies it can be observed that the application of tests for nested models is quite common, for example by means of likelihood ratio tests; the analysis, though, is much 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) R^2 criterion, while in maximum likelihood estimation a commonly used criterion is the information criterion proposed by Akaike (1973), 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 the latter category (cfr. Gourieroux and Monfort, cit.).

In contingent valuation analyses, 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 this field; and it might be added that such applications are very few in discrete data modeling in general. As put forth by Pesaran and Weeks (2000), the extra computational difficulties that the testing approach entails may explain why this path has been so neglected. However, 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 investigate on 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 2 gives a brief background about the KLIC and explains the 3 procedures above, section 3 describes the experimental setting of the simulation exercise, section 4 reports the results of the experiments and finally section 5 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:

$$\ell_0(y \mid x) = \prod_{i=1}^{n} \varphi_0(y_i \mid x_i),$$

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

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

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

$$\widetilde{K} = \frac{1}{n}\sum_{i=1}^{n} \log \varphi_0(y_i \mid x_i) - \frac{1}{n}\sum_{i=1}^{n} \log \varphi(y_i \mid x_i;\hat{\theta}_n),$$

where $\hat{\theta}_n$ is the maximum likelihood estimator of θ.

Since φ_0 is unknown $\widetilde{K}$ cannot be used; it can be noticed, though, that when two models are compared, the first term of $\widetilde{K}$ 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:

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

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

The drawback of using LR_n 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:

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

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

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

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:

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

against

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

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

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

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

where

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

-an adjusted LR statistic given by

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

$L\tilde{R}_n(\hat{\beta}_n, \hat{\gamma}_n) \equiv LR_n(\hat{\beta}_n, \hat{\gamma}_n) - \xi_n$, and ξ_n 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 H_0.

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 H_f ; but it also may be that the DGP belongs to H_g. 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 H_f is given by

$$S_f(\hat{\beta}_n, \hat{\gamma}_n) = \sqrt{n} T_f / \hat{v}_f ,$$

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

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

The expression $\hat{E}_f(\cdot)$ stands for the conditional expectations operator under H_f; it should be noted that $E_f\left(\frac{1}{n} LR_n(\hat{\beta}_n, \hat{\gamma}_n)\right)$ 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 the linear regression model. The difficulty lies in computing an estimate of $E_f\left(\frac{1}{n} LR_n(\hat{\beta}_n, \hat{\gamma}_n)\right)$,

because it entails finding an estimate of the pseudo true value γ^*, i.e the value that maximizes $E_f(\log g(y|x,\gamma))$.

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 CV 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 $\hat{\gamma}_n^*(R)$ 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:

$$\hat{E}_f\left(\frac{1}{n}LR_n(\hat{\beta}_n,\hat{\gamma}_n)\right)=\frac{1}{n}\sum_{i=1}^{n}\left\{(1-F_i(\hat{\beta}_n))\log\left(\frac{1-F_i(\hat{\beta}_n)}{1-G_i(\hat{\gamma}_n^*(R))}\right)+F_i(\hat{\beta}_n)\log\left(\frac{F_i(\hat{\beta}_n)}{G_i(\hat{\gamma}_n^*(R))}\right)\right\}.$$

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 $E_g\left(\frac{1}{n}LR_n(\hat{\beta}_n,\hat{\gamma}_n)\right)$ 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. Full details on the derivation of the Cox simulation based test statistic are given in Appendix 2.

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 their standard errors, 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 latent variable, individual willingness to pay, or *wtp* for brevity, 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).

For our experiment, we considered different DGP for the wtp, obtained from a linear functional form for the deterministic part of the model, which is common to all experiments, and an additive error term that is varied across experiments. The general model is the following:

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

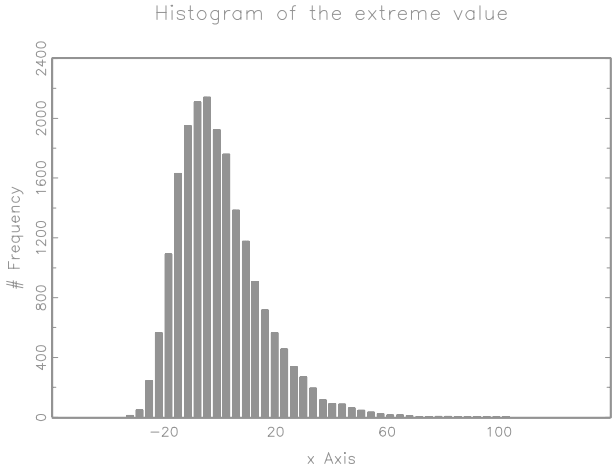
where x_1 and x_3 are continuous variables respectively ranging from 4 to 75 and from 0.5 to 1.5; while x_2 is a qualitative variable, taking values zero or one.

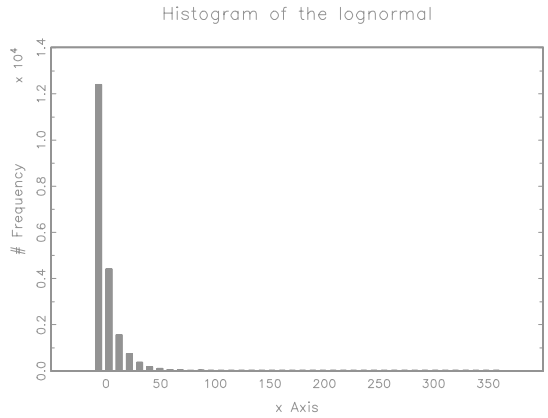
For the error term we investigate three different scenarios, allowing for differences in the degree of skewness of the distributions. In the first one the error term is distributed as a Normal with zero mean and standard deviation 15. In the second one we assume that the error term follows an extreme value distribution[1] with mean zero and standard deviation 15; while in the third scenario the error term is generated as a translated lognormal[2], obtained from a lognormal by subtracting its mean, so that the resulting error has mean

[1] The parameters of the extreme value distribution are: a=-$b\gamma$ and $b = 15\frac{\sqrt{6}}{\pi}$ where γ is Euler's constant which we approximate with the value 0.5772156649.

[2] The parameters of the lognormal are: $\mu = \ln\left(\frac{15}{\sqrt{e\cdot(e-1)}}\right)$, σ=1.

zero and standard deviation 15. The shape of the last two distributions is depicted in the histograms below which have been produced by generating samples of size 20000 from the two distributions described above.





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:

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

,

where Ψ represents generically one of the distributions hypothesized by the researcher; $\theta = (\delta, v)$, and I_i 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[3] models. As mentioned earlier, the first two models are frequently applied by contingent valuation practitioners: the underlying distributions for these two models are 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 complex, since it involves the computation of the quasi maximum likelihood estimates of each model assuming that the other model is the true DGP.

The purpose of this study is to assess the three approaches in situations that are typically encountered by applied researchers, rather than to investigate the large sample

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

properties of the tests. The sample sizes considered in the experiments, 300, 600, and 1000 observations, are representative of a small, medium and large contingent valuation data set.

4. Results

We now examine the results of our simulation for each of the three DGP (Normal, Extreme Value, translated Lognormal) and the three candidate models: probit, logit and weibit. 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.

In table 1 of each experiment we report the estimates of the parameters obtained from the two models to be compared. The number of replications in the Monte Carlo experiments was fixed at 300, but the actual number in each experiment depends on the rate of convergence –always very high, even though it never attains 100%; the differences in the number of successful replications explain the slight differences between estimates obtained under the same specification and the same data but in a different experiment.

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 $\pi/\sqrt{3}$ and $\pi/\sqrt{6}$ 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 $0.5772 \cdot v$ to the estimated constant by the weibit to compare it to the corresponding probit estimate. A general feature of all experiments is that the estimates of the parameters δ_3 and especially δ_4 are less efficient than the others; we are not able at this stage to provide an explanation of this discrepancy in the precision of the parameter estimates.

Similar parameter estimates produce similar estimates for the mean wtp value, as it can be observed in table 2 of each experiment; but it can be observed that the asymmetry of the extreme value distribution generally produces a lower value for the estimated median – whether it is more correct or biased depends on the true DGP. Choice of the median rather than the mean as a central tendency measure is often deemed to be preferable, both on statistics and economics grounds (cfr. Hanemann and Kanninen, cit., pp. 325-6): therefore, we will focus on the estimated value for this central tendency measure for the weibit

model. Closeness of the estimated mean and median wtp values to the true population value can be measured by the mean square error (MSE); this criterion can give us a measure of the goodness of fit of each specification. Since the sampling happened to be not well behaved, with the 600 sample being closer to the population values than the 1000 sample, in the same table we report also the MSE with respect to the sample WTP values.

Finally, table 3 of each experiment shows the conclusions drawn from each of the three methods under investigation: the Akaike information criterion, the Vuong test and the Cox test.

From table 3 of each experiment, we can derive the probability of rejecting the null for the Vuong and Cox tests when the nominal size is five percent. For the Vuong test, 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. The null hypothesis for each one of the two Cox tests is that one of the models is correctly specified, therefore the probability of rejecting each one of the two nulls should be computed considering the two cases when the corresponding null is rejected, either because the alternative is accepted, or because both are rejected.

To help the exposition, we use the following code for our experiments:

Pairs of models **DGP**	**Probit** vs **Logit**	**Probit** vs **Weibit**	**Logit** vs **Weibit**
Normal	**A-I**	**A-II**	**A-III**
Extreme value	**B-I**	**B-II**	**B-III**
Trans. Lognormal	**C-I**	**C-II**	**--**

We analyze the results by first considering the experiments where one of the models was correctly specified: i.e. experiments A-I and A-II, and experiments B-II and B-III.

As recalled above, in these cases the probability for the Vuong test to reject the null should approach one. In experiment A (probit correctly specified) it falls quite short of one even for n=1000: the probability is 23% in I (probit vs logit), and 52% in II (probit vs weibit). In experiment B (weibit correctly specified) the probability is 42% in II and 57% in III (logit vs weibit). In terms of power of the test, these results are not very encouraging for the Vuong test; 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.

Let's see now, for the same set of experiments, the performance of the Cox test. We can observe mixed results, but in general it performs better, in terms of power of the test, than the Vuong test, especially when the probit and the weibit are compared. In addition, the power improves, while the size of the tests decreases, with the sample size; the latter reaching levels of 10-13% for n=1000. In this sense our results are comparable to Weeks (2000)[4]. It remains to analyze the performance of the Akaike criterion: when one of the models is correctly specified, and the sample size is large enough (in our case, 600 or more) the Akaike criterion works very well when the weibit is compared with the other two; while when two similar models, probit and logit, are compared, the Akaike criterion is not able to pick up the correct model in more than 20% of the cases. For the small sample size this problem extends also to the comparisons involving the weibit model: with the exception of experiment B-III, the Akaike criterion may lead to choosing the wrong model in about the 20% of the cases.

Let's now consider the subset of experiments where both models are misspecified. These are experiments A-III, B-I, and both experiments C.

In experiment A-III, we compare the logit and the weibit, the DGP being Normal. The Akaike criterion will choose the logit over the weibit, with increasing frequency as the sample size increases. The Vuong test seems to point out that both models are equivalent in terms of KLIC; but as the sample size increases, the probability that the logit model is chosen increases as well. It can be noticed that the probability that the Vuong test selects

[4] It should be noted that Weeks compares alternative variants for the Cox statistic and does not use the simulated version of P&P for all sample sizes.

the “worse” model (in this case, the weibit) is close to zero. As for the Cox test, its behavior is not very satisfying for the small sample; while it improves with sample size, yet in this case the power of the first Cox test falls short of 53% even for n=1000.

Experiment B-I considers probit versus logit when the true DGP is Extreme Value. The Akaike criterion would choose logit or probit with almost the same probability when the sample size is small, but the frequency of the logit choice increases with sample size. Here the probability of rejecting the null for the Vuong test reaches 0.096 and both models are found to be equivalent 90% of the time, while the Cox test accepts both models over 65% of the time for n=1000. Here the Cox test lacks power since the probability of rejecting the null is below 23% for both tests for n=1000.

This situation is reversed when we consider the last experiments, under the translated Lognormal DGP. Here the Akaike criterion discriminates much more between models, leading to choosing the logit and the weibit over the probit. Also the Vuong test now shows a strong support for the same models and it chooses the logit around 85%, and the weibit 100% of the time for n=1000. While the selected models, especially the weibit, are indeed closer to the true DGP, none of them is correct: in this case the Vuong test, like the Akaike criterion, is not able to signal that both models are misspecified. Only the Cox test points out that in this situation none of the models is to be accepted, again improving with sample size, with the power of both Cox tests approaching one for the large samples.

Summing up, when both models are misspecified, the Akaike criterion is obviously dominated by the other two methods, since it does not give any information about the misspecification problem. For the other two methods, we get mixed results, depending on the DGP and the sample size. In general, the Cox test seems to perform better than the Vuong, especially for large samples; for the smaller sample size instead the Cox test is not so satisfying, since it often accepts the null when it is false.

5. Conclusions

From an operative point of view, it is important that the selection method be able to signal a possible misspecification when its consequences are more serious. In our context this could be answered by considering the MSE of the estimated mean and median. The

most serious consequences of misspecification can be observed in experiment C, where the data generated by a strongly skewed distribution (the translated Lognormal) is fitted with symmetric (probit and logit), or sligthly asymmetric (weibit) models. In this experiment we could see that the Cox test is the only one that, for an adequately large sample size, uncovers the misspecification problem.

Unfortunately, in other cases we could observe that the power of the Cox test is unacceptably low: this holds for example in experiment B-I, where misspecification leads to a relevant bias in the estimated median wtp. Also, the performance of the Cox test when the sample size is small is generally poor.

In conclusion, the results of our experiments can be interpreted as follows:

- the Akaike model selection criterion is unsatisfactory, since when both models are misspecified it cannot signal it; even when one model is correctly specified, for small size samples it presents a non negligible probability of choosing a wrong model. When the sample size is small, it may be preferable to use the Vuong test rather than the Akaike criterion, since it makes less mistakes in the choice of the model, and may suggest further checks of the specification in the (too frequent) case that the two models are deemed equivalent;
- when the sample size is large enough, it is preferable to use the Cox test, that, although more complex to implement, seems more effective and conclusive than the two model selection methods.

In this paper we have analyzed three methods for model selection based on the KLIC: as pointed out by Aznar Grasa (1989), this is just one of many alternative measures of closeness. Further research is required to find out if other methods are superior to those assessed in this study.

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

A. NORMAL DGP

Experiment A-I: Normal DGP, Probit vs Logit

Table A-I.1 *Parameter estimates[a] for normal DGP using H_f (normal) and H_g (logistic) across 300 replications[b].*

Parameters	Sample Size					
	300		600		1000	
	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)

a) Mean values and standard deviations (in parenthesis) over 300 replications.
b) The actual number of successful experiments was 284, 288 and 293 for the 300, 600 and 1000 sample size respectively.
c) The estimated scale parameter of the logit should be multiplied by $\pi/3^{1/2}$ for comparison with the corresponding probit estimate.

Table A-I.2 *Mean estimated wtp and their MSE*

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

a) MSE with respect to the population true mean wtp: 84.500
b) 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 A-I.3 *Conclusions drawn from each method (% frequency)*

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	0.260	0.000	0.740	
	1000	0.228	0.003	0.769	
Cox		H_f accepted H_g rejected	H_g accepted H_f 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 A-II: Normal DGP, Probit vs Weibit

Table A-II.1 *Parameter estimates[a] for normal DGP using H_f (normal) and H_g (extreme value) across 300 replications[b].*

Parameters	Sample Size					
	300		600		1000	
	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.

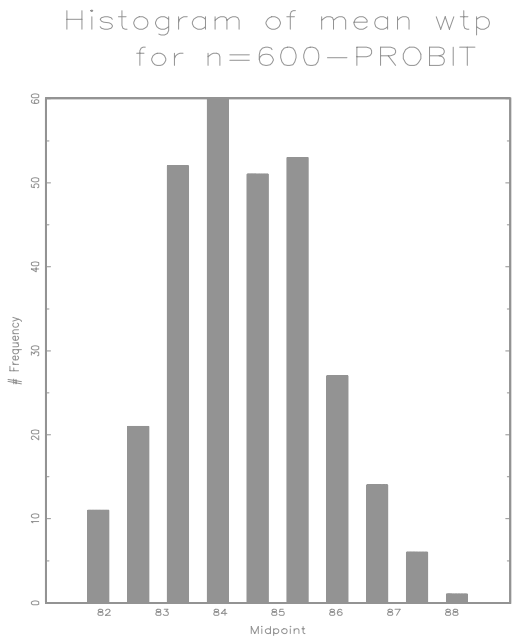
Table A-II.2 *Mean and Median[a] estimated wtp and their MSE*

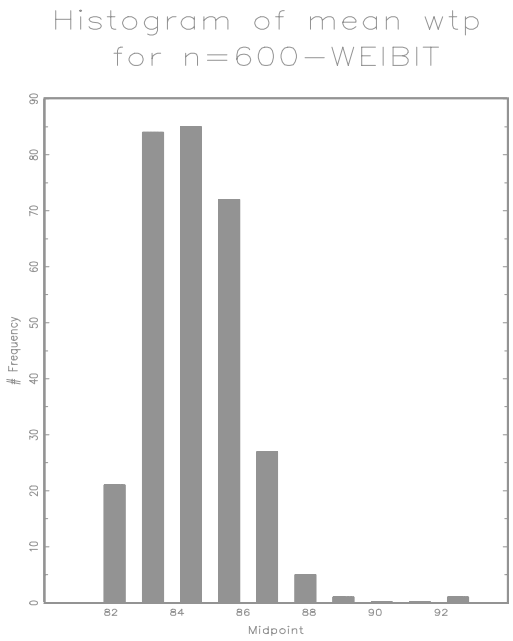
	H_f		H_g			
	Mean-Median (st.dev)	MSE[b] *SMSE[c]*	Mean (st.dev)	MSE *SMSE*	Median (st.dev)	MSE *SMSE*
300	86.227	6.242	86.469	8.549	83.755	5.110
	(1.808)	*3.262*	(2.165)	*4.703*	(2.138)	*10.988*
600	84.425	1.616	84.542	2.000	81.783	9.324
	(1.271)	*1.611*	(1.416)	*2.011*	*2.138*	*8.942*
1000	85.404	1.668	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 A-II.3 *Conclusions drawn from each method (% frequency)*

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_g accepted H_f 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 A-III: Normal DGP, Logit vs Weibit

Table A.III.1: Parameter estimates[a] for normal DGP using H_f (logistic) and H_g (extreme value) across 300 replications[b].

Parameters	Sample Size					
	300		600		1000	
	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 A-III.2: *Mean and Median*[a] *estimated wtp and their MSE*

	H_f		H_g			
	Mean-Median (st.dev)	MSE[b] *SMSE*[c]	Mean (st.dev)	MSE *SMSE*	Median (st.dev)	MSE *SMSE*
300	86.284	6.315	86.459	8.001	83.702	4.678
	(1.772)	*3.130*	(2.044)	*4.190*	(2.014)	*10.748*
600	84.548	1.936	84.548	2.199	81.900	8.828
	(1.393)	*1.948*	(1.476)	*2.228*	(1.441)	*8.463*
1000	85.526	1.932	85.526	2.154	82.795	3.881
	(0.940)	*0.885*	(0.978)	*0.973*	(0.988)	*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 A-III.3 *Conclusions drawn from each method (% frequency)*

Method		Conclusion			
Akaike		H_f is better		H_g is better	
	300	0.753		0.247	
	600	0.813		0.186	
	1000	0.921		0.079	
Vuong		H_f is better	H_g is better		H_f and H_g equivalent
	300	0.118	0.034		0.847
	600	0.217	0.020		0.762
	1000	0.333	0.003		0.663
Cox		H_f accepted H_g rejected	H_g accepted H_f rejected	Both H_f and H_g accepted	Both H_f and H_g rejected
	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

B) EXTREME VALUE DGP

Experiment B-I: Extreme Value DGP, Probit vs Logit

Table B-I.1: *Parameter estimates[a] for extreme value DGP using H_f (normal) and H_g (logistic) across 300 replications[b].*

Parameters	Sample Size					
	300		600		1000	
	H_f	H_g	H_f	H_g	H_f	H_g
δ_1	26.270	26.195	26.596	26.432	26.183	26.152
	(8.009)	(7.870)	(5.539)	(5.463)	(4.576)	(4.489)
δ_2	1.522	1.515	1.514	1.507	1.515	1.506
	(0.117)	(0.114)	(0.078)	(0.077)	(0.065)	(0.064)
δ_3	-2.783	-2.725	-3.140	-3.145	-3.226	-3.152
	(4.037)	(3.987)	(2.837)	(2.769)	(2.197)	(2.163)
δ_4	0.414	0.325	0.593	0.634	1.0	0.918
	(6.088)	(5.965)	(4.318)	(4.184)	(3.422)	(3.388)
v[c]	14.624	8.013	14.858	8.127	14.839	8.103
	(2.316)	(1.226)	(1.475)	(0.798)	(1.202)	(0.682)

a) Mean values and standard deviations (in parenthesis) over 300 replications.
b) The actual number of successful experiments was 274 for the 300, and 291 for the 600 and 1000 sample size.
c) The estimated scale parameter of the logit should be multiplied by $\pi/3^{1/2}$ for comparison with the corresponding probit estimate.

Table B-I.2: *Mean, median estimated wtp and their MSE*

	H_f			H_g		
	Mean-Median (st.dev)	MSE[a] *SMSE[b]* Mean	MSE[c] *SMSE[d]* Median	Mean-Median (st.dev)	MSE *SMSE* Mean	MSE *SMSE* Median
300	86.531	7.268	23.354	86.150	5.704	19.908
	(1.775)	*3.197*	*10.452*	(1.729)	*3.001*	*8.374*
600	84.594	1.568	8.107	84.198	1.641	6.228
	(1.250)	*1.586*	*8.476*	(1.247)	*1.603*	*6.541*
1000	85.583	2.234	13.645	85.166	1.471	10.829
	(1.032)	*1.077*	*7.772*	(1.015)	*1.112*	*5.753*

a) MSE with respect to the population true mean wtp: 84.5
b) 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.
c) MSE with respect to the population true median wtp: 82.03.
d) MSE with respect to the sample true median wtp: 83.827, 81.964, and 82.992 for the 300, 600 and 1000 sample size respectively.

Table B-I.3: *Conclusions drawn from each method (% frequency)*

Method		Conclusion			
Akaike		H_f is better		H_g is better	
	300	0.485		0.514	
	600	0.344		0.656	
	1000	0.251		0.749	
Vuong		H_f is better	H_g is better		H_f and H_g equivalent
	300	0.124	0.036		0.839
	600	0.048	0.054		0.896
	1000	0.031	0.065		0.904
Cox		H_f accepted H_g rejected	H_g accepted H_f rejected	Both H_f and H_g accepted	Both H_f and H_g rejected
	300	0.248	0.106	0.500	0.146
	600	0.158	0.140	0.642	0.058
	1000	0.120	0.168	0.656	0.054

Experiment B-II: Extreme value DGP, Probit vs Weibit

Table B-II.1: *Parameter estimates[a] for the extreme value DGP using H_f (normal) and H_g (extreme value) across 300 replications[b].*

Parameters	Sample Size					
	300		600		1000	
	H_f	H_g	H_f	H_g	H_f	H_g
δ_1	26.905	20.531	26.662	20.475	26.904	20.323
	(8.729)	(8.384)	(5.726)	(5.777)	(4.490)	(4.314)
δ_2	1.503	1.504	1.513	1.506	1.499	1.497
	(0.128)	(0.119)	(0.080)	(0.076)	(0.064)	(0.060)
δ_3	-2.935	-3.206	-3.005	-3.085	-2.954	-3.039
	(4.056)	(3.811)	(2.782)	(2.584)	(2.165)	(1.908)
δ_4	0.486	0.488	0.507	0.490	0.709	0.713
	(6.071)	(5.550)	(4.572)	(4.270)	(3.480)	(3.143)
v[c]	14.326	11.229	14.707	11.518	14.847	11.677
	(2.351)	(1.737)	(1.422)	(1.211)	(1.268)	(1.030)

a) Mean values and standard deviations (in parenthesis) over 300 replications.
b) The actual number of successful experiments was 294 for the 300 and 600 sample sizes and 290 for
c) the 1000 sample size.
d) 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.

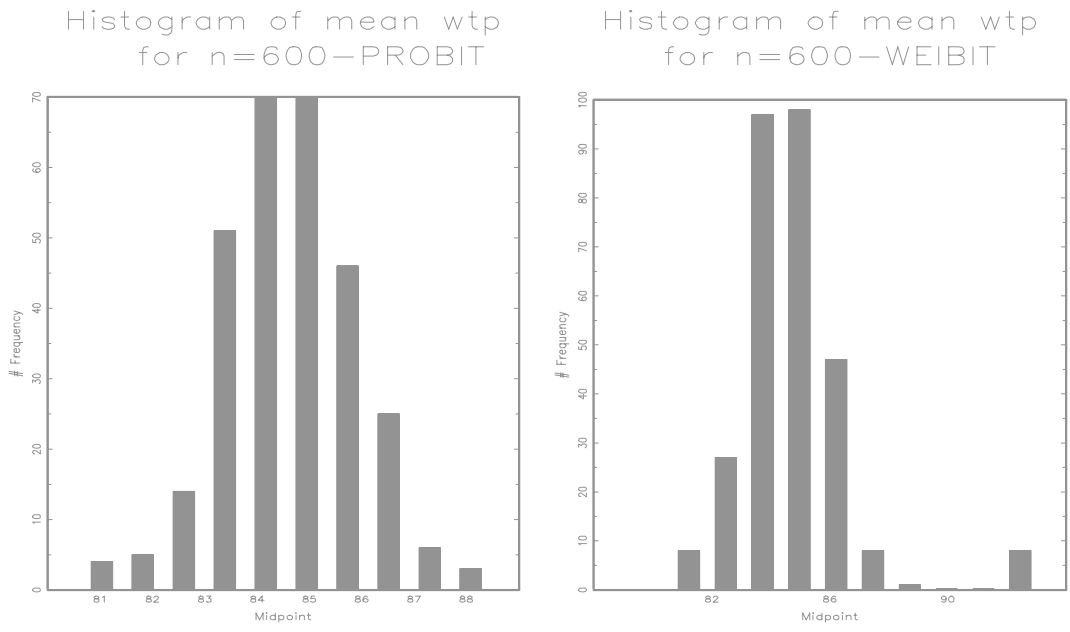
Table B-II.2: *Mean and Median[a] estimated wtp and their MSE*

	H_f			H_g			
	Mean-Median (st.dev)	MSE[b] *SMSE*[c] Mean	MSE[d] *SMSE*[e] Median	Mean (st.dev)	MSE *SMSE* Mean	Median (st.dev)	MSE *SMSE* Median
300	86.361	5.932	21.18	86.312	6.181	83.947	6.38
	(1.573)	*2.472*	*8.888*	(1.704)	*2.895*	(1.654)	*2.742*
600	84.623	1.605	8.288	84.722	3.575	82.295	3.208
	(1.262)	*1.627*	*8.661*	(1.881)	*3.612*	(1.775)	*3.250*
1000	85.583	2.091	13.502	85.619	3.426	83.159	3.190
	(0.960)	*0.934*	*7.629*	(1.477)	*2.200*	(1.391)	*1.956*

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.
d) MSE with respect to the population true median wtp: 82.03
e) MSE with respect to the sample true mean wtp: 83.827, 81.964, and 82.992 for the 300, 600 and 1000 sample size respectively.

Table B-II.3: *Conclusions drawn from each method (% frequency)*

Method		Conclusion			
Akaike		H_f is better		H_g is better	
	300	0.190		0.810	
	600	0.092		0.908	
	1000	0.059		0.941	
Vuong		H_f is better	H_g is better		H_f and H_g equivalent
	300	0.020	0.204		0.775
	600	0.017	0.353		0.629
	1000	0.017	0.400		0.583
Cox		H_f accepted H_g rejected	H_g accepted H_f rejected	Both H_f and H_g accepted	Both H_f and H_g rejected
	300	0.119	0.717	0.068	0.095
	600	0.030	0.826	0.020	0.122
	1000	0.010	0.862	0.003	0.124



Experiment B-III: Extreme value DGP, Logit vs Weibit

Table B-III.1: *Parameter estimates[a] for the extreme value DGP using H_f (logistic) and H_g (extreme value) across 300 replications[b].*

Parameters	Sample Size					
	300		600		1000	
	H_f	H_g	H_f	H_g	H_f	H_g
δ_1	26.364	20.052	26.291	20.280	26.798	20.269
	(8.567)	(8.479)	(5.553)	(5.648)	(4.523)	(4.389)
δ_2	1.515	1.519	1.506	1.506	1.496	1.502
	(0.112)	(0.110)	(0.081)	(0.078)	(0.062)	(0.059)
δ_3	-3.055	-3.248	-3.010	-3.087	-2.995	-3.049
	(4.152)	(3.974)	(2.783)	(2.670)	(2.168)	(1.966)
δ_4	0.351	0.386	0.712	0.596	0.574	0.556
	(6.283)	(5.884)	(4.439)	(4.168)	(3.443)	(3.193)
v[c]	8.033	11.485	8.065	11.501	8.145	11.656
	(1.205)	(1.656)	(0.787)	(1.183)	(0.737)	(0.995)

a) Mean values and standard deviations (in parenthesis) over 300 replications.
b) The actual number of successful experiments was 282, 294 and 286 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 B-III.2: *Mean and Median[a] estimated wtp and their MSE*

	H_f			H_g			
	Mean-Median (st.dev)	MSE[a] *SMSE*[b] Mean	MSE[c] *SMSE*[d] Median	Mean (st.dev)	MSE *SMSE* Mean	Median (st.dev)	MSE *SMSE* Median
300	86.077	5.851	19.7	86.468	8.010	84.048	7.841
	(1.836)	*3.407*	*8.425*	(2.037)	*4.166*	(1.950)	*3.838*
600	84.195	1.634	6.206	84.651	3.043	82.227	2.740
	(1.243)	*1.595*	*6.518*	(1.740)	*3.069*	(1.646)	*2.772*
1000	85.189	1.373	10.844	85.587	3.155	83.132	2.911
	(0.949)	*0.969*	*5.724*	(1.406)	*1.989*	(1.309)	*1.729*

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.
d) MSE with respect to the population true median wtp: 82.03
e) MSE with respect to the sample true median wtp: 83.827, 81.964, and 82.992 for the 300, 600 and 1000 sample size respectively.

Table B-III.3: *Conclusions drawn from each method (% frequency)*

Method		Conclusion			
Akaike		H_f is better		H_g is better	
	300	0.160		0.840	
	600	0.061		0.938	
	1000	0.031		0.969	
Vuong		H_f is better	H_g is better	H_f and H_g equivalent	
	300	0.003	0.330	0.667	
	600	0.017	0.476	0.507	
	1000	0.013	0.555	0.430	
Cox		H_f accepted H_g rejected	H_g accepted H_f rejected	Both H_f and H_g accepted	Both H_f and H_g rejected
	300	0.025	0.830	0.014	0.131
	600	0.006	0.877	0.010	0.105
	1000	0.003	0.888	0.000	0.108

C. LOGNORMAL DGP

Experiment C-I: Lognormal DGP, Probit vs Logit

Table C-I.1: *Parameter estimates[a] for the lognormal DGP using H_f (normal) and H_g (logistic) across 300 replications[b].*

Parameters	Sample Size					
	300		600		1000	
	H_f	H_g	H_f	H_g	H_f	H_g
δ_1	26.572	26.306	25.246	25.326	25.952	25.930
	(7.758)	(6.804)	(5.379)	(4.768)	(4.258)	(3.763)
δ_2	1.511	1.500	1.535	1.515	1.529	1.510
	(0.103)	(0.092)	(0.07)	(0.062)	(0.057)	(0.052)
δ_3	-2.813	-2.808	-2.952	-2.965	-3.095	-3.019
	(3.766)	(3.328)	(2.672)	(2.348)	(2.057)	(1.769)
δ_4	0.529	0.510	1.018	0.946	0.695	0.623
	(5.985)	(5.196)	(4.420)	(3.857)	(3.459)	(2.963)
v[c]	12.202	6.091	13.147	6.468	13.216	6.455
	(3.073)	(1.390)	(2.143)	(0.962)	(1.835)	(0.814)

a) Mean values and standard deviations (in parenthesis) over 300 replications.
b) The actual number of successful experiments was 258, 261 and 247 for the 300, 600 and 1000 sample size respectively.
c) The estimated scale parameter of the logit should be multiplied by $\pi/3^{1/2}$ for comparison with the corresponding probit estimate.

Table C-I.2: *Mean and Median estimated wtp and their MSE*

	H_f			H_g		
	Mean-Median (st.dev)	MSE[a] *SMSE[b]* Mean	MSE[c] *SMSE[d]* Median	Mean-Median (st.dev)	MSE *SMSE* Mean	MSE *SMSE* Median
300	86.483	6.456	44.588	85.773	3.755	35.496
	(1.591)	*2.559*	*24.555*	(1.463)	*2.401*	*18.006*
600	84.633	1.214	22.688	83.851	1.507	15.934
	(1.096)	*1.238*	*23.353*	(1.044)	*1.419*	*16.488*
1000	85.715	2.192	33.412	84.92	0.758	24.818
	(0.846)	*0.781*	*23.385*	(0.764)	*0.869*	*16.312*

a) MSE with respect to the population true mean wtp: 84.5
b) 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.
c) MSE with respect to the population true median wtp: 80.
d) MSE with respect to the sample true median wtp: 81.789, 79.926, and 80.954 for the 300, 600 and 1000 sample size respectively.

Table C-I.3: *Conclusions drawn from each method (% frequency)*

Method		Conclusion			
Akaike		H_f is better		H_g is better	
	300	0.097		0.903	
	600	0.015		0.984	
	1000	0.008		0.992	
Vuong		H_f is better	H_g is better		H_f and H_g equivalent
	300	0.011	0.422		0.565
	600	0.000	0.704		0.295
	1000	0.000	0.854		0.146
Cox		H_f accepted H_g rejected	H_g accepted H_f rejected	Both H_f and H_g accepted	Both H_f and H_g rejected
	300	0.050	0.259	0.232	0.457
	600	0.004	0.210	0.100	0.678
	1000	0.000	0.133	0.024	0.842

Experiment C-II: Lognormal DGP, Probit vs Weibit

Table C-II.1: *Parameter estimates[a] for lognormal DGP using H_f (normal) and H_g (extreme value) across 300 replications[b].*

Parameters	Sample Size					
	300		600		1000	
	H_f	H_g	H_f	H_g	H_f	H_g
δ_1	26.343	21.789	25.832	22.214	26.391	22.240
	(8.456)	(6.617)	(5.324)	(3.985)	(4.191)	(3.169)
δ_2	1.511	1.502	1.530	1.501	1.527	1.505
	(0.106)	(0.085)	(0.079)	(0.058)	(0.055)	(0.042)
δ_3	-2.825	-2.846	-2.876	-3.047	-2.848	-2.771
	(3.892)	(2.918)	(2.722)	(1.947)	(1.934)	(1.386)
δ_4	0.646	0.786	0.592	0.475	0.225	0.162
	(5.784)	(4.058)	(4.186)	(2.866)	(3.372)	(2.462)
v[c]	12.513	8.046	13.305	8.428	13.210	8.259
	(3.005)	(1.636)	(1.978)	(1.150)	(1.581)	(0.768)

a) Mean values and standard deviations (in parenthesis) over 300 replications.
b) The actual number of successful experiments was 285, 281 and 287 for the 300, 600 and 1000 sample sizes respectively.
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.

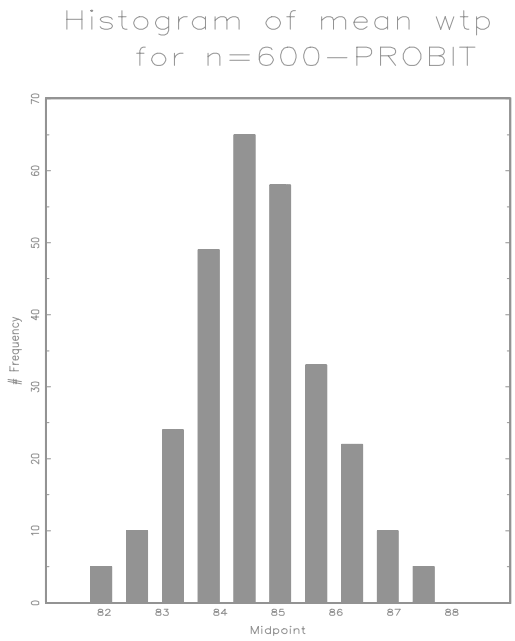
Table C-II.2: *Mean and Median[a] estimated wtp and their MSE*

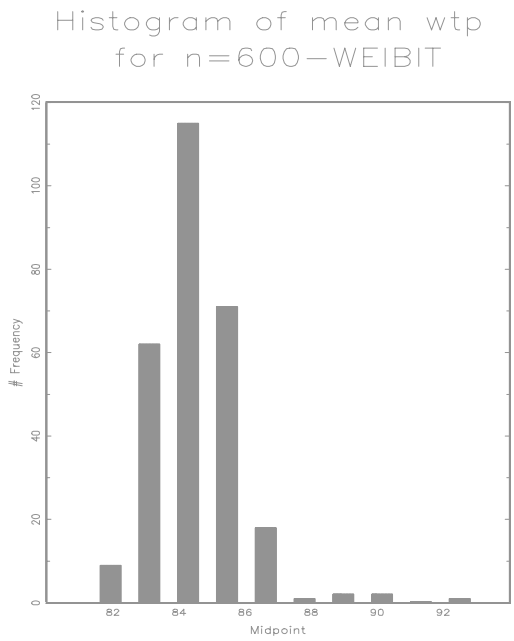
	H_f			H_g			
	Mean-Median (st.dev)	MSE[a] *SMSE[b]* Mean	MSE[c] *SMSE[c]* Median	Mean (st.dev)	MSE *SMSE*	Median (st.dev)	MSE *SMSE*
300	86.376	6.416	43.588	86.215	5.596	84.52	22.571
	(1.704)	*2.901*	*23.937*	(1.631)	*2.658*	(1.455)	*9.571*
600	84.642	1.312	22.866	84.518	1.734	82.742	8.892
	(1.138)	*1.337*	*23.533*	(1.319)	*1.742*	(1.167)	*9.288*
1000	85.780	2.298	34.102	85.511	1.588	83.771	14.689
	(0.813)	*0.763*	*23.951*	(0.753)	*0.568*	(0.671)	*8.383*

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.
d) MSE with respect to the population true median wtp: 80.
e) MSE with respect to the sample true median wtp: 81.789, 79.926, and 80.954 for the 300, 600 and 1000 sample size respectively.

Table C-II.3: *Conclusions drawn from each method (% frequency)*

Method		Conclusion			
Akaike		H_f is better		H_g is better	
	300	0.014		0.985	
	600	0.011		0.989	
	1000	0.000		1.000	
Vuong		H_f is better	H_g is better		H_f and H_g equivalent
	300	0.003	0.905		0.091
	600	0.003	0.982		0.014
	1000	0.000	1.000		0.000
Cox		H_f accepted H_g rejected	H_g accepted H_f rejected	Both H_f and H_g accepted	Both H_f and H_g rejected
	300	0.003	0.245	0.003	0.747
	600	0.000	0.078	0.003	0.918
	1000	0.000	0.020	0.000	0.980





APPENDIX 2

Derivation of the simulation based Cox test statistic.

Let F and G be two candidate cumulative distribution functions for the single bound model, with corresponding density functions f and g, which give rise to the two models for the willingness to pay below:

$H_f : Y_i = x_i'\delta_1 + \varepsilon_{1i}, \varepsilon_{1i} \sim F$ and $P(Y_i > t_i \mid x_i) = 1 - F((t_i - x_i'\delta_1)/v_1) = 1 - F_i(\beta)$, and

$H_g : Y_i = x_i'\delta_2 + \varepsilon_{2i}, \varepsilon_{2i} \sim G$ and $P(Y_i > t_i \mid x_i) = 1 - G((t_i - x_i'\delta_2)/v_2) = 1 - G_i(\gamma)$.

These two models differ only in the specification of the distribution of the error term. The test statistic for the null hypothesis that the true DGP belongs to H_f against the alternative that it belongs to H_g is given by

$$S_f(\hat{\beta}_n, \hat{\gamma}_n) = \sqrt{n} T_f / \hat{v}_f \text{ , where}$$

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

Under H_f the log-likelihood is given by,

$\log L(f(\beta)) = \sum_{i=1}^{n} I_i \log[1 - F_i(\beta)] + (1 - I_i)\log[F_i(\beta)]$, while under H_g we have,

$\log L(g(\gamma)) = \sum_{i=1}^{n} I_i \log[1 - G_i(\gamma)] + (1 - I_i)\log[G_i(\gamma)]$. Let $\hat{\beta}_n$ and $\hat{\gamma}_n$ be the corresponding maximum likelihood estimators, then the log-likelihood ratio is given by,

$$LR_n(\hat{\beta}_n, \hat{\gamma}_n) = \sum_{i=1}^{n} I_i \log\left(\frac{1 - F_i(\hat{\beta}_n)}{1 - G_i(\hat{\gamma}_n)}\right) + (1 - I_i)\log\left(\frac{F_i(\hat{\beta}_n)}{G_i(\hat{\gamma}_n)}\right).$$

In order to compute the second term in the numerator of the Cox statistic we need to find an estimator of that value of γ- call it γ^*- that maximizes the expected log-likelihood of H_g when H_f is true. The simulation method proposed by P&P(1993) consists of generating R samples of $n \times 1$ indicators according to the distribution $F(\hat{\beta}_n)$, i.e. for each of the R replications we have $I_j = (I_{1j}, \ldots I_{nj}), j = 1, \ldots R$ where

$I_{ij} = \begin{cases} 1 & \text{if } 1 - F_i(\hat{\beta}_n) > U \\ 0 & \text{otherwise} \end{cases}$, and U is drawn from a uniform distribution from (0,1). For each one of the R samples I_j is used as the new data to compute the estimator $\hat{\gamma}_j^*$ which maximizes the expression below,

$$\sum_{i=1}^{n} I_{ij} \log[1 - G_i(\gamma)] + (1 - I_{ij}) \log G_i(\gamma).$$

The average of the estimates $\hat{\gamma}_j^*$ over the R replications, i.e. $\hat{\gamma}^*(R) = \frac{1}{R} \sum_{i=1}^{R} \hat{\gamma}_j$, can be used as an estimate of γ^*, therefore we have,

$$\hat{E}_f\left(\frac{1}{n} LR_n(\hat{\beta}_n, \hat{\gamma}_n)\right) = \frac{1}{n}\left[\sum_{i=1}^{n} (1 - F_i(\hat{\beta}_n)) \log\left[\frac{1 - F_i(\hat{\beta}_n)}{1 - G_i(\hat{\gamma}^*(R))}\right] + F_i(\hat{\beta}_n) \log\left[\frac{F_i(\hat{\beta}_n)}{G_i(\hat{\gamma}^*(R))}\right]\right].$$

The denominator of the Cox statistic- $\hat{v}_f$- is obtained as the standard error (not corrected for the loss of degrees of freedom) of a regression of $I_i \log\left[\frac{1 - F_i(\hat{\beta}_n)}{1 - G_i(\hat{\gamma}_n)}\right] + (1 - I_i) \log\left[\frac{F_i(\hat{\beta}_n)}{G_i(\hat{\gamma}_n)}\right]$ on a constant and the vector of derivatives

$$\left\{\frac{\partial\left(I_i \log(1 - F_i(\hat{\beta}_n)) + (1 - I_i) \log(F_i(\hat{\beta}_n))\right)}{\partial \beta}\right\}.$$

The Cox statistic when we reverse the roles of the null and alternative hypotheses can be derived in a similar fashion by reversing the roles of *F* and *G*, and it requires simulating samples of indicators under the assumption that *G* is correct.