Testing predictive performance of binary choice models[1]

Bas Donkers
Econometric Institute and Department of Marketing
Erasmus University Rotterdam

Bertrand Melenberg
Department of Econometrics
Tilburg University

Econometric Institute Report EI 2002-01

Abstract

Binary choice models occur frequently in economic modeling. A measure of the predictive performance of binary choice models that is often reported is the hit rate of a model. This paper develops a test for the outperformance of a predictor for binary outcomes over a naive prediction method, which predicts the outcome that is most often observed. This is done for a general class of prediction models, including the well known Probit and Logit models. In many cases the test is easy to compute. The test is then applied and compared to a general test of Pesaran and Timmermann (1992) for dependence between predictors and realizations.

Keywords: Binary choice, Predictive performance, Testing, Marketing.
JEL-classification: C12, C25, M31.

[1] We thank Richard Paap for helpful discussions.

1 Introduction

Binary choice models are often used in economic modeling, either because data is available only in a binary format, think of yes/no or above/below, etc., or one is mainly interested in whether something is bought or not, whether the stock market increases or not, etc. Other situations in which binary choice models are used include labor supply decisions, product purchase decisions, market entry decisions, and many more.

The evaluation of econometric models dealing with binary outcomes is not straightforward. Various measures for the goodness of fit exist, see Windmeijer (1995) and Cramer (1999). With respect to the (in-sample) predictive performance of binary choice models, the hit rate of the model is a frequently reported measure (see Harrison, 1998, Neelamiegham and Jain, 1999, Birchenhall et al., 1999, and Shipchandler and Moore, 2000). The (in-sample) hit rate is defined as the percentage of the observations (in-sample) that is correctly predicted by the model. However, a high hit rate does not always imply good prediction properties of the model. The most important reason for this is that in many applications of binary choice models the sample has an uneven distribution among the two possible outcomes. When 90% of the observations have the same outcome, it is not difficult to predict 90% of the data correctly, while this might be rather difficult when the variable of interest has a more equal distribution.

For this reason it makes sense to use the (in-sample) hit rate of a naive predictor, which predicts the outcome that is most often observed in the data, as a benchmark for the evaluation of the predictive performance of a model under consideration.[2] Such a naive prediction results from most binary choice models when no explanatory variables (except the constant term) are used. We call a prediction model that predicts correctly a larger number of observations than the naive model a prediction model that has (positive) predictive performance. This paper develops a test for predictive performance, i.e., a test for the hypothesis that a certain model outperforms the predictive performance of the naive prediction model. This test enables researchers not only to compare the hit rate of a model with the hit rate of the naive predictor, but also to test whether the model under consideration predicts (in-sample) statistically significantly better than a naive model. It turns out that in many cases, including the regular Probit and Logit models, the test is very easy to compute.

In the literature some other benchmark models have been used, which

[2] As Birchenhall et al. (1999) note, it has become traditional to compare the results from probability models for business-cycle regimes to the "naive" predictor.

can be outperformed with lower hitrates than the simple benchmark model we propose. The simplest model predicts randomly with equal probabilities on both outcomes, resulting, on average, in 50% correctly predicted observations. A more sophisticated test is the test developed in Pesaran and Timmermann (1992), who test whether there is dependence between the predictions of a model and the outcomes (and whose test can be specialized to the binary choice case).[3] Although such dependence is, of course, a desirable feature of a predictor, it does not guarantee a high number of correct predictions. However, as we will show, a predictor that outperforms the naive model also has positive dependence between the predictor and the outcomes. Therefore, dependence is a necessary, but not a sufficient condition for predictive performance. To distinguish between the two tests we call the test of Pesaran and Timmermann (1992) a test for predictive dependence (or predictor dependence) and the test that will be developed in this paper a test for predictive performance. A detailed discussion of the Pesaran and Timmermann test for the binary choice model is presented in Franses (2000).

Much of the theoretical literature on the performance of prediction models has focused on the evaluation of out of sample predictions, see Diebold and Mariano (1995) and McCracken(2000). However, often one would like to select a prediction model for out of sample prediction, based on its in sample performance. We develop our test for the evaluation of in sample predictive performance and show how it is adapted to evaluate the out of sample predictive performance, comparable to McCracken (2000).

As an application we consider a model explaining whether a household owns a certain type of insurance. In principle, a good prediction model should predict well both in-sample and out-of-sample. Therefore, a test for predictive performance should result in similar conclusions for the estimation and the validation sample. Both Pesaran and Timmermann's and our test are applied to an estimation sample and a validation sample. In the application the difference between the two tests becomes very clear. It turns out that our test results in similar conclusions for the estimation and the validation sample, while this is not the case for the test of predictor dependence.[4]

The structure of the paper is as follows: Section 2 develops the test statistic. The application of the test is presented in Section 3, while Section 4 concludes. Formal proofs are presented in the appendix.

[3] This test is similar to the test of Henriksson and Merton (1981).

[4] This result might be due to different small sample properties, but the sample size used is representative of many empirical applications.

2 The test

2.1 Definition

Our interest is in whether a prediction model, M, outperforms a naive prediction model in terms of the number of correct predictions. Let $i = 1, \ldots, N$ denote the observations in a random sample $(y_i, x_i')'$, where $y_i \in \{0, 1\}$ denotes the realization of the endogenous variable for observation i and let $\widehat{P}_i^M \in \{0, 1\}$ denote the prediction of model M for observation i, based on x_i and possibly estimated parameters (as indicated by the hat). Define the (average) hit rate of a model, M, as:

$$H^M = \frac{1}{N} \sum_{i=1}^{N} \left(y_i \times \widehat{P}_i^M + (1 - y_i) \times (1 - \widehat{P}_i^M) \right), \tag{1}$$

so H^M is the fraction of the observations that is correctly predicted by the model.

The naive prediction model, which we shall indicate by the superscript S (of simple), predicts the same for each observation and this prediction is the realization that has been observed most frequently in the estimation sample. Without loss of generality we assume that the naive model predicts 1, so the hit rate of the naive model equals $H^S = \frac{1}{N}\Sigma_{i=1}^{N} y_i$. Our test is based on the difference in hit rates between the model under consideration and the naive model:

$$\begin{aligned} T &\equiv H^M - H^S \\ &= \frac{1}{N} \sum_{i=1}^{N} \left(y_i \times \widehat{P}_i^M + (1 - y_i) \times (1 - \widehat{P}_i^M) \right) - \frac{1}{N} \sum_{i=1}^{N} y_i \\ &= \frac{1}{N} \sum_{i=1}^{N} \left((1 - 2y_i) \times (1 - \widehat{P}_i^M) \right). \end{aligned}$$

Under appropriate conditions, the sample average hit rate H^M will converge to its population analogue $H_*^M = E\{y_i P_i^M + (1 - y_i)(1 - P_i^M)\}$ as N goes to infinity (with P_i^M the population analogue of $\widehat{P}_i^M$). Similarly, H^S will converge to $H_*^S = P\{y_i = 1\}$. The hypotheses we are interested in are $H_0 : H_*^M \leq H_*^S$, stating that the prediction model under consideration performs at best equally well as the naive model in terms of average hit rates, and $H_1 : H_*^M > H_*^S$, stating that the prediction model performs better than the naive model. The alternative hypothesis corresponds to what we define as positive prediction performance of the prediction model.

So far we have not discussed the predictor, $\widehat{P}_i^M$, and its population analogue, P_i^M. We consider predictors that can be written in the following way:

$$\widehat{P}_i^M = I_{(\hat{k},\infty)}(x_i'\widehat{\theta}).$$

Here, $I_A(.)$ denotes the usual indicator function, and the hats indicate that the parameters are possibly estimated. Thus, the predictor equals 1 if the estimated index $x_i'\hat{\theta}$ exceeds a threshold level $\hat{k}$, which can be estimated or fixed in advance. This class of predictors includes the predictors based on the standard Probit and Logit models. The population analogue of $\widehat{P}_i^M$ is defined as

$$P_i^M \equiv I_{(k_0,\infty)}(x_i'\theta_0),$$

with k_0 and θ_0 the (pseudo-)true values of $\widehat{k}$ and $\widehat{\theta}$, respectively.

2.2 Limit distribution

In order to derive the limit distribution of the test statistic under the assumption $H_*^M = H_*^S$, we make the following assumptions.

A1 $(y_i, x_i')'$, $i = 1, ..., N$ is a random sample from the population of interest.

This is a standard assumption in a cross sectional analysis. In case of time series data it can be relaxed to stationarity and ergodicity, but at the cost of a more complicated limit distribution of the test statistic.

A2 $(\widehat{\theta}', \widehat{k})'$ is a $\sqrt{N}-$consistent estimator of $(\theta_0', k_0)'$.

This assumption is twofold. First of all, we assume that $(\widehat{\theta}', \widehat{k})'$ is a consistent estimator of its (pseudo-)true value $(\theta_0', k_0)'$. In addition, we assume $\sqrt{N}-$consistency. In case of, for example, Probit or Logit models, this assumption will be satisfied.

In the following two assumptions we suppress the sub-index i referring to observation i.

A3 (i) $x = (x_1, \widetilde{x}')'$, $\theta_0 = (\theta_{01}, \widetilde{\theta}_0')'$, $\theta_{01} \neq 0$;

(ii) $F(x_1, \widetilde{x}) = P\{y = 1 \mid x_1, \widetilde{x}\}$ is continuous in its first argument;
(iii) the distribution $x_1 \mid \widetilde{x}$ has a density $g_{x_1|\widetilde{x}}$
with respect to the Lebesgue measure;
(iv) the density function $g_{x_1|\widetilde{x}}$ is continuous;
(v) $0 < P\{x_i'\theta_0 > k_0\} < 1$.

This assumption requires the existence of at least one continuously distributed variable x_1, whose (pseudo-true) coefficient θ_{01} in the predictor P_i^M is unequal to zero, where x_1, conditional upon $\widetilde{x}$ has a continuous density with respect to the Lebesgue-measure. In addition, the assumption guarantees sufficient smoothness conditions on the underlying population distribution. The final part (v) ensures that, at least asymptotically, there is variation in the predictions of the model, i.e. the models predictions, P_i^M, differ from the naive models predictions with positive probability. When the predictions of the two models are identical, the test statistic is degenerate.

As final assumption we make

A4 $F(\frac{k_0 - \widetilde{x}'\widetilde{\theta}_0}{\theta_{01}}, \tilde{x}) = 0.5$.

This assumption can be interpreted as an assumption about the threshold level, k_0, requiring that when $x_i'\theta_0$ equals the threshold level k_0, both events are equally likely to occur. This assumption greatly simplifies the limit distribution of the test statistic. Notice that in terms of, for instance, a correctly specified standard Probit model of the form $P\{y = 1 \mid x_1, \widetilde{x}\} = \Phi(\theta_{01}x_1 + \widetilde{\theta}_0'\widetilde{x})$, with Φ the standard normal distribution function, this assumption will be satisfied for $k_0 = 0$, since in this case

$$F(\frac{k_0 - \widetilde{x}'\widetilde{\theta}_0}{\theta_{01}}, \tilde{x}) = \Phi(\theta_{01}\left(\frac{k_0 - \widetilde{x}'\widetilde{\theta}_0}{\theta_{01}}\right) + \widetilde{\theta}_0'\widetilde{x}) = \Phi(k_0).$$

Thus, for the standard Probit (and Logit) models, assumption A4 requires using a threshold $k_0 = 0$, as is usually done. Examples of situations where $F(\frac{k - \widetilde{x}'\widetilde{\theta}_0}{\theta_{01}}, \tilde{x})$ can be different from 0.5 are predictors that are based on misspecified models, or, for instance, predictors that use the rule discussed in Cramer (1997). However, this rule, in general, reduces the total number of correct predictions, which contrasts with our aim to predict as well as possible. As we will discuss below, it is still feasible to obtain the asymptotic properties of the test statistic in these situations, but additional assumptions have to be made.

To motivate the limit distribution under A1-A4, suppose we would know the parameters $(\theta_0', k_0)'$ appearing in the predictor; then we could have used as predictor $P_i^M = I_{(k_0,\infty)}(x_i'\theta_0)$. The limit distribution of the statistic T under the assumption $H_*^M - H_*^S = 0$ would be

$$\sqrt{N}T \xrightarrow{d} N(0, E\{(1-2y_i)^2(1-P_i^M)^2\}).$$

But $E\{(1-2y_i)^2(1-P_i^M)^2\} = E\{(1-P_i^M)\}$, which can be estimated by $\frac{1}{N}\sum_{i=1}^N(1-P_i^M)$. In practice, we have to estimate $(\theta_0', k_0)'$ (and P_i^M). As a consequence, we will obtain a correction term in the limit distribution of the statistic T. As shown in the Appendix, this correction term vanishes under assumption A4. So, we obtain the following theorem:

Theorem
Under assumptions A1-A4 and $H_*^M = H_*^S$

$$\frac{T}{\sqrt{\frac{1}{N^2}\sum_{i=1}^N(1-\widehat{P}_i^M)}} \xrightarrow{d} N(0,1). \tag{2}$$

Proof: See Appendix

To indicate what happens if we do not impose assumption A4, consider the correction term in the limit distribution of the test statistic, as derived in the appendix

$$E\{(1-2F(\frac{k_0-\widetilde{x}'\widetilde{\theta}_0}{\theta_{01}},\tilde{x}))g_{x_1|\tilde{x}}(\frac{k_0-\widetilde{x}'\widetilde{\theta}_0}{\theta_{01}}) \times$$
$$\begin{pmatrix} -(k_0-\widetilde{x}'\widetilde{\theta}_0)/\theta_{01}^2\vdots & -\widetilde{x}'/\theta_{01}\vdots & 1/\theta_{01}\end{pmatrix}\}\sqrt{N}\begin{pmatrix}\widehat{\theta}_1-\theta_{01}\\ \widehat{\widetilde{\theta}}-\widetilde{\theta}_0\\ \widehat{k}-k_0\end{pmatrix}.$$

From this correction term we see that, without assumption A4, we need additional assumptions about the estimates $\hat{\theta} = (\widehat{\theta}_1, \widehat{\widetilde{\theta}}')'$ and $\hat{k}$ and we also need an estimate of $g_{x_1|\tilde{x}}(\frac{k-\widetilde{x}'\widetilde{\theta}_0}{\theta_{01}})$, the conditional density of x_1, given $\tilde{x}$. Moreover, depending on the choice of the model, we might have to estimate the distribution function F. Since in applications, A4 is usually imposed (under the assumption of a correctly specified model), we do not consider the case where A4 is violated in further detail.

The same test can also be used for out-of-sample testing: let sample A with sample size N_A be used to estimate $(\theta_0', k_0)'$, and let sample B with sample size N_B be used to predict whether $y_i = 0$ or $y_i = 1$, using as prediction

$\widehat{P}_i^M$, with the estimate of $(\theta_0', k_0)'$ based upon sample A. Under assumptions A1-A4, with A1 in terms of sample B, and A2 in terms of sample A, together with the assumption that $N_B/N_A \rightarrow c < \infty$, the theorem remains valid (with N replaced by N_B).[5]

2.3 Relationship between prediction performance and positive dependence.

The test we propose has a completely different starting point compared to the test developed by Pesaran and Timmermann (1992), which tests for dependence between the predictors and the quantities to be predicted. However, when a model has positive predictive performance (in terms of our definition), there has to be dependence between the predictions and the quantities to be predicted. This can be shown as follows. Suppose that the predictor (P_i^M) and the quantity to be predicted (y_i) are independent. Moreover, we still assume (without loss of generality) that the naive model predicts 1, so (if the sample size is large enough) $E\{y_i\} \geq \frac{1}{2}$. From the independence of the predictor and the predicted quantity it now follows that

$$\begin{aligned} H_*^M - H_*^S &= E\left\{(1-2y_i) \times (1-P_i^M)\right\} \\ &= E\{1-2y_i\} \times E\{1-P_i^M\}. \end{aligned}$$

Substituting $E\{y_i\} \geq \frac{1}{2}$ shows that $H_*^M - H_*^S \leq 0$; thus, there is no positive predictive performance without dependence. Conversely, however, if all you know is that there is dependence between predictor and quantity to be predicted, there is obviously no guarantee that $H_*^M - H_*^S > 0$.

Predictor dependence is therefore a necessary, but not a sufficient condition for predictive performance. Moreover, using the naive model as a benchmark model turns out to demand more from a prediction model than the benchmarks that have been used.

3 Application

The application is based on the research in Verhoef and Donkers (2001), who investigate the prediction of customer potential value of the customers of an insurance company. The potential value of a customer is the potential profit

[5] In fact, all we need is that N_B/N_A stays bounded away from ∞. Moreover, if $N_B/N_A \rightarrow 0$, the correction terms vanishes without assumption A4. See also McCracken (2000) for a comparable derivation of these out-of-sample results in a more abstract context.

a company can make from this customer (Grant and Schlesinger, 1995). An estimate of a customer's potential value can be used, for example, to increase the service level to customers with a high potential value. This might lead to a higher share of such a customer's potential value that is earned by the company, and consequently to higher profits for the company.

In the insurance industry, the potential value of a customer is the total profit earned on the insurance policies purchased by a customer. In this context predictions are needed for the ownership of the different types of insurances by the companies' customers. Based on these predictions the potential value of a customer can be computed. Moreover, information on who owns which insurance policies is highly relevant for the company under consideration, since the company is a direct writer. As a direct writer, the insurance company only has direct contact with its customers; no use is made of, for example, intermediate insurance agencies. Often the insurance company sells new insurance policies by offering a particular insurance policy to customers in a letter. The success of these mailings, which are called direct mailings, as a marketing instrument depends crucially on whether the respondents are willing to own the insurance policy offered or not.

The insurance company sells the following types of insurance policies: car, damage, disability, funeral, furniture, health, house, liability, legal aid, life, travel, and continuous travel insurance. For reasons of confidentiality, we do not identify the insurance types in the analysis. A survey has been held among the customers of the company about which types of these insurances they own. A total of 1565 customers of the insurance company have completely answered this survey. In practice one would use more observations for model estimation than for validation of the model. However, to fit in this paper, we decided to estimate the model on about half the sample – 800 observations – and use the other half – 765 observations – as a validation sample. In this way possible differences in significance levels of the tests between the estimation and validation sample are not caused by differences in the sample sizes.

For each type of insurance, a separate Probit model is used with as explanatory variables a small set of demographic variables and information on which insurances are purchased with the company under consideration. The Probit model has been estimated with Maximum Likelihood for each insurance type separately. Based on the resulting parameter estimates, a customer is predicted to own a certain type of insurance when the probability of ownership according to the Probit model exceeds 0.5. In terms of the notation of this paper, the predictor for ownership of insurance type j by customer i is $I_{(0,\infty)}(x_i'\hat{\theta}_j)$, where x_i is the vector of explanatory variables used and $\hat{\theta}_j$ is the ML-estimate of θ_j in the Probit model for ownership of insurance type

j. In this model the threshold k is not estimated but fixed at 0.

Table 1 presents sample characteristics, hit rates, and test statistics for eight different types of insurance policies that are sold by the insurance company for the estimation sample. The first row of this table reports the fraction of the households that owns the particular type of insurance, i.e., the fraction of the population with $y_i = 1$. The second row reports the fraction of households predicted to own such an insurance policy by the model, i.e., the fraction of the population with $\hat{P}_i^M = 1$. Rows three and four report the hit rates of the naive and the Probit model, H^S and H^M, respectively. The fifth row presents the standardized test statistic for predictive performance, presented in (2), while the last row presents the test statistic for predictor dependence, which asymptotically has a $N(0, 1)$ distribution, see equation (6) in Pesaran and Timmermann (1992).

From the last two rows of Table 1 it is clear that the two tests arrive at different conclusions for which insurance types the Probit model results in good predictions for the estimation sample. The test for predictive performance indicates that the Probit model has predictive performance for insurance type 5, and, depending on the desired level of significance, also for insurance type 6. The test for predictor dependence indicates that there is positive dependence between the predictions and the realizations for insurance types 1, 2, 3, 5, 6, and 8. It is clear that the test on predictive performance rejects the null hypothesis less often than the test on predictor dependence. Notice here that the tests are based on the same sample, so the test results for the different insurance types are not independent.

The question that remains is how well the model actually performs in out-of-sample prediction, which is what prediction models are generally used for. Table 2 presents the sample statistics and test results for the validation sample, like Table 1 does for the estimation sample. The results of the test for predictive performance and the test for predictor dependence for the validation sample are presented in the last two rows of the table. The test on predictive performance finds strong evidence against the null hypothesis for insurance type 5 and weak evidence for this for insurance type 6. These outcomes are perfectly in line with the findings in the estimation sample. The behavior of the test on predictor dependence is less stable across the estimation and validation sample. For four out of the six insurance types, for which the hypothesis of no dependence was rejected in the estimation sample, the test yields the same conclusion in the validation sample.

Although this in itself might not be very worrying, the test results for insurance types 1 and 8 are. For these insurance types strong evidence for positive dependence between the predictions and the realizations was found in the estimation sample. This is in contrast with the results for the validation sample, where the conclusion for these types of insurance would be that there is no dependence and, if anything, it would be negative dependence instead of positive dependence. One reason for the instability of the conclusions might be the fact that the asymptotic properties of the test are nor accurate enough in the finite sample used. However, an estimation sample of 800 observations is not that small and increasing the estimation sample to 1200 observations did not change the findings. Moreover, results in the estimation sample are highly significant. Notice also that the results for insurance type 4 in the validation sample illustrate that positive predictor dependence, although insignificant, does not imply outperformance of the naive model.

4 Conclusion

In the empirical economics literature researchers using binary choice models frequently report the hit rate of their model. The benchmarks against which the predictive performance of the model can be tested, however, only imposed low standards onto the models. Benchmark models are either random prediction, resulting in a hit rate of 50%, or random prediction, conditional on the number of times each outcome is predicted, which boils down to a test on dependence of the predictor and the realization. Models that pass these tests, however, are sometimes beaten by a very simple prediction model, which just predicts the outcome that is observed most frequently.

This paper develops a test for predictive performance where the hit rate of such a naive model is the benchmark. When a model passes this test, it also satisfies the criteria imposed by the other two tests, so the test is more demanding. The advantages of the test are that it tests a highly relevant aspect of prediction models, which is the number of correct predictions, and in most practical situations the test is very easy to compute. The large number of papers that report the hit rate of their models can now accompany this hit rate with a significance level for predictive performance of the model.

In the application there is a clear difference between our test and the test on predictor dependence as it is developed by Pesaran and Timmermann (1992). Since prediction performance is a characteristic that has relevance for both estimation and validation samples, one would expect that tests come to similar conclusions when applied to an estimation and a validation sample. For the test on predictive performance this is found in the application, while for the predictor dependence tests we found highly significant positive values in the estimation sample but negative values in the validation sample.

Appendix. Proofs

In this appendix we derive the limit distribution of the test statistic

$$\begin{aligned}\sqrt{N}T &= \sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(1-2y_i)\times\left(1-\widehat{P}_i^M\right)\\ &= \sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(1-2y_i)\times\left(1-I_{(\widehat{k},\infty)}(x_i'\widehat{\theta})\right)\\ &= \sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(1-2y_i)\times I_{(-\infty,\widehat{k}]}(x_i'\widehat{\theta})\end{aligned}$$

under Assumptions A1-A4, together with $H_*^M - H_*^S = 0$. In this appendix we shall make use of the following notation:

$\widehat{\theta}$: estimator; θ_0 : (pseudo-)true value; θ : any parameter value;

$\widehat{k}$: estimator; k_0 : (pseudo-)true value; k : any parameter value;

$x = (x_1, \widetilde{x}')'$; $\theta = (\theta_1, \widetilde{\theta}')'$; $\theta_0 = (\theta_{01}, \widetilde{\theta}_0')'$;

$h_\theta = (h_{\theta 1}, \widetilde{h}_\theta')'$; $h_{\theta n} = (h_{\theta n1}, \widetilde{h}_{\theta n}')'$ (of the same dimension as θ);

P_N : empirical distribution function (y, x);

P : (population) distribution (y, x); P_x : (population) distribution x;

$P(y = 1 \mid x) = F(x) = F(x_1, \widetilde{x})$;

$dP_x = g_{x_1|\widetilde{x}}dx_1 dP_{\widetilde{x}}$;

$\widetilde{X}$: Support $P_{\widetilde{x}}$.

First, consider the following decomposition of the test statistic:

$$
\begin{aligned}
&\sqrt{N}T = \sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(1-2y_i) \times I_{(-\infty,\widehat{k}]}(x_i'\widehat{\theta}) \\
&= \sqrt{N}\int(1-2y) \times I_{(-\infty,\widehat{k}]}(x'\widehat{\theta})dP_N \\
&= \sqrt{N}\int((1-2y) \times I_{(-\infty,\widehat{k}]}(x'\widehat{\theta}) - (1-2y) \times I_{(-\infty,k_0]}(x'\theta_0))d(P_N - P) \\
&\quad + \sqrt{N}\int((1-2y) \times I_{(-\infty,\widehat{k}]}(x'\widehat{\theta}) - (1-2y) \times I_{(-\infty,k_0]}(x'\theta_0))dP \\
&\quad + \sqrt{N}\int(1-2y) \times I_{(-\infty,k_0]}(x'\theta_0)dP_N.
\end{aligned}
$$

In the sequel, we shall investigate the behaviour of the three terms in this decomposition.

We start with the first term

$$
\sqrt{N}\int((1-2y) \times I_{(-\infty,\widehat{k}]}(x'\widehat{\theta}) - (1-2y) \times I_{(-\infty,k_0]}(x'\theta_0))d(P_N - P).
$$

This term converges to zero in probability. To show this, we shall apply, for instance, Pakes & Pollard (1989, Lemma (2.17)): The set of functions $\{f(y,x;\theta,k) = (1-2y) \times I_{(-\infty,0]}(x'\theta - k);\ \theta \in R^k,\ k \in R\}$ is clearly a Euclidean class (terminology Pakes & Pollard), with envelope $G(y,x) = 1$. Moreover, assuming $\theta_{01} > 0$, and θ_1 sufficiently close to θ_{01}, so that we can assume $\theta_1 > 0$ as well, we have

$$
\begin{aligned}
&\int[(1-2y)\times I_{(-\infty,0]}(x'\theta-k)-(1-2y)\times I_{(-\infty,0]}(x'\theta_0-k_0)]^2dP\\
&=\int(1-2y)^2[I_{(-\infty,0]}(x'\theta-k)-I_{(-\infty,0]}(x'\theta_0-k_0)]^2dP\\
&=\int\left[I_{(-\infty,0]}(x'\theta-k)-2I_{(-\infty,0]}(x'\theta-k)\times I_{(-\infty,0]}(x'\theta_0-k_0)+I_{(-\infty,0]}(x'\theta_0-k_0)\right]dP\\
&=\int_{\widetilde{X}}\left[\int\limits_{-\infty}^{(k-\widetilde{x}'\widetilde{\theta})/\theta_1}g_{x_1|\widetilde{x}}(x_1)dx_1-2\int\limits_{-\infty}^{\min\{(k-\widetilde{x}'\widetilde{\theta})/\theta_1,(k_0-\widetilde{x}'\widetilde{\theta}_0)/\theta_{10}\}}g_{x_1|\widetilde{x}}(x_1)dx_1+\right.\\
&\qquad\left.\int\limits_{-\infty}^{(k_0-\widetilde{x}'\widetilde{\theta}_0)/\theta_{10}}g_{x_1|\widetilde{x}}(x_1)dx_1\right]dP_{\widetilde{x}},
\end{aligned}
$$

which clearly converges to 0 as $(\theta',k)'\longrightarrow(\theta_0',k_0)'$, showing $L_2(P)-$continuity

(in the terminology of Pakes & Pollard). Thus, according to their Lemma (2.17), we have for each sequence of positive numbers $\{\delta_N\}$ converging to 0 that

$$
\sup_{|(\theta',k)'-(\theta_0',k_0)'|<\delta_N}[\sqrt{N}\int((1-2y)\times I_{(-\infty,k]}(x'\theta)-
(1-2y)\times I_{(-\infty,k_0]}(x'\theta_0))d(P_N-P)\big]\xrightarrow{p}0.\quad(*)
$$

Define

$$
Q_N=\sqrt{N}\int((1-2y)\times I_{(-\infty,\widehat{k}]}(x'\widehat{\theta})-(1-2y)\times I_{(-\infty,k_0]}(x'\theta_0))d(P_N-P).
$$

Then

$$
\begin{aligned}
&P\left\{|Q_N|>\epsilon\right\}\\
&\leq P\left\{|Q_N|>\epsilon,\ |\ (\widehat{\theta}',\widehat{k})'-(\theta_0',k_0)'\ |<\delta_N\right\}\\
&\qquad\qquad+P\left\{|Q_N|>\epsilon,\ |\ (\widehat{\theta}',\widehat{k})'-(\theta_0',k_0)'\ |\geq\delta_N\right\}\\
&\leq P\{\sup_{|(\theta',k)'-(\theta_0',k_0)'|<\delta_N}|Q_N|>\epsilon\}+P\{|\ (\widehat{\theta}',\widehat{k})'-(\theta_0',k_0)'\ |\geq\delta_N\}.
\end{aligned}
$$

The first term at the right hand side converges to zero due to $(*)$, the second term converges to zero due to the $\sqrt{N}-$consistency of $(\widehat{\theta}', \widehat{k})'$.

Next, consider the second term in the decomposition

$$\sqrt{N} \int ((1-2y) \times I_{(-\infty,\widehat{k}]}(x'\widehat{\theta}) - (1-2y) \times I_{(-\infty,k_0]}(x'\theta_0))dP.$$

For $\theta_{01} > 0$, this term has the same limit distribution as

$$E\{(1-2F(\frac{k_0 - \widetilde{x}'\widetilde{\theta}_0}{\theta_{01}}, \widetilde{x}))g_{x_1|\widetilde{x}}(\frac{k_0 - \widetilde{x}'\widetilde{\theta}_0}{\theta_{01}}) \begin{pmatrix} -(k_0 - \widetilde{x}'\widetilde{\theta}_0)/\theta_{01}^2 & -\widetilde{x}'/\theta_{01} & 1/\theta_{01} \end{pmatrix}\} \times$$
$$\sqrt{N} \begin{pmatrix} \widehat{\theta}_1 - \theta_{01} \\ \widehat{\widetilde{\theta}} - \widetilde{\theta}_0 \\ \widehat{k} - k_0 \end{pmatrix}.$$

This follows from (Hadamard) differentiability of

$$(\theta, k) \rightarrow \int ((1-2y) \times I_{(-\infty,0]}(x'\theta - k))dP,$$

with derivative

$$\begin{aligned} (h_\theta, h_k) \quad \longrightarrow \quad & E\left\{\left(1 - 2F(\frac{k - \widetilde{x}'\widetilde{\theta}_0}{\theta_{01}}, \widetilde{x})\right) g_{x_1|\widetilde{x}}(\frac{k - \widetilde{x}'\widetilde{\theta}_0}{\theta_{01}}) \times \right. \\ & \left. \begin{pmatrix} -(k - \widetilde{x}'\widetilde{\theta}_0)/\theta_{01}^2 & -\widetilde{x}'/\theta_{01} & 1/\theta_{01} \end{pmatrix} \right\} \times \begin{pmatrix} h_{\theta 1} \\ h_{\widetilde{\theta}} \\ h_k \end{pmatrix} \end{aligned}$$

Indeed, let $(h_{\theta n}, h_{kn}) \longrightarrow (h_\theta, h_k)$ and $\varepsilon \downarrow 0$, then

$$\frac{1}{\varepsilon}[\int((1-2y)\times I_{(-\infty,k+\varepsilon h_{kn}]}(x'(\theta_0+\varepsilon h_{\theta n})))dP-\int((1-2y)\times I_{(-\infty,k_0]}(x'\theta_0))dP]$$
$$=\frac{1}{\varepsilon}[\int((1-2F(x))\times I_{(-\infty,k+\varepsilon h_{kn}]}(x'(\theta_0+\varepsilon h_{\theta n})))dP_x-$$
$$\int((1-2F(x))\times I_{(-\infty,k_0]}(x'\theta_0))dP_x]$$
$$=\frac{1}{\varepsilon}[\int_{\widetilde{X}}\left[\int_{-\infty}^{((k+\varepsilon h_{kn})-\widetilde{x}'(\widetilde{\theta}_0+\varepsilon h_{\widetilde{\theta}n}))/(\theta_{01}+\varepsilon h_{\theta n1})}((1-2F(x))g_{x_1|\widetilde{x}}(x_1)dx_1-\right.$$
$$\left.\int_{-\infty}^{(k_0-\widetilde{x}'\widetilde{\theta}_0)/\theta_{01}}((1-2F(x))g_{x_1|\widetilde{x}}(x_1)dx_1\right]dP_{\widetilde{x}}\longrightarrow$$
$$E\{(1-2F(\frac{k_0-\widetilde{x}'\widetilde{\theta}_0}{\theta_{01}},\widetilde{x}))g_{x_1|\widetilde{x}}(\frac{k_0-\widetilde{x}'\widetilde{\theta}_0}{\theta_{01}})\begin{pmatrix}-(k-\widetilde{x}'\widetilde{\theta}_0)/\theta_{01}^2 & -\widetilde{x}'/\theta_{01} & 1/\theta_{01}\end{pmatrix}\}\begin{pmatrix}h_{\theta 1}\\ h_{\widetilde{\theta}}\\ h_k\end{pmatrix}.$$

Finally, consider the term

$$\sqrt{N}\int(1-2y)\times I_{(-\infty,k_0]}(x_i'\theta_0)dP_N=\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(1-2y_i)\times I_{(-\infty,k_0]}(x_i'\theta_0).$$

Notice that $E\{(1-2y_i)\times I_{(-\infty,k_0]}(x_i'\theta_0)\}=H_*^M-H_*^S$, so that, as discussed in the main text, under the assumption $H_*^M-H_*^S=0$ this term converges in distribution to $N(0,E\{1-P_i^M\})$.

To complete the proof, we need to verify

$$\frac{1}{N}\sum_{i=1}^{N}\left(1-\widehat{P}_i^M\right)\xrightarrow{p}E\{1-P_i^M\}.$$

Consider

$$
\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{N}\left(1-\widehat{P}_i^M\right)-E\{1-P_i^M\}\right| \\
&\leq \left|\frac{1}{N}\sum_{i=1}^{N}\left(1-\widehat{P}_i^M\right)-E\{1-\widehat{P}_i^M\}\right| + \left|E\{1-\widehat{P}_i^M\}-E\{1-P_i^M\}\right| \\
&\leq \sup_{(\theta',k)}\int\left(1-I_{(-\infty,k]}(x'\theta)\right)d(P_N-P) + \left|E\{1-\widehat{P}_i^M\}-E\{1-P_i^M\}\right|.
\end{aligned}
$$

The first term on the right hand side converges to zero, due to lemma (2.8) of Pakes and Pollard (1989), since the set of functions $\{g(y,x;\theta,k) = 1 - I_{(-\infty,0]}(x'\theta - k);\ \theta \in R^k,\ k \in R\}$ is clearly a Euclidean class (terminology Pakes & Pollard), with envelope $H(y,x) = 1$. The second term of the right hand side converges to zero due to consistency of $(\widehat{\theta}', \widehat{k})'$ in combination with continuity of $(\theta', k)' \to E\{1 - I_{(k,\infty)}(x'\theta)\}$. Indeed,

$$
\begin{aligned}
E\{1-I_{(k,\infty)}(x'\theta)\} &= \int I_{(-\infty,0]}(x'\theta-k)dP \\
&= \int_{\widetilde{X}}\left[\int\limits_{-\infty}^{(k-\widetilde{x}'\widetilde{\theta})/\theta_1} g_{x_1|\widetilde{x}}(x_1)dx_1\right]dP_{\widetilde{x}},
\end{aligned}
$$

from which continuity follows, given our assumptions.

5 References

Birchenhall, C.R., H. Jessen, D.R. Osborn, and P. Simpson (1999), "Predicting U.S. Business-Cycle Regimes," Journal of Business & Economic Statistics, 17, 313-323.

Cramer, J.S. (1999), "Predictive Performance of Binary Logit Models in Unbalanced Samples," The Statistician, 48, 85-94.

Diebold, F.C., and R.S. Mariano (1995), "Comparing presictive performance," Journal of Business & Economic Statistics, 13, 253-263.

Franses, P.H. (2000), "A test for the hit rate in binary response models," International Journal of Market Research, 42, 239-245.

Grant, H.W.H. and L.A. Schlesinger (1995), "Realize Your Customers Full Profit Potential", Harvard Business Review, 73, 59-75

Harrison, G.W. (1998), "Mortgage Lending in Boston: a Reconsideration of the evidence," Economic Inquiry, 36, 29-38.

Henriksson, R.D. and R.C. Merton (1981), "On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills," Journal of Business, 54, 513-533.

McCracken, M.W. (2000), "Robust out-of-sample inference," Journal of Econometrics, 99, 195-223.

Neelamiegham, R., and D. Jain (1999), "Consumer Choice Process for Experience Goods: an Econometric Model and Analysis," Journal of Marketing Research, 36, 373-386.

Pakes and Pollard (1989), "Simulation and the asymptotics of optimization estimators", Econometrica, 57, 1027-1057.

Pesaran, M.H., and A. Timmermann (1992), "A simple nonparametric test of predictive performance," Journal of Business & Economic Statistics, 10, 461-465.

Shipchandler, Z.E., and J.S. Moore (2000), "Factors Influencing Foreign Firm Performance in the U.S. Market," American Business Review, 18, 62-68.

Verhoef, P.C. and B. Donkers (2001), "Estimating customer potential with an application to the insurance industry," Decision Support Systems, forthcoming

Table 1: Descriptive statistics and test results for eight types of insurance policies. Estimation sample. (N=800)

	Type of insurance							
	1	2	3	4	5	6	7	8
Sample fraction with y_i=1	0.875	0.691	0.646	0.626	0.595	0.504	0.431	0.393
Sample fraction with P_i^M=1	0.999	0.995	0.955	0.963	0.698	0.448	0.053	0.020
H^S	0.875	0.691	0.646	0.626	0.595	0.504	0.569	0.607
H^M	0.876	0.694	0.649	0.629	0.645	0.551	0.574	0.615
Predictive performance test	0.100	0.127	0.119	0.116	2.222*	1.907*	0.215	0.339
Predictor dependence test	2.649**	1.916*	2.236*	1.184	6.745**	2.940**	1.565	2.443**

Note: * significant at 5%, ** significant at 1%

Table 2: Descriptive statistics and test results for eight types of insurance policies. Validation sample. (N=765)

	Type of insurance							
	1	2	3	4	5	6	7	8
Sample fraction with y_i=1	0.895	0.718	0.647	0.651	0.550	0.514	0.420	0.429
Sample fraction with P_i^M=1	0.995	0.991	0.950	0.953	0.660	0.452	0.054	0.020
H^S	0.895	0.718	0.647	0.651	0.550	0.514	0.580	0.571
H^M	0.890	0.719	0.650	0.638	0.618	0.554	0.571	0.567
Predictive performance	-0.447	0.068	0.122	-0.612	2.804**	1.607	-0.391	-0.166
Predictor dependence	-0.686	1.708*	2.296*	0.156	6.001**	3.091**	-0.066	-0.227

Note: * significant at 5%, ** significant at 1%