Using the Estimated AUC to Test the Adequacy of Binary Predictors*

Yu-Chin Hsu† Robert P. Lieli‡

October, 2015

Abstract

We consider using parametric and nonparametric estimators of AUC, the area under an ROC curve, to test the hypothesis that a predictive index combined with a range of cutoffs performs no better than pure chance in forecasting a binary outcome. We show that if the predictive index is an estimated, rather than fixed, function of the underlying covariates, then testing AUC=1/2 based on standard asymptotic results (such as the limiting distribution of the Mann-Whitney U-statistics) can lead to severe overrejection of the null. The problem arises because first stage estimation overfits the model for the index in a way that artificially boosts the in-sample performance of the classifiers. Under these circumstances the bootstrap also fails to properly approximate the null distribution of the empirical AUC. We propose tests with chi-squared critical values that control size asymptotically.

Keywords: binary classification, ROC curve, area under ROC curve, model diagnostics, hypothesis testing

JEL codes: C12, C25, C52, C53

*An early version of this paper was made public under the title "Inference for ROC Curves Based on Estimated Predictive Indices: A Note on Testing AUC=0.5"

†Institute of Economics, Academia Sinica, Taiwan. Email: ychsu@econ.sinica.edu.tw

‡Department of Economics, Central European University, Budapest and Magyar Nemzeti Bank. Email: lielir@ceu.hu

1 Introduction

The receiver operating characteristic (ROC) curve is a widely used statistical tool to assess the quality of binary forecasts. As the name suggests, it originates from the engineering literature on signal detection (e.g., Egan 1975, Green and Swets 1966), and is now routinely employed in fields such as medical diagnostics, meteorology, pattern recognition, etc. In recent years ROC analysis has become increasingly common in financial and economic applications as well (Berge and Jorda 2011; Carole and Putnins 2014; Jorda and Taylor 2011, 2013; Lahiri and Wang 2013; Sreedhar and Dittmar 2010).

Though its relevance is often debated (e.g., Hand 2009), the area under the ROC curve (AUC) is ubiquitous as a statistic to characterize the overall predictive power of binary forecasts. It gives the probability that for a randomly chosen pair from the subpopulation of positive and negative outcomes, the positive outcome is associated with a higher value of the predictive index used in constructing the ROC curve (Bamber 1975, Hanley and McNeil 1982). If the index is statistically independent of the outcome, this probability is 0.5 by symmetry, while for a perfect predictor it is unity. Intermediate values of AUC closer to one are generally taken as a signal of better overall predictive power.

The literature on the statistical properties of the empirical ROC curve, and AUC in particular, is now rather large, but is scattered around in journals of many different disciplines. The typical framework for statistical inference takes observations on the binary outcome and a scalar predictive index as the raw data. In such a setting the area under the empirical ROC curve is closely related to the Mann-Whitney U-statistic, whose asymptotic distribution theory is well understood (see, e.g., Lehmann 1999, Ch. 6). Nevertheless, in many practical applications, especially in economics, binary forecasts are derived from predictive indices that are themselves outputs of a statistical model with estimated parameters. For example, given a vector of potential predictors, a researcher may estimate the conditional probability of a positive outcome using a logit regression, and then use the fitted probabilities to construct the ROC curve. The goal of this paper is to demonstrate that the first stage estimation step has nontrivial implications for how one should conduct in-sample tests of the hypothesis that AUC=0.5 versus AUC>0.5. This is a fundamental test akin to the test

of the "overall significance" of a linear regression model. In the latter case, acceptance of the null means that the regressors are not capable of explaining a significant portion of the variation in the dependent variable; in the former, acceptance is interpreted as the predictive model being no better than flipping an unbalanced coin. Hence, rejection of the null is as low a bar as one can set for the usefulness of the proposed predictor.

Our first contribution is to provide insight, through analytical examples and Monte Carlo simulations, into why and to what extent traditional inference procedures for AUC fail under the null of independence. Perhaps most surprisingly, this includes the failure of the bootstrap. As we show, using the same data set for estimation and inference causes severe size inflation, because first stage estimation overfits the model for the predictive index in a way that artificially boosts in-sample classification performance. We contend that presently these points are not well understood in the econometrics research community. For instance, after surveying the relevant literature, Jorda and Taylor (2011) state that in the presence of estimated parameters asymptotic normality of the empirical AUC remains valid "except that the variance of the AUC would need to reflect parameter estimation uncertainty as well—an issue that can be easily resolved in practice by using the bootstrap". While these claims are generally true if the model in question has nontrivial predictive power, they fail under independence. The following conclusion reached by *ibid.* is therefore wrong: "[t]o test against the random null, the large sample results [...] can be used to construct Wald tests for the null that AUC=1/2. If the [test] is based on a model with estimated parameters, then the bootstrap percentiles [...] should be used." The problem is analogous to testing the value of the autoregressive parameter in an AR(1) model. If the process is stationary under the null, the OLS estimator is asymptotically normal, but this distribution is no longer valid under the null of a unit root.

Our second contribution is more constructive. We propose asymptotic tests of no predictive power (implying AUC=1/2) that are valid when classification is based on an estimated predictive index. The first stage model may be a linear regression, a logit or probit regression with a linear index, or a similar conditional probability model with some other c.d.f. as the link function. The test statistic is easily computed from the estimated AUC (either the usual

nonparametric estimate or a parametric one), the sample proportion of positive outcomes, and the sample size. The asymptotic null distribution is chi-squared with degrees of freedom equal to the number of explanatory variables in the first stage model. We provide a complete proof for the asymptotic null distribution in the parametric case, but for the nonparametric test the proof, for now, is restricted to a special case. In fact, the Monte Carlo evidence suggests that for the nonparametric test the proposed null distribution is not generally valid without further conditions on the distribution of the predictors. Nevertheless, the same set of results also show that if there are size distortions, they only make the test somewhat more conservative, and are not very large in magnitude. Therefore, the test can still serve as a useful rule of thumb in practice. The derivation of more precise theoretical results for the nonparametric test is of course work in progress.

The papers most closely related to our setup are Demler, Pencina and D'Agostino (2011, 2012). In the former, the authors relate the statistical significance of additional covariates used in linear discriminant analysis (LDA) to statistically significant improvements in AUC under the assumption of joint normality. Given the straightforward mapping between regression and LDA, their setup is in a way more general than the one considered here, though they do not explicitly state whether their results extend to the situation when the additional covariates constitute the full set. While *ibid.* emphasize that their general results are dependent on normality, we argue that for the special case at hand the normality assumption is not essential. In particular, as stated above, a Wald-type test computed on the basis of a parametric AUC estimator is asymptotically valid even if the normality assumption underlying the estimator does not hold. In addition, we also study the nonparametric version of the test computed directly from the area under the empirical ROC curve.

The second paper by Demler et al. is concerned with using the DeLong et al. (1988) test to compare the AUCs of nested models with estimated parameters. However, this comparison does not include the special case when the smaller model is degenerate. The authors observe the failure of asymptotic normality of the estimated AUC differential under the null of no improvement, but they do not provide such detailed insight into the underlying reasons as we do here. Interestingly, they find that asymptotic non-normality in their setup causes the

DeLong test to be overly conservative in contrast to the overrejection problem documented here.

Finally, neither of the Demler et al. papers considers the bootstrap distribution of the empirical AUC and shows that it does not consistently approximate the actual sampling distribution under the null of no predictive power and in the presence of estimated parameters.

The rest of the paper is organized as follows. Section 2 introduces ROC curves, discusses the estimation of AUC, and summarizes standard asymptotic results available in the literature. In Section 3 we present analytical examples and Monte Carlo evidence showing that standard asymptotic theory generally fails under the conditions considered here. Section 4 does the same for the bootstrap. The proposed tests of the AUC=1/2 null hypothesis are presented in Section 5.

2 Binary prediction and the ROC curve: basic definitions and results

2.1 Binary predictors and cutoff rules

Let $Y \in \{0, 1\}$ be a binary outcome. Given a k-dimensional vector X of covariates (predictors), a classifier is a function $\hat{Y} = \hat{y}(X)$ that maps the possible values of X into $\{0, 1\}$, i.e., produces an outcome forecast based on X. (We will also refer to classifiers as binary predictors or decision rules.) Classifiers based on "cutoff rules" arise naturally in many situations and are particularly important in practice. These are of the form $\hat{Y} = 1(g(X) > c)$, where $g(X)$ is a scalar predictive index based on X and c is an appropriate threshold or cutoff.

Example 1 If a scalar measurement X exceeds a certain threshold c, it is classified as a signal otherwise it is treated as noise. Here $g(X) = X$. ∎

Example 2 Suppose that there is a loss function $\ell(\hat{y}, y)$, $\hat{y}, y \in \{0, 1\}$, associated with a binary prediction problem and the goal is to construct a decision rule that minimizes expected

loss. In particular, for any given value of X the forecaster wants to solve the problem

$$\min_{\hat{y}\in\{0,1\}} E[\ell(\hat{y}, Y) \mid X].$$

If $\ell(1,1) < \ell(0,1)$ and $\ell(0,0) < \ell(1,0)$, then it is easy to show that the optimal decision rule is of the form

"predict the outcome 1 if and only if $P(Y = 1 \mid X) > c$",

where the cutoff $c \in (0,1)$ depends only on the loss function ℓ (e.g., Elliott and Lieli 2013). Thus, cutoff rules are theoretically optimal in a wide range of settings, and the conditional probability $p(X) = P(Y = 1 \mid X)$ serves as an optimal predictive index. ∎

In Example 1 the predictive index is a primitive—it is raw data that the researcher directly observes. In Example 2 the optimal index is a theoretically fixed function of the covariates. Nevertheless, in practice $p(X)$ is typically unknown, and needs to be estimated, e.g., by a logit regression.

2.2 The population ROC curve

Two quantities that characterize the performance of a classifier are the true positive rate (TPR) and false positive rate (FPR), defined as

$$TPR = P(\hat{Y} = 1 \mid Y = 1) \text{ and } FPR = P(\hat{Y} = 1 \mid Y = 0).$$

Given a predictive index $g(X)$, consider the family of cutoff rules $\{1(g(X) > c) : c \in \mathbb{R}\}$. For a fixed value of c, the associated TPR and FPR are given by

$$TPR(c) = P(g(X) > c \mid Y = 1) \text{ and } FPR(c) = P(g(X) > c \mid Y = 0).$$

The set of points traced out in the unit square $[0,1] \times [0,1]$ by the pair $(FPR(c), TPR(c))$ as c varies between minus infinity and infinity is called the (population) receiver operating characteristic (ROC) curve associated with the index $g(X)$. See Figure 1 for an illustration.[1]

[1] In light of Example 2, one can think of the ROC curve as a loss function free way of evaluating the predictive power of the index $g(X)$ for Y. More precisely, the ROC curve considers all possible loss functions (cutoffs) simultaneously; hence, it is an appropriate evaluation tool in situations in which it is not possible or desirable to commit to a specific loss function.

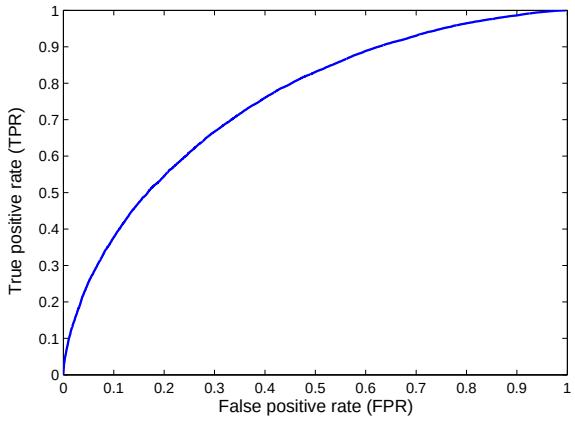

Figure 1: ROC curve

The ROC curve is clearly invariant to strictly increasing transformations of the predictive index, i.e., the rules $1(g(X) > c)$ and $1(f(g(X)) > c)$ produce the same ROC curves for $f(\cdot)$ strictly increasing.

For a given value of FPR one would generally like to maximize TPR and, conversely, for a given TPR one would like to minimize FPR. Thus, the "bulgier" the ROC curve is toward the northwest, the stronger the general predictive power of the underlying index is. This is the intuitive reason why the area under the ROC curve, abbreviated as AUC or AUROC, can be considered as an overall measure of classification performance. More precisely, let Z_1 and Z_0 be two independent random variables with $Z_1 \sim g(X)|Y = 1$ and $Z_0 \sim g(X)|Y = 0$. Then

$$\text{AUC} = P(Z_1 > Z_0) + 0.5P(Z_0 = Z_1);$$

see, e.g., Bamber (1975). If, in particular, the (conditional) distribution of $g(X)$ is continuous, then AUC is simply $P(Z_1 > Z_0)$.

Closed form expressions for AUC are also available under appropriate distributional assumptions. For example, if $g(X)|Y = j \sim N(\mu_j, \sigma_j^2)$, $j = 0, 1$, then

$$\text{AUC} = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2 + \sigma_1^2}}\right), \tag{1}$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution; see, e.g., Pepe (2003).

Suppose that $g(X)$ has no predictive power, i.e. that it is independent of Y. In this case

$$FPR(c) = TPR(c) = P(g(X) > c),$$

so that the ROC curve runs along the main diagonal of the unit square and the area under it is 1/2. Intuitively, one can imagine tracing out this degenerate ROC curve by flipping unbalanced coins with various head probabilities and predicting $Y = 1$ if head actually occurs. Hence, ROC curves situated below the main diagonal of the unit square represent inadmissible decision rules in that they perform uniformly worse than classification based on pure chance. Nevertheless, in such cases $g(X)$ is actually informative about Y. The problem is that the cutoff rules $1(g(X) > c)$ use this information the wrong way—they get the outcome labels reversed. Switching to the decision rule $1(-g(X) > c)$ will cause the ROC curve to be reflected over the point $(1/2, 1/2)$ and the reflected curve will run above the diagonal. Only an ROC curve coinciding with the diagonal means that $g(X)$ is completely uninformative about Y.

2.3 The empirical ROC curve and estimators of AUC

The population quantities defined in the preceding section have natural sample counterparts. Given an i.i.d. sequence of observations $\{(X_i, Y_i)\}_{i=1}^n$, one can construct the empirical ROC curve based on sample estimates of TPR and FPR:[2]

$$\widehat{TPR}(c) = \frac{1}{n_1}\sum_{i=1}^{n} 1(g(X_i) > c)Y_i \text{ and } \widehat{FPR}(c) = \frac{1}{n_0}\sum_{i=1}^{n} 1(g(X_i) > c)(1 - Y_i),$$

where $n_1 = \sum_{i=1}^n Y_i$ and $n_0 = \sum_{i=1}^n (1 - Y_i)$. As c varies between plus and minus infinity, the pair $(\widehat{FPR}(c), \widehat{TPR}(c))$ takes on a finite number of values in the unit square in a successive manner. If successive points are connected by straight line segments, one obtains the empirical ROC curve, which is typically a step function.[3]

[2]We will maintain the random sampling assumption throughout the rest of the paper along with the assumption that X has a nonsingular variance-covariance matrix.

[3]Two successive points are typically connected by either a horizontal or a vertical line as only one of the coordinates will differ. However, if the $Y = 1$ and $Y = 0$ subsamples contain observations that share

Let eAUC denote the area under the empirical ROC curve (the prefix e stands for empirical). This statistic has an interpretation analogous to the population AUC. Let $\{Z_{0,i}\}_{i=1}^{n_0}$ and $\{Z_{1,j}\}_{j=1}^{n_1}$ denote the values of the predictive index $g(X)$ over the $Y = 0$ and $Y = 1$ subsamples, respectively. Define

$$\hat{U} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} 1(Z_{0,i} < Z_{1,j}) \text{ and } \hat{U}' = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} 1(Z_{0,i} = Z_{1,j}) \tag{2}$$

so that $\hat{U}$ is the sample proportion of $(Z_{0,i}, Z_{1,j})$ pairs with $Z_{0,i} < Z_{1,j}$ and $\hat{U}'$ is the sample proportion of ties between the two groups. In general, $e\text{AUC} = \hat{U} + 0.5\hat{U}'$, but of course if $g(X)$ is a continuous random variable, then $e\text{AUC} = \hat{U}$ with probability one.

The quantity $\hat{U}$ is known as the (two-sample) Mann-Whitney U-statistic and its asymptotic theory is well-developed. Drawing on the pertaining statistics literature, Bamber (1975) states a general asymptotic normality result for eAUC valid for continuous as well as discrete $g(X)$; specifically, $\sqrt{n}(e\text{AUC} - \text{AUC}) \overset{a}{\sim} N(0, V)$, where the asymptotic variance V is the limit of n times the expression given by Bamber's equation (6). Thus, given a consistent estimator $\hat{V}_n$ for V, one can use the pivotal statistic

$$\frac{\sqrt{n}(e\text{AUC} - \text{AUC})}{\sqrt{\hat{V}_n}} \overset{a}{\sim} N(0, 1), \tag{3}$$

to test hypotheses about AUC. Here we will utilize two special cases of this result: (i) if $g(X)$ is continuous and Y is independent of $g(X)$, then one can take

$$\hat{V}_n = \frac{n^2}{12 n_0 n_1}, \tag{4}$$

and (ii) if $g(X) \in \{0, 1\}$ and Y is independent of $g(X)$, then

$$\hat{V}_n = \frac{n^2 \hat{p}_x (1 - \hat{p}_x)}{4 n_0 n_1}, \tag{5}$$

where $\hat{p}_x = n^{-1} \sum_{i=1}^{n} g(X_i)$.

As an alternative to eAUC, one can employ a parametric estimator based on a closed formed expression for AUC. In particular, we will subsequently consider the sample analog

the same value for the predictive index $g(X)$), then some connecting line segments will have a finite positive slope. In this case there are slightly different ways to construct the empirical ROC curve, but this is the most conventional. See Fawcett (2004) for details.

of (1), denoted as pAUC (the prefix p stands for parametric). For $g(\cdot)$ fixed, the asymptotic theory of pAUC is entirely standard, and under independence of X and Y it yields

$$\sqrt{n}(p\text{AUC} - 1/2) \overset{a}{\sim} N\left(0, \frac{n^2}{4\pi n_0 n_1}\right). \tag{6}$$

A quick proof is given in Appendix A for easy reference.

3 The possible failure of standard asymptotics

The standard asymptotic theory for eAUC and pAUC presented in the previous section does not automatically extend to all situations in which the function $g(\cdot)$ depends on parameters estimated from the same data set. While the estimation effect is asymptotically negligible in most situations,[4] we will next show that asymptotic normality does fail under the additional hypothesis that X and Y are independent. In this case the bootstrap also fails to properly approximate the relevant null distribution. These claims are demonstrated through simple but insightful analytical examples and Monte Carlo simulations.

3.1 Analytical examples

Let X be a scalar predictor and consider the empirical ROC curves induced by the decision rules

$$\text{Rule}(+X)\text{: } \hat{Y} = 1(X_i > c) \text{ and } \text{Rule}(-X)\text{: } \hat{Y} = 1(-X_i > c).$$

Let $e\text{AUC}_X$ denote the area under the former and $e\text{AUC}_{-X}$ the area under the latter. As mentioned in Section 2.2, the two curves are symmetric about the point $(1/2, 1/2)$ so that $e\text{AUC}_X = 1 - e\text{AUC}_{-X}$. For $g(x) = \beta_0 + \beta_1 x$, the decision rule $1(g(X_i) > c)$ is clearly equivalent to $1(X_i > c)$ for any $\beta_1 > 0$ and to $1(-X_i > c)$ for any $\beta_1 < 0$.

Suppose that we let an OLS regression "pick" the value of β_1; that is, we employ the decision rule $1(\hat{g}(X_i) > c)$ with

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \text{ and } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

[4] See our companion paper, Hsu and Lieli (2015) for a precise regularity conditions and a further discussion of this issue.

Hence, the rule $1(\hat{g}(X_i) > c)$ is a random "mixture" between Rule$(+X)$ and Rule$(-X)$.[5] Let $e\text{AUC}_M$ denote the area under the associated ROC curve (the subscript M stands for mixture or model). Clearly,

$$e\text{AUC}_M \;=\; (e\text{AUC}_X) \cdot 1(\hat{\beta}_1 > 0) + (e\text{AUC}_{-X}) \cdot 1(\hat{\beta}_1 < 0). \tag{7}$$

While $e\text{AUC}_X$ and $e\text{AUC}_{-X}$ are asymptotically normal, this is not generally true for $e\text{AUC}_M$. The reason is that the sign of $\hat{\beta}_1$ is correlated with the in-sample classification performance of Rule$(+X)$ versus Rule$(-X)$. That is, $\hat{\beta}_1$ is likely to be positive when $e\text{AUC}_X > 1/2$ ($\Rightarrow e\text{AUC}_{-X} < 1/2$) and negative in the opposite case. This implies that $e\text{AUC}_M$ is more likely to be over $1/2$ than below $1/2$, i.e., the distribution of $e\text{AUC}_M$ is not symmetric around $1/2$.

In particular, when X and Y are independent, Rule$(+X)$ and Rule$(-X)$ are equally useless in the population. Nevertheless, in finite samples one of the rules will still slightly outperform the other just by random variation. While the value of $\hat{\beta}_1$ is likely to be close to zero, it will not be exactly zero, meaning that the preceding arguments apply even for large n. This suggests that $\sqrt{n}(e\text{AUC}_M - 1/2)$ cannot generally have a normal limit distribution with mean zero.

We formalize the intuition outlined above in three special cases: when X is binary, when X is uniform over [0,1], and when X is a vector and AUC is estimated parametrically. The following elementary lemma helps establish a connection between the sign of $\hat{\beta}_1$ and the event $e\text{AUC}_X > 1/2$.

Lemma 1 *Let $\hat{\beta}_0 \in \mathbb{R}$ and $\hat{\beta}_1 \in \mathbb{R}^k$ denote the estimated coefficients from a linear regression of $Y \in \{0,1\}$ on a constant and $X \in \mathbb{R}^k$. Then:*

(i) $\hat{\beta}_1 = \hat{p}(1-\hat{p})\hat{M}^{-1}(\bar{X}_1 - \bar{X}_0)$, where $\hat{M} = n^{-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'$, $\hat{p}$ is the sample mean of Y, and $\bar{X}_j$, $j = 0, 1$, is the sample mean of X in the $Y = j$ subsample.

(ii) $(\bar{X}_1 - \bar{X}_0)'\hat{\beta}_1 > 0$; in particular, if X is a scalar, then $sign(\hat{\beta}_1) = sign(\bar{X}_1 - \bar{X}_0)$.

[5]The probability of the event $\hat{\beta}_1 = 0$ is zero or asymptotically zero in all but some very non-generic situations. By the law of the iterated logarithm, this is true even when $\hat{\beta}_1$ converges to zero in probability. We will therefore ignore this event.

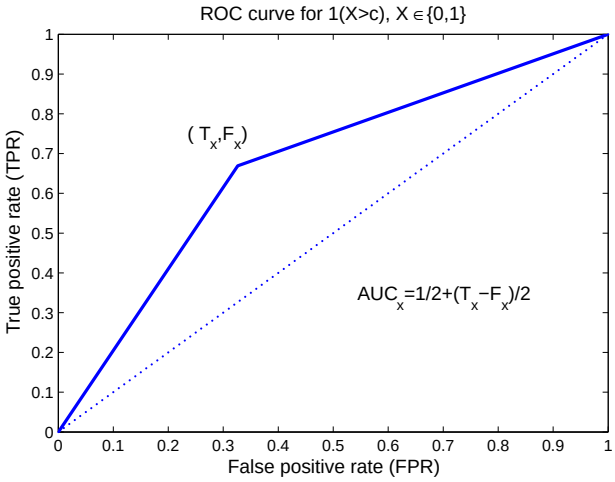

Figure 2: ROC curve with binary predictor

Part (i) of Lemma 1 follows from the general formula for the OLS estimator and some algebra. Part (ii) is a consequence of $\hat{M}$ being positive definite.

Example 3 Suppose that $X \in \{0,1\}$. For Rule(X), the pair $(\widehat{FPR}(c), \widehat{TPR}(c))$ takes on three possible values: $(1,1)$, $(0,0)$ and $(\hat{F}_X, \hat{T}_X)$, where

$$\widehat{F}_X = \frac{1}{n_0}\sum_{i=1}^{n} X_i(1-Y_i) \quad \text{and} \quad \widehat{T}_X = \frac{1}{n_1}\sum_{i=1}^{n} X_i Y_i.$$

One obtains the empirical ROC curve by connecting these points by a straight line; see Figure 2 for illustration. It is easy to verify that the area under the ROC curve is given by $\text{AUC}_X = 1/2 + (T_X - F_X)/2$. As $\hat{T}_X = \bar{X}_1$ and $\hat{F}_X = \bar{X}_0$, Lemma 1 part (ii) implies $sign(\hat{\beta}_1) = sign(e\text{AUC}_X - 1/2)$. In other words, the events $\hat{\beta}_1 > 0$ and $e\text{AUC}_X > 1/2$ are perfectly correlated. It follows that $\sqrt{n}(e\text{AUC}_M - 1/2)$ is *always* strictly greater than zero and its limit distribution is given by the absolute value of a normal random variable. To see this formally, rewrite equation (7) as

$$\begin{aligned} e\text{AUC}_M &= (e\text{AUC}_X)\cdot 1(\hat{\beta}_1 > 0) + (1 - e\text{AUC}_X)\cdot 1(\hat{\beta} < 0) \\ &= 1/2 + (e\text{AUC}_X - 1/2)[1(\hat{\beta}_1 > 0) - 1(\hat{\beta}_1 < 0)] \\ &= 1/2 + (e\text{AUC}_X - 1/2)[1(e\text{AUC}_X > 1/2) - 1(e\text{AUC}_X < 1/2)] \\ &= 1/2 + |e\text{AUC}_X - 1/2|. \end{aligned}$$

Therefore, by (5),

$$\sqrt{n}(e\text{AUC}_M - 1/2) \overset{a}{\sim} \left| N\left(0, \frac{n\hat{p}_x(1-\hat{p}_x)}{4n_0n_1}\right)\right|,$$

as claimed. Thus, a one-sided t-test of AUC=1/2 vs. AUC>1/2 based on (3) and (5) is twice as likely to reject the null as the chosen nominal size. ∎

Example 4 If X is a continuous random variable, $e\text{AUC}_X$ coincides with the Mann-Whitney U-statistic denoted in (2) as $\hat{U}$ (set $g(x) = x$). The general influence function representation of $\hat{U}$ is developed, for example, in Lehmann (1999, Ch. 6) and is restated in Appendix A for convenience. Under the additional assumption that X is uniform on $[0, 1]$ and X is independent of Y, this representation implies

$$\sqrt{n}(e\text{AUC}_X - 1/2) = \sqrt{n}(\bar{X}_1 - \bar{X}_0) + o_p(1);$$

set $F(x) = x$ and $g(x) = x$ in equation (18) in Appendix A. Then, by Lemma 1 part (ii), the sign of the l.h.s. coincides with the sign of $\hat{\beta}_1$ with probability approaching one, i.e., the events $\hat{\beta}_1 > 0$ and $e\text{AUC}_X > 1/2$ are again perfectly correlated, albeit asymptotically. By the same argument as in Example 3, the large sample distribution of $\sqrt{n}(e\text{AUC}_M - 1/2)$ is then the absolute value of a normal with mean zero and variance given by (4). Again, this means that a one-sided t-test of AUC=1/2 has actual size twice the chosen nominal size. ∎

Examples 3 and 4 are special in that the correlation between the events $\hat{\beta}_1 > 0$ and $e\text{AUC}_X > 1/2$ is (near) perfect under the null of independence. Simulations confirm that this is not generally true for other X-distributions, resulting in a positive probability that $e\text{AUC}_M < 1/2$. Thus, the limit distribution of $\sqrt{n}(e\text{AUC}_M - 1/2)$ is not necessarily the absolute value of a mean zero normal. Nevertheless, even an imperfect correlation between the two events in question is sufficient to ruin the standard limit results stated in Section 2.3 and lead to an overrejection of the null of independence.

Example 5 In contrast to the previous examples, let X be a k-dimensional vector with $X \mid Y = j \sim N(\mu_j, M)$, $j = 0, 1$. Since our focus is on testing the independence of X and Y, the assumption of a constant variance-covariance matrix M is not restrictive. By

the properties of the multivariate normal, $X'\beta_1 \mid Y = j \sim N(\mu_j'\beta_1, \beta_1'M\beta_1)$ for any given $\beta_1 \in \mathbb{R}^k$, $\beta_1 \neq 0$. Then, by formula (1), the population AUC of the decision rule $1(X'\beta_1 > c)$ is given by

$$\text{AUC} = \Phi\left(\frac{(\mu_1 - \mu_0)'\beta}{\sqrt{2\beta_1'M\beta_1}}\right).$$

Now suppose that we use the sample decision rules $1(X_i'\hat{\beta}_1 > c)$, where $\hat{\beta}_1$ is the vector of slope coefficients from an OLS regression of Y on X and a constant. Instead of computing the area under the empirical ROC curve, AUC can be estimated parametrically by

$$p\text{AUC} = \Phi\left(\frac{(\bar{X}_1 - \bar{X}_0)'\hat{\beta}_1}{\sqrt{2\hat{\beta}_1'\hat{M}\hat{\beta}_1}}\right), \tag{8}$$

where $\hat{M}$ and $\bar{X}_j$ are as in Lemma 1. It follows immediately from part (ii) of Lemma 1 that $p\text{AUC} > 1/2$. This implies that if X and Y are independent, result (6) cannot hold. ∎

3.2 Monte Carlo evidence

To further illustrate the severity as well as the generality of the overrejection problem, we also present a small Monte Carlo exercise. In all data generating processes considered here X and Y are independent. The key parameter is the dimension of X; we present results for $\dim(X) = 1, 2, 3, 10$. We specify a number of different distributions for X, including some cases where the components of X are correlated. For each specification of X, $p = P(Y = 1)$ is fixed at two different levels, 0.5 and 0.85.

We draw 10,000 random samples of size $n = 100$, $n = 500$ and 5000 from the distribution of (Y, X). For each sample, we estimate the linear regression of Y on X and a constant and construct the empirical ROC curve based on the fitted values. First we compute the area under this curve (eAUC) then we use the sample analog of the parametric formula (1) to approximate the same area (pAUC). We then test the hypothesis $H_0 : AUC = 1/2$ against $H_1 : AUC > 1/2$ at the $\alpha = 5\%$ and 10% nominal significance levels using the traditional normal null distributions presented in Section 2.3. Actual rejection rates over the 10,000 Monte Carlo repetitions are presented in Table 1 ($\alpha = 5\%$) and Table 2 ($\alpha = 10\%$), in the columns labeled 'Trad' as in 'traditional inference'.

The two most apparent features of the results are that (i) the overrejection problem is severe and (ii) the degree of size distortion depends mostly on the dimension of X. Intuitively speaking, if the dimension of X is higher, a first stage regression is more likely to create enough spurious matches between Y_i and $1(\hat{g}(X_i) > c)$ to boost the in-sample AUC beyond the usual normal critical values. For example, in the scalar case actual size is twice the nominal size, while for $\dim(X) = 3$, the actual size of the test is around 30-45% when $\alpha = 5\%$, and 50-67% when $\alpha = 10\%$. For $\dim(X) = 10$, rejection is practically certain.

There are also some more subtle patterns to the results involving eAUC. When X is uniform[0,1], we know that the asymptotic null distribution of $e\text{AUC}_M - 1/2$ is the absolute value of a mean zero normal, so the result that actual size is twice the nominal size is well understood. However, when, say, X is χ^2_1, there is roughly a 20% chance that $e\text{AUC}_M < 1/2$ even in large samples, yet the same result persists. For dim(X)>1, rejection rates are somewhat smaller for the heavily right-skewed distributions with unbounded support (chi-squared and lognormal) and larger for the distributions with bounded support (uniform and beta). Rejection rates for iid normal X are close to the latter. These differences might also have to do with the fact that for bounded X-distributions and the normal the probability of the event $e\text{AUC}_M < 1/2$ is small, while it is non-negligible for the outlier-prone distributions.

Rejection rates based on pAUC are virtually the same across all distributions in samples large enough ($n = 5000$). This is not surprising in light of the fact that testing $AUC = 1/2$ based on pAUC is equivalent to testing $E(X|Y = 1) = E(X|Y = 0)$ using differences in sample means. By the central limit theorem this difference will have the same limiting distribution regardless of the actual distribution of X.

4 Bootstrap failure

We will draw on Example 3, where X is binary, along with Monte Carlo simulations to illustrate the failure of the bootstrap for eAUC under first stage estimation and the independence of X and Y. The intuition is as follows. It is possible for the fixed decision rules Rule(X) or Rule($-X$) to have an eAUC less than 1/2 under the null; in fact, one of these values is necessarily less than 1/2. As shown Example 3, this is no longer so for the OLS mixture of

these two rules—first stage estimation essentially reduces the parameter space for AUC to the interval $[1/2, 1]$. Thus, under the null the true value of AUC is on the boundary of the relevant parameter space. This is a situation in which the bootstrap is known to be prone to failure.

4.1 Analytical example

Consider the setting of Example 3. Let the random variables $\{(Y_i, X_i)\}_{i=1}^n$ represent the original data, drawn from any probability distribution P on $\{0,1\}^2$. (Independence is not imposed at this point.) Let P_n^* denote the empirical distribution of the sample, i.e., P_n^* is a probability measure defined on $\{0,1\}^2$ that puts mass $1/n$ on each of the n points $\{(Y_i, X_i)\}_{i=1}^n$. The dependence of P_n^* on the original data means that it is a random probability measure. Let $\{(Y_i^*, X_i^*)\}_{i=1}^n$ denote a random sample drawn from P_n^*. As usual, statistics computed from the bootstrap sample will also be denoted by a star superscript.

We want to show that the bootstrap distribution

$$P_n^*[\sqrt{n}(e\text{AUC}_M^* - e\text{AUC}_M) \leq z] \tag{9}$$

does not consistently approximate the distribution

$$P[\sqrt{n}(e\text{AUC}_M - 1/2) \leq z] \tag{10}$$

under the null hypothesis that X and Y are independent. The argument is based on the observation that under the null $e\text{AUC}_M$ approaches $1/2$ from above so that (10) is exactly zero for $z = 0$ and any n. On the other hand, we will demonstrate that for $z = 0$ the sequence of random probabilities given in (9) does *not* converge to zero (in P-probability) as n gets large.

The following lemma will be useful.

Lemma 2 *For a random sample $\{(Y_i, X_i)\}_{i=1}^n$ drawn from P, $\sqrt{n}(\hat{T} - T)$ and $\sqrt{n}(\hat{F} - F)$ are independent, asymptotically normal random variables with mean zero and variance*

$$V_T = T_X(1 - T_X)/p \quad \textit{and} \quad V_F = F_X(1 - F_X)/(1 - p),$$

respectively.

While Lemma 2 is fairly elementary, a proof is provided in Appendix A for convenience.

First, for a given data set, consider the probability of drawing a bootstrap sample such that $e\text{AUC}_M^* < e\text{AUC}_M$. For concreteness, suppose that $e\text{AUC}_X = 1/2 + (\hat{T}_X - \hat{F}_X)/2 > 1/2$ in the original data so that $e\text{AUC}_M = e\text{AUC}_X$. In this case a sufficient (but not necessary) condition for the event of interest is

$$0 < \hat{T}_X^* - \hat{F}_X^* < \hat{T}_X - \hat{F}_X \;\Rightarrow\; e\text{AUC}_M^* = e\text{AUC}_X^* = 1/2 + (\hat{T}_X^* - \hat{F}_X^*)/2 < e\text{AUC}_X = e\text{AUC}_M.$$

Therefore,

$$\begin{aligned} & P_n^*\left(e\text{AUC}_M^* < e\text{AUC}_M\right) \\ & > P_n^*\left(0 < \hat{T}_X^* - \hat{F}_X^* < \hat{T}_X - \hat{F}_X\right) \\ & = P_n^*\left(-\sqrt{n}(\hat{T}_X - \hat{F}_X) < \sqrt{n}[\hat{T}_X^* - \hat{F}_X^* - (\hat{T}_X - \hat{F}_X)] < 0\right) \end{aligned} \tag{11}$$

Conditional on the original data, P_n^* is just a fixed probability measure on $\{0,1\}^2$, so one can apply Lemma 2 to the bootstrapped statistics $\sqrt{n}(\hat{T}_X^* - \hat{T}_X)$ and $\sqrt{n}(\hat{F}_X^* - \hat{F}_X)$. It follows that for large n, the bootstrap distribution of the random variable

$$\sqrt{n}[\hat{T}_X^* - \hat{F}_X^* - (\hat{T}_X - \hat{F}_X)]$$

is approximately normal with mean zero and variance $\hat{V}_T + \hat{V}_F$, where $\hat{V}_T = \hat{T}_X(1 - \hat{T}_X)/\hat{p}$ and $\hat{V}_F = \hat{F}_X(1 - \hat{F}_X)/(1 - \hat{p})$. Therefore, the probability under (11) is asymptotically equal to

$$\frac{1}{2} - \Phi\left(-\frac{\sqrt{n}(\hat{T}_X - \hat{F}_X)}{\sqrt{\hat{V}_T + \hat{V}_F}}\right). \tag{12}$$

Similarly, if $e\text{AUC}_X < 1/2$ in the original data, then (11) is asymptotically equal to

$$\frac{1}{2} - \Phi\left(-\frac{\sqrt{n}(\hat{F}_X - \hat{T}_X)}{\sqrt{\hat{V}_T + \hat{V}_F}}\right). \tag{13}$$

Next, we will again regard the original data as a random draw from P. This means that the probabilities under (12) and (13) are random quantities rather than fixed numbers. In particular, if X and Y are independent under P, then $T_X = F_X = P(X = 1)$, so

$$Z_n \equiv \frac{\sqrt{n}(\hat{T}_X - \hat{F}_X)}{\sqrt{\hat{V}_T + \hat{V}_F}}$$

has a standard normal limiting distribution by Lemma 2 and the fact that $\hat{V}_T$ and $\hat{V}_F$ are consistent for V_T and V_F, respectively.

Combining (9), (11), (12) and (13) for $z = 0$ yields

$$\begin{aligned} & P_n^*[\sqrt{n}(e\text{AUC}_M^* - e\text{AUC}_M) \leq 0] \\ & = P_n^*\left(e\text{AUC}_M^* < e\text{AUC}_M\right) \\ & > 1/2 - \Phi(-|Z_n|) \end{aligned} \tag{14}$$

for large n. As Z_n is asymptotically standard normal under P, (14) clearly does not go to zero in probability as $n \to \infty$.

4.2 Monte Carlo evidence

The example presented in the previous section shows that the event $\sqrt{n}(e\text{AUC}_M^* - e\text{AUC}_M) < 0$ has potentially large positive probability even in large samples. This suggests that the test

> "reject H_0 : AUC=1/2 if $\sqrt{n}(e\text{AUC}_M - 1/2)$ is greater than, say, the 95th percentile of the distribution of $\sqrt{n}(e\text{AUC}_M^* - e\text{AUC}_M)$"

is likely to have asymptotic size larger than 5%, so the bootstrap does not solve the overrejection problem.

We investigate this prediction for a small subset of the Monte Carlo specifications described in Section 3.2 with the difference that we use the bootstrap, i.e., the decision rule stated above, to conduct inference. (We draw 300 bootstrap samples in each Monte Carlo cycle.) Results are reported in Table 3. Comparing this table with the appropriate sections of Tables 1 and 2, we see that for $\dim(X) > 1$ the bootstrap reduces the overrejection problem to some degree in comparison with the analytical tests, but the remaining size distortion is still very severe. There is no improvement in the scalar case. Use of the bootstrap to conduct inference about the hypothesis AUC=1/2 is therefore not recommended in the presence of first stage estimation.

5 Tests of AUC=1/2 with first stage estimation

We propose two tests of the hypothesis AUC=1/2 that are asymptotically valid for decision rules with estimated parameters. One of the tests is based on pAUC and the other is on eAUC.

5.1 Theoretical results

We summarize our claims in the following proposition.

Proposition 1 *Let $X \in \mathbb{R}^k$ with at least one continuous component. Let $\hat{g}(X) = \hat{\beta}_0 + X'\hat{\beta}_1$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated regression coefficients from an OLS regression of Y on a constant and X. Consider the empirical ROC curve constructed from the family of cutoff rules $1(\hat{g}(X_i) > c)$.*

(a) Let eAUC denote the area under the ROC curve. Under the null hypothesis that X and Y are independent,

$$12\hat{p}(1-\hat{p})n(e\text{AUC} - 1/2)^2 \overset{a}{\sim} \chi^2_k.$$

(b) Let pAUC denote the parametric approximation to the area under the ROC curve given by equation (8). Under the null hypothesis that X and Y are independent,

$$4\pi\hat{p}(1-\hat{p})n(p\text{AUC} - 1/2)^2 \longrightarrow_d \chi^2_k. \tag{15}$$

Comments

1. Proposition 1 remains unchanged even if the first stage model is a logit-type regression, i.e., $\hat{g}(x) = G(\hat{\beta}_0 + \hat{\beta}_1' x)$, where $G(\cdot)$ is a cdf (e.g., logistic) and $\hat{\beta}_0$ and $\hat{\beta}_1$ are maximum likelihood estimates. The reason is that under the null of independence, the estimated coefficients are first-order asymptotically equivalent to the OLS regression coefficients up to a scalar multiple.

2. While the formula for pAUC is derived under normality, the test stated in part (b) can be used generally even if the data are non-normal. In fact, as shown below in

Comment 7, (15) is equivalent to simply testing $E(X \mid Y = 0) = E(X \mid Y = 1)$ or, by Lemma 1, the hypothesis that $\beta_1 = 0$. The only difference is that the statistic used to construct the test has the interpretation of being an estimator of AUC. The parametric estimator is consistent under the null, but may be inconsistent under the alternative (AUC>1/2) if the normality assumption does not actually hold. Nevertheless, the test still has power against alternatives with $E(X \mid Y = 0) \neq E(X \mid Y = 1)$.

3. Proposition 1 part (b) is related to results by Demler, Pencina and D'Agostino (2011), who are interested in testing whether additional covariates used in a linear discriminant analysis cause a statistically significant increase in the sample AUC (see the introduction for a more detailed discussion).

4. The proposition quantifies the impact of overfitting due to pre-estimation. Suppose that $n = 100$, $k = 3$, $\hat{p} = 0.5$, and that the researcher uses a regression model to construct the predictive index for Y. The 95th percentile of a χ^2_3 distribution is 7.815. Using Proposition 1 part (a), one needs an eAUC greater than $0.5 + \sqrt{7.815/300} = 0.661$ to reject the null of independence at the 5% level. In contrast, if one naively uses a one-sided t-test based on (3) and (4), one will reject the null already when the the empirical AUC exceeds $0.5 + 1.645/\sqrt{300} = 0.595$.

5. In this version of the paper we do not yet have a complete proof of Proposition 1 part (a). In fact, the simulation results presented in the next section suggest that the claim is not precisely true without some restrictions on the distribution of X (e.g., bounded support). Nevertheless, the same set of results also show that if there are size distortions, they only make the test more conservative, and are not very large in magnitude. Therefore, even if Proposition 1 part (a) is not literally true in general, it can still serve as a useful rule of thumb in practice. (Hence the subtle difference in notation between the two parts.) Nevertheless, as shown below in Comment 6, part (a) holds precisely for continuous scalar X. This case is mostly of academic interest as for scalar X there is no real point in estimating a first stage regression before constructing the ROC curve.

6. *Proof of Proposition 1 part* (a) *for dim*(X)*=1.* Let X be a continuously distributed scalar variable. In Example 3 we have shown that

$$e\text{AUC}_M = 1/2 + (e\text{AUC}_X - 1/2)[1(\hat{\beta}_1 > 0) - 1(\hat{\beta}_1 < 0)].$$

The derivation of this equation only depended on the fact that at X was a scalar and did not make any use of the binary assumption made in Example 3. As Rule(X) has no estimated parameters, equations (3) and (4) apply to $e\text{AUC}_X$ under the null of independence, and they yield

$$\begin{aligned} \frac{n(e\text{AUC}_M - 1/2)^2}{n^2/12n_0n_1} &= 12\hat{p}(1-\hat{p})n(e\text{AUC}_X - 1/2)^2[1(\hat{\beta}_1 > 0) - 1(\hat{\beta}_1 < 0)]^2 \\ &= 12\hat{p}(1-\hat{p})n(e\text{AUC}_X - 1/2)^2 \xrightarrow{H_0}_d [N(0,1)]^2 = \chi^2_1, \end{aligned}$$

where the second equality follows because the difference of the two indicator functions is either 1 or -1. The proof of the scalar case is complete. ∎

7. *Proof of Proposition 1 part* (b). Substituting the expression for $\hat{\beta}_1$ given in part (i) of Lemma 1 into equation (8) gives

$$p\text{AUC} = \Phi\left(\sqrt{\frac{1}{2}(\bar{X}_1 - \bar{X}_0)'\hat{M}^{-1}(\bar{X}_1 - \bar{X}_0)'}\right).$$

A Taylor expansion of $\Phi(x)$ around $x = 0$ shows that under the null,

$$\sqrt{n}(p\text{AUC} - 1/2) = \phi(0)\sqrt{\frac{n}{2}(\bar{X}_1 - \bar{X}_0)'\hat{M}^{-1}(\bar{X}_1 - \bar{X}_0)'} + o_p(1). \tag{16}$$

By standard asymptotic theory,

$$\sqrt{n}(\bar{X}_1 - \bar{X}_0) \to_d N\left(0, \frac{1}{p(1-p)}M\right)$$

under the null, where M is the population variance-covariance matrix of X. If M is nonsingular, it follows from standard results on normal quadratic forms that

$$\hat{p}(1-\hat{p})n(\bar{X}_1 - \bar{X}_0)'\hat{M}^{-1}(\bar{X}_1 - \bar{X}_0) \xrightarrow{H_0}_d \chi^2_k. \tag{17}$$

Squaring equation (16) and combining it with (17) completes the proof of Proposition 1 part (b). ∎

8. At first glance one might think that the nonparametric test is potentially consistent for a broader range of alternatives (violations of independence) than the parametric one. This is not so because of the first stage estimation step. The nonparametric test is consistent only against alternatives with $\beta_1 \neq 0$. But by the population version of Lemma 1 part (i), $\beta_1 \neq 0$ is equivalent to $E(X \mid Y = 0) \neq E(X \mid Y = 1)$, as the variance-covariance matrix of X is nonsingular.

9. In the multivariate case ($\dim(X) > 1$) the asymptotic null distribution of the nonparametric test cannot easily be derived from the influence function representation given in Appendix A. If one sets $g(X) = X'\hat{\beta}_1$, the cdf F also becomes dependent on $\hat{\beta}_1$, and for $\hat{\beta}_1 \to_p 0$, it degenerates toward the cdf of unit mass on zero. This implies that in a Taylor expansion of F around zero, there will be terms of the form $0 \times \infty$. If one uses a normalization to stabilize the variance of $X'\hat{\beta}_1$, higher order terms will not generally vanish.

5.2 Monte Carlo evidence

Using both parts of Proposition 1, we conduct inference for the full set of data generating processes introduced in Section 3.2. In all cases X and Y are independent. Actual rejection rates of the hypothesis AUC=1/2 are reported in Table 1 ($\alpha = 5\%$) and Table 2 ($\alpha = 10\%$) in the columns labeled 'Corr(ected)'. In discussing the results, we will also refer to the tests based on eAUC and pAUC as tests (a) and (b), respectively.

The contrast between the corrected tests and the traditional ones is rather striking. In short, tests (a) and (b) both completely eliminate the overrejection problem. The asymptotic size of the test based on pAUC is spot-on in all cases (see $n = 5000$), which is not exactly surprising given that it is basically a version of the well-established Wald-test. The test based on eAUC also has accurate-seeming asymptotic size for covariate distributions with bounded support and the normal, but there is evidence of underrejection for the chi-squared and lognormal distributions. Nevertheless, in most of these cases, the test is just a couple of percentage points more conservative in large samples than its nominal significance level. Increasing the dimension of X makes the degree of size distortion larger, but even for

$\dim(X) = 10$, X iid lognormal, actual asymptotic size is about 40% of the nominal size. It is likely that the observed size problems have to do with the outlier-proneness of the chi-squared and lognormal distributions rather than skewness per se (the beta(2,1) distribution is also skewed but it has bounded support).

Both tests show good size control even in smaller samples ($n = 100, 500$); overrejection is never a problem and the degree of underrejection is still generally tolerable when it occurs. A notable difference is that now test (b) is also likely to underreject for X chi-squared and lognormal, especially when the distribution of Y is unbalanced ($p = 0.85$). In fact, in the latter case test (a) may have more accurate size when the sample size is small ($n = 100$).

An interesting aspect of test (a) is that in those cases when it underrejects, the degree of size distortion seems rather insensitive to the sample size; that is, the actual size for $n = 100$ is already similar to the asymptotic size ($n = 5000$). This suggests that underrejection is tied to very specific features to the covariate distributions, which makes the general characterization of the null distribution of eAUC potentially very challenging. In contrast, when test (b) underrejects in small samples, the central limit theorem eventually eliminates the size distortion, and the asymptotic size is very accurate in all cases.

6 Conclusion

We are concerned with testing the null hypothesis that the area under a sample ROC curve is 1/2, which means that the underlying predictors do no better than chance in classifying the outcome. We have used some analytical examples and Monte Carlo simulations to demonstrate that if a sample ROC curve is constructed from a model estimated on the same data set, then, under the null, (i) the estimates of AUC (parametric or nonparametric) do not follow the normal limit distribution derived from conventional asymptotic theory; (ii) even bootstrap-based inference produces misleading results. The Monte Carlo evidence demonstrates that the upward size distortion is potentially severe and depends mostly on the number of estimated parameters (at least for models based on a linear index).

We propose two asymptotic tests as a solution to the problem at hand. The test based on a parametric estimate of AUC has a chi-squared null distribution, and is equivalent to a Wald

test of the joint significance of the coefficients in the first stage model. The nonparametric version of the test, based on the area under the empirical ROC curve, uses the same chi-squared critical values. Even though the theoretical justification behind the latter test is incomplete at the moment, the Monte Carlo evidence shows very good size control that is occasionally on the conservative side. A better understanding of the asymptotic null distribution of the nonparametric test is obviously a priority for future research. While we argued that the two tests should be consistent for essentially the same set of alternatives, a more careful comparison of the finite sample or local power properties would also be of interest, both from a theoretical and a practical standpoint.

A quick way of avoiding the overfitting problem altogether is not to use the same data set for estimation (training) and evaluation (validation). While this is a legitimate strategy in practice, explicit characterization of the first stage estimation effect is still valuable. One reason for this is that splitting the existing data set involves loss of power, which might be substantial in applications such as forecasting recessions, where the amount of available data is necessarily limited. Furthermore, results may depend on the exact sample split used unless both samples are sufficiently large or a full-fledged cross validation exercise is conducted.

References

[1] Anjali D.N. and P. Bossaerts (2014): "Risk and Reward Preferences under Time Pressure". *Review of Finance* 18: 999-1022.

[2] Bamber, D. (1975): "The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph". *Journal of Mathematical Psychology* 12: 387-415.

[3] Carole, C-F. and T.J. Putnins (2014): "Stock Price Manipulation: Prevalence and Determinants". *Review of Finance* 18: 23-66.

[4] DeLong, E.R., D.M. DeLong and D.L. Clarke-Pearson (1988): "Comparing areas under two or more correlated receiver operating characteristic curves: a nonparametric approach". *Biometrics* 44: 837-845.

[5] Demler, O.V., M.J. Pencina and R.B. D'Agostino, Sr. (2011): "Equivalance of improvement in area under ROC curve and linear discriminant analysis coefficient under assumption of normality". *Statistics in Medicine* 30: 1410-1418.

[6] Demler, O.V., M.J. Pencina and R.B. D'Agostino, Sr. (2012): "Misuse of DeLong test to compare AUCs for nested models". *Statistics in Medicine* 31: 2577-2587.

[7] Egan, J.P. (1975): *Signal Detection Theory and ROC Analysis*. Academic Press: New York.

[8] Elliott, G. and R.P. Lieli (2013): "Predicting Binary Outcomes". *Journal of Econometrics* 174: 15-26.

[9] Fawcett, T. (2004): "ROC Graphs: Notes and Practical Considerations for Researchers". Technical report, HP Laboratories.

[10] Green, D.M. and J.A. Swets (1966): *Signal Detection Theory and Psychophysics*. Wiley: New York.

[11] Hand, D.J. (2009): "Measuring classifier performance: a coherent alternative to the area under the ROC curve." *Machine Learning* 77: 103-123.

[12] Hanley, J.A. and B.J. McNeil (1982): "The meaning and use of the area under a receiver operating characteristic (ROC) curve". *Radiology* 143: 29-36.

[13] Hsu, Y-C. and R.P. Lieli (2015): "Inference for ROC curves based on estimated predictive indices". Working paper.

[14] Jorda, O. and A.M. Taylor (2011): "Performance Evaluation of Zero Net-Investment Strategies". NBER Working Paper 17150.

[15] Jorda, O. and A.M. Taylor (2013): "The Time for Austerity: Estimating the Average Treatment Effect of Fiscal Policy". NBER Working Paper 19414.

[16] Lahiri, K. and J.G. Wang (2013): "Evaluating Probability Forecasts for GDP Declines Using Alternative Methodologies". *International Journal of Forecasting* 29: 175190.

[17] Lehmann, E.L. (1999): *Elements of Large Sample Theory.* Springer: New York.

[18] Pepe, M.S. (2003): *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford: Oxford University Press.

[19] Sreedhar, T.B. and A.K. Dittmar (2010): "Why Do Firms Use Private Equity to Opt Out of Public Markets?". *Rev. Financ. Stud.* 23: 1771-1818.

Appendix A: Proofs, derivations and some technical results

A.1 The asymptotic distribution of pAUC for fixed $g(\cdot)$ and under independence

For simplicity, set $\sigma_0^2 = \sigma_1^2 := \sigma^2 > 0$ in formula (1). As we are interested in the distribution of pAUC under the null that X and Y are independent, this simplification is not restrictive. Let $\hat{\mu}_j$ denote the sample mean of $g(X)$ in the $Y = j$ subsample and let $\hat{\sigma}^2$ be the pooled sample variance of $g(X)$. Put $p\text{AUC} = \Phi\big((\hat{\mu}_1 - \hat{\mu}_0)/\sqrt{2\hat{\sigma}^2}\big)$. Under the null,

$$\frac{\sqrt{n}(\hat{\mu}_1 - \hat{\mu}_0)}{\sqrt{2\hat{\sigma}^2}} = \frac{\sqrt{n}(\hat{\mu}_1 - \hat{\mu}_0)}{\sqrt{2\sigma^2}} + o_p(1) \overset{a}{\sim} N\left(0, \frac{n}{2n_0} + \frac{n}{2n_1}\right).$$

By the delta method,

$$\sqrt{n}(p\text{AUC} - \Phi(0)) = \sqrt{n}(p\text{AUC} - 1/2) \overset{a}{\sim} N\left(0, \frac{n[\Phi'(0)]^2}{2n_0} + \frac{n[\Phi'(0)]^2}{2n_1}\right),$$

which is the same as (6) since $n = n_1 + n_0$ and $\Phi'(0) = 1/\sqrt{2\pi}$.

A.2 The influence function representation of $\hat{U}$

Let eAUC be the area under the empirical ROC curve associated with the decision rule $1(g(X_i) > c)$. Suppose that the distributions $g(X)|Y = j$ are absolutely continuous with cdfs where $F_j(\cdot)$, $j = 0, 1$. In this case $e\text{AUC} = \hat{U}$. Let $\{Z_{0,i}\}_{i=1}^{n_0}$ and $\{Z_{1,j}\}_{j=1}^{n_1}$ denote the values of the predictive index $g(X)$ over the $Y = 0$ and $Y = 1$ subsamples, respectively. Applying formula (6.1.75) of Lehmann (1999) with $\phi(a, b) = 1(a < b)$ yields

$$\sqrt{n}(e\text{AUC} - \theta) \quad = \quad \sqrt{n}\left\{\frac{1}{n_0}\sum_{i=1}^{n_0}[1 - F_1(Z_{0,i}) - \theta] + \frac{1}{n_1}\sum_{j=1}^{n_1}[F_0(Z_{1,j}) - \theta]\right\} + o_p(1),$$

where θ is the area under the population ROC curve. If X and Y are independent, then $F_0 = F_1 := F$ and $\theta = 1/2$ yielding

$$\begin{aligned}\sqrt{n}(e\text{AUC} - 1/2) \quad &= \quad \sqrt{n}\left\{\frac{1}{n_1}\sum_{j=1}^{n_1}F(Z_{1,j}) - \frac{1}{n_0}\sum_{i=1}^{n_0}F(Z_{0,i})\right\} + o_p(1) \qquad (18)\\ &= \quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{Y_i}{p} - \frac{1 - Y_i}{1 - p}\right)(F(g(X_i)) - 1/2) + o_p(1).\end{aligned}$$

A.3 The proof of Lemma 2: the asymptotic properties of $\hat{T}_X$ and $\hat{F}_X$

That $\hat{T}_X$ and $\hat{F}_X$ are independent follows immediately from the fact that the former is the sample average of the X values in the $Y = 1$ subsample, while the latter is the sample average of the X values in the $Y = 0$ subsample. It is straightforward to verify that the estimators $\hat{T}_X$ and $\hat{F}_X$ are asymptotically linear with

influence function representations

$$\begin{aligned}
\sqrt{n}(\hat{T}_X - T_X) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{X_i Y_i}{p} - T_X - \frac{T_X}{p}(Y_i - p) \right] + o_p(1) \\
&\equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_T(X_i, Y_i) + o_p(1) \\
\sqrt{n}(\hat{F}_X - F_X) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{X_i(1 - Y_i)}{1 - p} - F_X + \frac{F_X}{1 - p}(Y_i - p) \right] + o_p(1) \\
&\equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_F(X_i, Y_i) + o_p(1)
\end{aligned}$$

where $p = P(Y = 1)$. As $\{\psi_T(X_i, Y_i)\}$ and $\{\psi_F(X_i, Y_i)\}$ are i.i.d. sequences with

$$\begin{aligned}
&E[\psi_T(X_i, Y_i)] = E[\psi_F(X_i, Y_i)] = 0 \\
&V_T \equiv E\{[\psi_T(X_i, Y_i)]^2\} = T_X(1 - T_X)/p \\
&V_F \equiv E\{[\psi_F(X_i, Y_i)]^2\} = F_X(1 - F_X)/(1 - p),
\end{aligned}$$

the result follows immediately from the central limit theorem.

Appendix B: Tables

Table 1: Rejection rates of $H_0: AUC = 1/2$ under independence of X and Y at $\alpha = 5\%$

		$n = 100$				$n = 500$				$n = 5000$			
		Trad.		Corr.		Trad.		Corr.		Trad.		Corr.	
dim(X)=1	p	eAUC	pAUC	eAUC	pAUC	eAUC	pAUC	eAUC	pAUC	eAUC	pAUC	eAUC	pAUC
$X \sim N(0,1)$	0.5	0.102	0.101	0.050	0.050	0.097	0.098	0.051	0.050	0.097	0.100	0.052	0.052
	0.85	0.102	0.100	0.049	0.049	0.099	0.099	0.051	0.049	0.099	0.102	0.049	0.048
$X \sim U[0,1]$	0.5	0.102	0.100	0.051	0.050	0.102	0.101	0.050	0.051	0.105	0.104	0.050	0.050
	0.85	0.102	0.103	0.051	0.053	0.099	0.099	0.050	0.050	0.096	0.096	0.047	0.048
$X \sim \beta(2,1)$	0.5	0.097	0.095	0.046	0.045	0.104	0.102	0.054	0.052	0.099	0.102	0.048	0.047
	0.85	0.103	0.102	0.054	0.054	0.102	0.102	0.053	0.052	0.097	0.097	0.047	0.048
$X \sim \chi^2_1$	0.5	0.101	0.099	0.051	0.047	0.100	0.099	0.052	0.050	0.096	0.096	0.048	0.048
	0.85	0.102	0.102	0.052	0.044	0.102	0.095	0.052	0.050	0.102	0.101	0.052	0.054
$X \sim e^{N(0,1)}$	0.5	0.099	0.095	0.052	0.041	0.096	0.099	0.047	0.047	0.096	0.100	0.050	0.050
	0.85	0.104	0.085	0.053	0.033	0.098	0.097	0.050	0.045	0.100	0.098	0.050	0.051
dim(X)=2													
$X_i \sim$iid $N(0,1)$	0.5	0.257	0.260	0.049	0.050	0.254	0.260	0.049	0.049	0.249	0.253	0.046	0.048
	0.85	0.259	0.261	0.047	0.048	0.255	0.257	0.049	0.050	0.254	0.255	0.049	0.048
$X_i \sim$iid $U[0,1]$	0.5	0.263	0.262	0.050	0.051	0.258	0.260	0.051	0.051	0.258	0.258	0.053	0.054
	0.85	0.262	0.262	0.048	0.052	0.257	0.259	0.049	0.051	0.258	0.260	0.052	0.051
$X_i \sim$iid $\beta(2,1)$	0.5	0.258	0.261	0.051	0.051	0.251	0.256	0.048	0.049	0.260	0.262	0.050	0.048
	0.85	0.249	0.250	0.044	0.047	0.255	0.262	0.050	0.051	0.258	0.261	0.052	0.051
$X_i \sim$iid χ^2_1	0.5	0.220	0.263	0.044	0.047	0.221	0.260	0.041	0.051	0.207	0.257	0.040	0.052
	0.85	0.223	0.257	0.040	0.033	0.210	0.253	0.039	0.046	0.210	0.266	0.044	0.051
$X_i \sim$iid $e^{N(0,1)}$	0.5	0.204	0.265	0.036	0.038	0.200	0.269	0.036	0.048	0.188	0.255	0.036	0.051
	0.85	0.212	0.232	0.041	0.023	0.197	0.254	0.036	0.040	0.192	0.262	0.037	0.050
dim(X)=3													
$X_i \sim$iid $N(0,1)$	0.5	0.431	0.444	0.050	0.050	0.426	0.439	0.048	0.049	0.419	0.436	0.050	0.052
	0.85	0.432	0.439	0.045	0.046	0.425	0.440	0.048	0.049	0.418	0.436	0.046	0.050
$X_i \sim$iid $U[0,1]$	0.5	0.441	0.443	0.050	0.050	0.433	0.438	0.051	0.051	0.432	0.438	0.048	0.049
	0.85	0.443	0.444	0.047	0.050	0.434	0.439	0.049	0.050	0.426	0.433	0.051	0.052
$X_i \sim$iid $\beta(2,1)$	0.5	0.434	0.445	0.050	0.050	0.426	0.440	0.052	0.051	0.430	0.440	0.046	0.048
	0.85	0.438	0.447	0.045	0.049	0.432	0.443	0.051	0.053	0.425	0.437	0.044	0.048
$X_i \sim$iid χ^2_1	0.5	0.372	0.460	0.041	0.047	0.360	0.440	0.041	0.054	0.352	0.437	0.039	0.051
	0.85	0.378	0.437	0.032	0.026	0.354	0.437	0.038	0.047	0.350	0.443	0.037	0.049
$X_i \sim$iid $e^{N(0,1)}$	0.5	0.346	0.465	0.035	0.041	0.309	0.447	0.031	0.045	0.295	0.441	0.030	0.049
	0.85	0.341	0.405	0.030	0.017	0.319	0.443	0.032	0.034	0.299	0.435	0.031	0.046
dim(X)=3													
$X_i = \sum_{s=1}^{i} Z_s$	0.5	0.435	0.447	0.049	0.050	0.429	0.443	0.050	0.051	0.428	0.440	0.050	0.053
	0.85	0.432	0.443	0.046	0.048	0.426	0.440	0.047	0.049	0.432	0.446	0.047	0.049
$X_i = \sum_{s=1}^{i} U_s$	0.5	0.442	0.446	0.050	0.052	0.437	0.441	0.051	0.051	0.437	0.440	0.049	0.049
	0.85	0.447	0.449	0.046	0.051	0.434	0.439	0.049	0.052	0.428	0.431	0.049	0.049
$X_i = \sum_{s=1}^{i} \beta_s$	0.5	0.436	0.446	0.051	0.052	0.428	0.438	0.050	0.052	0.430	0.444	0.049	0.052
	0.85	0.444	0.450	0.044	0.050	0.425	0.439	0.046	0.049	0.424	0.435	0.047	0.049
$X_i = \sum_{s=1}^{i} K_s$	0.5	0.359	0.451	0.040	0.046	0.356	0.447	0.037	0.047	0.347	0.436	0.034	0.049
	0.85	0.367	0.429	0.036	0.025	0.353	0.440	0.036	0.044	0.356	0.444	0.037	0.054
$X_i = \sum_{s=1}^{i} L_s$	0.5	0.332	0.452	0.037	0.040	0.313	0.455	0.033	0.049	0.299	0.444	0.030	0.049
	0.85	0.345	0.404	0.028	0.015	0.304	0.425	0.032	0.033	0.295	0.432	0.029	0.049
dim(X)=10													
$X_i \sim$iid $N(0,1)$	0.5	0.983	0.990	0.047	0.052	0.981	0.987	0.045	0.053	0.977	0.987	0.048	0.054
	0.85	0.982	0.988	0.024	0.029	0.979	0.986	0.041	0.047	0.978	0.988	0.044	0.049
$X_i \sim$iid $U[0,1]$	0.5	0.983	0.989	0.050	0.058	0.981	0.986	0.049	0.053	0.982	0.987	0.044	0.048
	0.85	0.983	0.988	0.023	0.030	0.983	0.988	0.040	0.048	0.983	0.990	0.046	0.050
$X_i \sim$iid $e^{N(0,1)}$	0.5	0.962	0.993	0.033	0.041	0.925	0.991	0.024	0.047	0.899	0.988	0.020	0.048
	0.85	0.950	0.984	0.009	0.004	0.925	0.989	0.023	0.022	0.905	0.988	0.021	0.042

Note: eAUC is the area under the empirical ROC curve constructed from prediction rules of the form $\hat{Y} = 1(\hat{p} > c)$, where $\hat{p}$ is the fitted value from an OLS regression of Y on X and a constant. pAUC is the normal-parametric estimate of AUC based on the same index $\hat{p}$. The columns labeled 'Trad(itional)' contain empirical rejection rates from traditional tests using normal critical values, while the columns labeled 'Corr(ected)' are based on the proposed tests with chi-squared critical values. Z_s, U_s, β_s, K_s and L_s denote iid standard normal, uniform$[0,1]$, $\beta(2,1)$, χ^2_1 and lognormal$(0,1)$ random variables, respectively.

Table 2: Rejection rates of $H_0: AUC = 1/2$ under independence of X and Y at $\alpha = 10\%$

		$n = 100$				$n = 500$				$n = 5000$			
		Trad.		Corr.		Trad.		Corr.		Trad.		Corr.	
dim(X)=1	p	eAUC	pAUC	eAUC	pAUC	eAUC	pAUC	eAUC	pAUC	eAUC	pAUC	eAUC	pAUC
$X \sim N(0,1)$	0.5	0.202	0.201	0.101	0.098	0.199	0.197	0.102	0.098	0.200	0.202	0.097	0.100
	0.85	0.206	0.203	0.102	0.099	0.199	0.199	0.101	0.099	0.198	0.202	0.099	0.102
$X \sim U[0,1]$	0.5	0.205	0.202	0.104	0.102	0.201	0.200	0.100	0.101	0.200	0.201	0.105	0.104
	0.85	0.205	0.202	0.104	0.103	0.202	0.201	0.101	0.099	0.200	0.200	0.096	0.096
$X \sim \beta(2,1)$	0.5	0.202	0.195	0.097	0.095	0.195	0.198	0.100	0.100	0.205	0.205	0.099	0.102
	0.85	0.207	0.206	0.103	0.102	0.203	0.200	0.102	0.103	0.206	0.200	0.097	0.097
$X \sim \chi^2_1$	0.5	0.198	0.205	0.101	0.099	0.197	0.199	0.101	0.099	0.192	0.200	0.098	0.098
	0.85	0.203	0.209	0.103	0.102	0.199	0.198	0.103	0.095	0.198	0.204	0.103	0.101
$X \sim e^{N(0,1)}$	0.5	0.201	0.210	0.101	0.095	0.185	0.208	0.098	0.099	0.190	0.198	0.099	0.100
	0.85	0.201	0.199	0.106	0.085	0.190	0.211	0.100	0.097	0.192	0.200	0.103	0.098
dim(X)=2													
$X_i \sim$iid $N(0,1)$	0.5	0.435	0.442	0.100	0.100	0.429	0.438	0.098	0.101	0.426	0.435	0.096	0.097
	0.85	0.439	0.443	0.098	0.099	0.430	0.441	0.098	0.100	0.430	0.440	0.099	0.100
$X_i \sim$iid $U[0,1]$	0.5	0.444	0.441	0.102	0.102	0.441	0.443	0.102	0.100	0.431	0.436	0.103	0.102
	0.85	0.447	0.444	0.099	0.104	0.441	0.443	0.100	0.100	0.440	0.440	0.102	0.103
$X_i \sim$iid $\beta(2,1)$	0.5	0.440	0.446	0.103	0.103	0.434	0.443	0.096	0.097	0.435	0.441	0.099	0.100
	0.85	0.426	0.432	0.094	0.097	0.435	0.444	0.100	0.101	0.430	0.438	0.100	0.104
$X_i \sim$iidχ^2_1	0.5	0.373	0.455	0.086	0.100	0.367	0.449	0.088	0.101	0.362	0.440	0.080	0.100
	0.85	0.385	0.454	0.085	0.086	0.362	0.442	0.082	0.096	0.363	0.443	0.085	0.101
$X_i \sim$iid$e^{N(0,1)}$	0.5	0.348	0.460	0.077	0.091	0.345	0.463	0.074	0.102	0.325	0.442	0.072	0.099
	0.85	0.362	0.440	0.079	0.064	0.340	0.450	0.076	0.088	0.330	0.444	0.072	0.098
dim(X)=3													
$X_i \sim$iid $N(0,1)$	0.5	0.640	0.655	0.100	0.102	0.633	0.653	0.096	0.100	0.625	0.645	0.097	0.104
	0.85	0.639	0.654	0.095	0.097	0.632	0.650	0.094	0.099	0.626	0.647	0.099	0.102
$X_i \sim$iid $U[0,1]$	0.5	0.650	0.655	0.100	0.100	0.644	0.650	0.100	0.100	0.641	0.651	0.098	0.098
	0.85	0.653	0.655	0.097	0.100	0.646	0.652	0.097	0.101	0.643	0.651	0.098	0.102
$X_i \sim$iid $\beta(2,1)$	0.5	0.640	0.657	0.100	0.102	0.635	0.648	0.101	0.102	0.631	0.651	0.097	0.100
	0.85	0.649	0.661	0.096	0.102	0.637	0.650	0.101	0.104	0.632	0.648	0.095	0.097
$X_i \sim$iid χ^2_1	0.5	0.551	0.672	0.084	0.102	0.540	0.654	0.082	0.103	0.536	0.651	0.079	0.098
	0.85	0.573	0.662	0.074	0.067	0.540	0.646	0.078	0.097	0.536	0.651	0.076	0.101
$X_i \sim$iid $e^{N(0,1)}$	0.5	0.522	0.683	0.073	0.090	0.479	0.662	0.067	0.096	0.466	0.652	0.063	0.102
	0.85	0.531	0.655	0.066	0.051	0.497	0.666	0.070	0.080	0.464	0.645	0.060	0.093
dim(X)=3													
$X_i = \sum_{j=1}^{i} Z_j$	0.5	0.640	0.657	0.098	0.102	0.635	0.655	0.100	0.101	0.630	0.650	0.097	0.100
	0.85	0.638	0.652	0.096	0.099	0.634	0.653	0.095	0.099	0.639	0.657	0.098	0.101
$X_i = \sum_{j=1}^{i} U_j$	0.5	0.650	0.655	0.102	0.102	0.644	0.649	0.101	0.101	0.642	0.654	0.096	0.098
	0.85	0.655	0.657	0.097	0.104	0.649	0.654	0.098	0.100	0.637	0.643	0.096	0.096
$X_i = \sum_{s=1}^{i} \beta_s$	0.5	0.639	0.653	0.105	0.106	0.639	0.658	0.101	0.102	0.640	0.651	0.101	0.104
	0.85	0.646	0.658	0.096	0.101	0.631	0.648	0.101	0.103	0.636	0.648	0.096	0.099
$X_i = \sum_{s=1}^{i} K_s$	0.5	0.550	0.659	0.077	0.097	0.542	0.663	0.077	0.099	0.527	0.643	0.074	0.094
	0.85	0.555	0.659	0.075	0.070	0.539	0.647	0.073	0.093	0.545	0.652	0.076	0.105
$X_i = \sum_{s=1}^{i} L_s$	0.5	0.511	0.684	0.072	0.089	0.484	0.665	0.069	0.099	0.460	0.658	0.063	0.097
	0.85	0.530	0.654	0.061	0.047	0.481	0.648	0.064	0.080	0.470	0.643	0.062	0.098
dim(X)=10													
$X_i \sim$iid $N(0,1)$	0.5	0.998	0.999	0.097	0.109	0.996	0.998	0.092	0.105	0.996	0.998	0.096	0.105
	0.85	0.996	0.999	0.060	0.070	0.995	0.998	0.084	0.097	0.997	0.998	0.089	0.097
$X_i \sim$iid $U[0,1]$	0.5	0.997	0.998	0.100	0.112	0.997	0.998	0.092	0.101	0.997	0.998	0.089	0.099
	0.85	0.997	0.999	0.063	0.079	0.997	0.999	0.094	0.104	0.997	0.999	0.094	0.102
$X_i \sim$iid $e^{N(0,1)}$	0.5	0.990	0.999	0.075	0.092	0.974	0.999	0.048	0.094	0.962	0.998	0.043	0.098
	0.85	0.986	0.999	0.029	0.016	0.976	0.999	0.048	0.055	0.963	0.998	0.042	0.091

Note: eAUC is the area under the empirical ROC curve constructed from prediction rules of the form $\hat{Y} = 1(\hat{p} > c)$, where $\hat{p}$ is the fitted value from an OLS regression of Y on X and a constant. pAUC is the normal-parametric estimate of AUC based on the same index $\hat{p}$. The columns labeled 'Trad(itional)' contain empirical rejection rates from traditional tests using normal critical values, while the columns labeled 'Corr(ected)' are based on the proposed tests with chi-squared critical values. Z_s, U_s, β_s, K_s and L_s denote iid standard normal, uniform$[0,1]$, $\beta(2,1)$, χ^2_1 and lognormal$(0,1)$ random variables, respectively.

Table 3: Actual rejection rates of $H_0: AUC = 1/2$ based on the bootstrap

		$n = 100$				$n = 500$			
		$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.1$	
$\dim(X)$=1	p	eAUC	pAUC	eAUC	pAUC	eAUC	pAUC	eAUC	pAUC
$X \sim N(0,1)$	0.5	0.110	0.113	0.201	0.210	0.105	0.102	0.211	0.211
$X \sim U[0,1]$	0.5	0.107	0.100	0.207	0.201	0.107	0.104	0.210	0.208
$\dim(X)$=2									
$X_i \sim$iid $N(0,1)$	0.5	0.191	0.198	0.313	0.319	0.193	0.186	0.315	0.321
$X_i \sim$iid $U[0,1]$	0.5	0.205	0.194	0.327	0.321	0.197	0.196	0.321	0.323
$\dim(X)$=3									
$X_i \sim$iid $N(0,1)$	0.5	0.275	0.278	0.401	0.412	0.275	0.272	0.406	0.408
$X_i \sim$iid $U[0,1]$	0.5	0.277	0.270	0.413	0.401	0.271	0.267	0.403	0.407
$\dim(X)$=3									
$X_i = \sum_{s=1}^{i} Z_s$	0.5	0.276	0.283	0.398	0.411	0.262	0.271	0.394	0.411
$X_i = \sum_{s=1}^{i} U_s$	0.5	0.279	0.262	0.413	0.410	0.262	0.270	0.405	0.409