CENTRAL EUROPEAN UNIVERSITY

DEPARTMENT OF ECONOMICS AND BUSINESS WORKING PAPERS

economics.ceu.edu

Using the Area Under an Estimated ROC Curve to Test the Adequacy of Binary Predictors

by

Robert P. Lieli[1]

Yu-Chin Hsu[2]

2018/1

[1] Department of Economics, Central European University, Budapest. Email: lielir@ceu.edu

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

Abstract

We consider using the area under an empirical receiver operating characteristic (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. This corresponds to the null hypothesis that the area in question, denoted as AUC, is 1/2. We show that if the predictive index comes from a first stage regression model estimated over the same data set, then testing the null based on standard asymptotic normality results leads to severe size distortion in general settings. We then analytically derive the proper asymptotic null distribution of the empirical AUC in a special case; namely, when the first stage regressors are Bernoulli random variables. This distribution can be utilized to construct a fully in-sample test of H0 : AUC = 1=2 with correct size and more power than out-of-sample tests based on sample splitting, though practical application becomes cumbersome with more than two regressors.

Keywords: area under the ROC curve, overfitting, in-sample hypothesis testing, binary classification, model evaluation

Acknowledgements: This article is based on two earlier working papers; one with the same title and one titled “The Null Distribution of the Empirical AUC for Classifiers with Estimated Parameters: a Special Case.” We thank two anonymous referees and the associate editor for useful comments. All errors are our responsibility.

1 Introduction

The nonparametric empirical receiver operating characteristic (ROC) curve is a standard statistical tool used to evaluate the performance of a binary classifier composed of a predictive index (signal) and a cutoff that determines whether a given index value corresponds to a positive or negative outcome prediction. Varying the cutoff induces a tradeoff between the probability of producing a true positive vs. false positive prediction; the ROC curve gives a complete summary of these tradeoffs. The concept originates from the engineering literature on signal detection (e.g., Green and Swets 1966, Egan 1975), but is now routinely employed in fields such as medical diagnostics, meteorology, pattern recognition, etc. In recent years ROC analysis has become more common in financial and economic applications as well.

There are a number of statistics associated with an ROC curve that characterize the performance of the underlying forecasting model in various ways; see, e.g., Pepe (2003, Ch. 4). The area under the curve (AUC), in particular, can be thought of as a measure of overlap between the conditional distributions of the predictive index in the positive vs. the negative state. If there is no overlap at all (i.e., the signal perfectly discriminates), then the theoretical AUC is 1, while if there is perfect overlap (i.e., the signal is not informative), then AUC is 1/2. This last result is due to the fact that if the signal is statistically independent of the outcome, the population ROC curve coincides with the 45° diagonal of the unit square.

Though its relevance is sometimes debated (e.g., Hand 2009), AUC is ubiquitous as a statistic to characterize the overall predictive power of binary forecasting models much like an R^2 statistic is used to gauge the fit of a linear regression. As well known, an R^2 close to zero means that the regressors are not capable of explaining the variation in the dependent variable. Similarly, an empirical AUC close to 1/2 means that the classifier is no better than flipping an unbalanced coin. Hence, rejection of the null that AUC $= 1/2$ constitutes as low a bar as one can set for the usefulness of a binary prediction model.

There are well-known results in the statistical literature that describe the asymptotic distribution of the empirical AUC (Bamber 1975) or the difference between two empirical AUCs (DeLong et al. 1988). The setting in which these results are derived assumes that the signal used in constructing the ROC curve is either directly observed (it is 'raw data') or it is a fixed (non-random) function of 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 developed (e.g., Lehmann 1999, Ch. 3, 6). Nevertheless, in many practical applications binary forecasts are derived from predictive indices that are themselves outputs of a statistical model with pre-estimated parameters. For example, given a multitude of potential predictors, a researcher may first estimate the conditional probability of a positive outcome using a linear probability model or a logit regression, and then use the fitted values as a predictive index to construct an ROC curve over the same sample.

The first contribution of this paper is to demonstrate, both analytically and through simulations, that such first-stage estimation has nontrivial implications for how one should conduct in-sample tests of the hypothesis that AUC $= 1/2$ versus AUC $> 1/2$. In particular, the standard asymptotic normality result for the empirical AUC fails, and the traditional test based on it is *severely* oversized. Intuitively speaking, the problem is that even with uninformative predictors, first stage estimation tends to find pure chance patterns in the data, and 'overfits' the model in a way that artificially boosts the area under the in-sample ROC curve.

We contend that the problems arising from pre-estimation are not well understood in the applied literature. For example, in an influential paper, published in a top economics journal, Schularik and Taylor (2012) predict financial crisis episodes in various countries by regressing a crisis indicator on lagged credit growth, country fixed effects, etc. In evaluating their model they state that "[t]he AUC provides a simple test against the null value of 0.5 with an *asymptotic normal distribution*, and for our baseline model [the in-sample] AUC=0.717 with a standard error of just 0.0349" (emphasis added). As we will show, the asymptotic distribution is non-normal in this situation, and the reported standard errors are likely to be downward biased. Thus, judging whether the empirical AUC is significantly different from $1/2$ based on a standard t-test can be very misleading.

A paper that points out a problem similar to ours is Demler et al. (2012). The authors are concerned with using the DeLong et al. (1988) test to compare the in-sample AUCs of nested models with pre-estimated parameters. They observe the failure of the asymptotic normality of the estimated AUC differential under the null of no improvement, but their results do not encompass ours for a number of reasons. Most importantly, their model comparison does not include the special case in which the smaller model is degenerate (the authors are actually not very clear about this point). What they find is that pre-estimation causes the DeLong test to be overly conservative, i.e., it often fails to recognize improvements in predictive power. While this may be true when the smaller model is already informative, it stands in stark contrast with the case where the benchmark is 'no predictive power'. Instead of being conservative, the traditional test (equivalent to the DeLong test) severely *overrejects*, making it look as if the model easily beats this benchmark. Thus, if a researcher naively extrapolates the Demler et al. (2012) findings to this case, she will trust a rejection even more, because she will believe it was made by a conservative test![1]

Our second major contribution is more constructive. We analytically characterize the asymptotic null distribution of the empirical AUC in the special case where the first stage model is an OLS regression and the regressors are Bernoulli random variables. (No comparable result is given by Demler et al. 2012.) The results apply equally to a first stage logit regression or linear discriminant analysis. While our characterization, in principle, allows for any number of Bernoulli predictors, it will be clear that writing down the asymptotic distribution explicitly becomes very cumbersome as the number of regressors increases. We therefore state the asymptotic distribution of the empirical AUC completely explicitly only in the two-regressor case. We also present Monte Carlo simulations to assess the quality of our asymptotic approximations.

From a practical standpoint, our results make it obvious that traditional tests of H_0 : AUC $= 1/2$ can be very misleading when some model parameters are estimated in-sample. Nevertheless, there are ways to get around this problem in practice. The easiest solution is to estimate the predictive model on a training sample, and to test it on an independent evaluation sample. Traditional inference based on asymptotic normality remains valid in this case.

Our third contribution, therefore, is to investigate the relative merits of in-sample vs.

[1] We conjecture that the conservativeness result by Demler et al. (2012) is overturned already when the smaller model has weak predictive power rather than none. We do not investigate this conjecture here.

out-of-sample tests of H_0 : AUC = 1/2. As we show through simulations, splitting the sample entails a non-trivial power loss, and whenever applicable, our analytical results help avoid this loss by facilitating valid in-sample inference using all available observations. We also provide an empirical illustration using German loan data (the outcome to be predicted is the borrower's standing). These exercises focus on the case with two Bernoulli regressors as the asymptotic theory is best developed in this (admittedly limited) setting.

The rest of the paper is organized as follows. We introduce the ROC curve and the standard test of AUC = 1/2 in Section 2. Section 3 documents and explains the failure of this test for models with parameters pre-estimated over the same sample. In Section 4 we derive the asymptotic distribution of the empirical AUC in the special setting discussed above. Section 5 presents Monte Carlo evidence on (i) the accuracy of our asymptotic approximation (Section 5.1); and (ii) the power of in-sample vs. out-of-sample tests of H_0 (Section 5.2). We present the illustrative application in Section 6. Section 7 concludes. Proofs are provided in a technical Appendix.

2 Binary predictors and the ROC curve

2.1 The population ROC curve and AUC

Let $Y \in \{0, 1\}$ be a binary outcome. Given a d-dimensional vector X of covariates (predictors), a classifier is a function that maps the possible values of X into $\{0, 1\}$, i.e., produces a point forecast of Y 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

$$\hat{Y}(c) = 1(g(X) > c), \tag{1}$$

where $g : \mathbb{R}^k \to \mathbb{R}$ is a fixed real valued function that maps X into scalar predictive index $g(X)$, and c is a cutoff for predicting one vs. zero.

Example 1 If a test result X exceeds a certain threshold c, a doctor issues a positive diagnosis, otherwise a negative one. Here $g(X) = X$. Whether the condition is actually present (Y) will be confirmed later on. ∎

Example 2 Suppose that there is a vector X of predictors. Let $\ell(\hat{y}, y)$ be a loss function that specifies the cost of predicting $Y = \hat{y} \in \{0,1\}$ when the actual outcome is $Y = y \in \{0,1\}$. For any given X the prediction that minimizes expected loss can be obtained by solving

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

If $\ell(1,1) < \ell(0,1)$ and $\ell(0,0) < \ell(1,0)$, then it is straightforward 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 (see, e.g., Elliott and Lieli 2013). Thus, cutoff rules are theoretically optimal in a wide range of settings, and the information content of X about Y is best summarized by the conditional probability $g(X) = P(Y = 1 \mid X)$. ∎

If one varies the cutoff c in (1) between plus and minus infinity, the pair of probabilities

$$(F(c), T(c)) = \Big(\mathbb{P}(\hat{Y}(c) = 1 \mid Y = 0), \mathbb{P}(\hat{Y}(c) = 1 \mid Y = 1)\Big),$$

called the false positive rate and true positive rate, respectively, trace out the population ROC curve in the unit square $[0,1]\times[0,1]$. For a given false positive rate one would generally like to maximize the true positive rate and, conversely, for a given true positive rate, one would like to minimize the false positive rate. Thus, the 'bulgier' the ROC curve is toward the northwest, at least for a suitable range of cutoffs, the more informative the underlying index is. This is an intuitive reason why AUC, the area under the ROC curve, is considered an overall measure of classification performance.[2] A typical population ROC curve is shown in Figure 1.

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

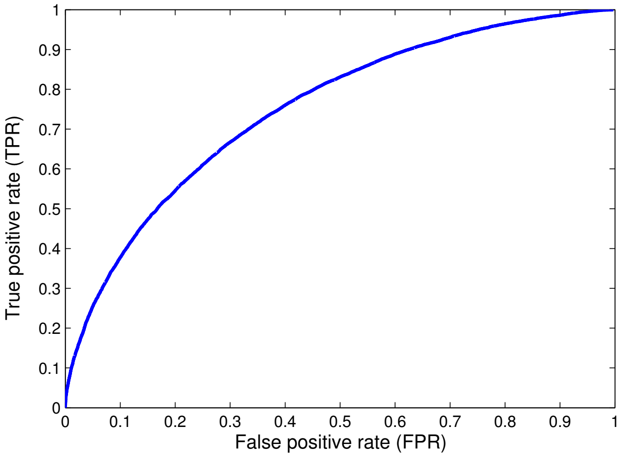

Figure 1: A population ROC curve

To state a more precise interpretation of AUC, let Z_1 and Z_0 be two independent random variables with $Z_1 \sim g(X)\big|Y = 1$ and $Z_0 \sim g(X)\big|Y = 0$. Then, as shown by Bamber (1975),

$$\mathrm{AUC} = \mathbb{P}(Z_1 > Z_0) + 0.5\mathbb{P}(Z_0 = Z_1). \tag{2}$$

If, in particular, the (conditional) distribution of $g(X)$ is continuous, then AUC is simply $\mathbb{P}(Z_1 > Z_0)$. Thus, AUC gives the probability of a randomly chosen positive outcome to be associated with a higher predictive index than a randomly chosen negative outcome.

Suppose that $g(X)$ has no predictive power, i.e. that it is independent of Y. In this case $F(c) = T(c) = \mathbb{P}(g(X) > c)$, so that the ROC curve runs along the 45°-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.

2.2 The empirical ROC curve and standard inference about AUC

Given a random sample of observations $\{(X_i, Y_i)\}_{i=1}^n$, the empirical ROC curve is constructed from the empirical false positive rate $\hat{F}(c)$ and the empirical true positive rate $\hat{T}(c)$:

$$\hat{F}(c) = \frac{1}{n_0}\sum_{i=1}^n 1(g(X_i) > c)(1 - Y_i) \quad \text{and} \quad \hat{T}(c) = \frac{1}{n_1}\sum_{i=1}^n 1(g(X_i) > c)Y_i,$$

where $n_1 = \sum_{i=1}^{n} Y_i$ and $n_0 = n - n_1$. As c varies between plus and minus infinity, the pair $(\hat{F}(c), \hat{T}(c))$ 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. The empirical AUC, the area under the empirical ROC curve, is denoted as eAUC.

The empirical AUC has an interpretation analogous to (2). Let $\{X_{0,i}\}_{i=1}^{n_0}$ and $\{X_{1,j}\}_{j=1}^{n_1}$ denote the predictor values over the $Y = 0$ and $Y = 1$ subsamples, respectively. Define

$$\hat{U} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} 1\{g(X_{0,i}) < g(X_{1,j})\} \text{ and } \hat{U}' = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} 1\{g(X_{0,i}) = g(X_{1,j})\} \quad (3)$$

so that $\hat{U}$ is the sample proportion of observation pairs chosen from the two subsamples with $g(X_{0,i}) < g(X_{1,j})$ and $\hat{U}'$ is the sample proportion of ties. Then $e\text{AUC} = \hat{U} + 0.5\hat{U}'$. This formula reveals a close relationship between eAUC and the (two-sample) Mann-Whitney U-statistic, whose asymptotic theory is well developed (see Bamber 1975, Lehmann 1999, Ch. 3, 6).

The null hypothesis we will focus on testing in the rest of this paper is $H_0 : \text{AUC} = 1/2$, using the empirical AUC as the test statistic. More precisely, we will maintain the slightly stronger underlying hypothesis that the predictive index $g(X)$ is statistically independent of the outcome Y because X itself is uninformative about Y.[3] Drawing on the U-statistics literature, Bamber (1975) states a general asymptotic normality result for eAUC. If, in particular, Y is independent of X, the result can be simplified to

$$\frac{e\text{AUC} - 1/2}{\sqrt{\frac{Bn}{4n_0 n_1}}} \overset{a}{\sim} N(0,1), \quad (4)$$

where B is a suitable constant. For example, if $g(X)$ is continuously distributed, then $B = 1/3$; the general formula and an estimator for B is provided by Bamber (1975).[4]

[3] In theory, it is possible that $g(X)$ is not independent of Y but AUC=1/2 because part of the population ROC curve runs below the main diagonal and part of it runs above. However, in this case one could switch to a decision rule with $AUC > 1/2$ simply by flipping the outcome predictions for those cutoffs values for which $T(c) < F(c)$.

[4] Let Z_1, Z_2 and Z_3 denote independent random variables with the same distribution as $g(X)$. Then, under the null, $B = P(Z_1, Z_2 < Z_3) + P(Z_3 < Z_1, Z_2) - 2P(Z_1 < Z_3 < Z_2)$.

3 The failure of standard inference about AUC

In Example 2 the optimal index is, theoretically, a fixed function of the covariates. Nevertheless, in practice $P(Y = 1 \mid X)$ is typically unknown, and needs to be estimated, e.g., by a linear probability model or logit regression. In the first case one approximates the conditional probability function by a linear projection, i.e.,

$$Y = \alpha + X'\beta + U,$$

where the error U is uncorrelated with X, and the coefficients α and β are estimated by OLS from a random sample $\{(X_i, Y_i)\}_{i=1}^n$. The predictive index for Y_i is given by the fitted value $\hat{g}(X_i) = \hat{\alpha} + X_i'\hat{\beta}$, where the coefficients $\hat{\alpha}$ and $\hat{\beta}$ depend on the entire sample. The index $\hat{g}(X_i)$, which can be interpreted as an estimate of the ex-ante conditional probability that $Y = 1$, is then used as in (1) to construct the sample ROC curve.

In the second case the model is

$$Y = 1(\alpha + X'\beta + U > 0),$$

where $-U$ is assumed to follow the logistic distribution and be independent of X. The coefficients α and β are estimated by ML. The predictive index for Y_i is given by the fitted value $\hat{g}(X_i) = \Lambda(\hat{\alpha} + X_i'\hat{\beta})$, where $\Lambda(\cdot)$ is the logistic c.d.f., and $\hat{\alpha}$ and $\hat{\beta}$ depend on the entire sample.[5]

When the predictive index contains parameters estimated in-sample, the observations $\{(\hat{g}(X_i), Y_i)\}_{i=1}^n$ are no longer independent, and it is not clear whether the standard asymptotic theory for eAUC presented in Section 2.2 applies. In particular, under the null of independence, $\hat{\beta}$ converges to $\beta = 0$ at the rate of $\sqrt{n}$, but of course $\hat{\beta}$ will never exactly be zero in finite samples. As we will shortly show, this is enough to prevent eAUC from being asymptotically normal.

In the rest of the paper we will focus on linear regression as the first stage predictive model. Nevertheless, all results apply equally to a logit regression, as the two are asymptotically equivalent under the null in the sense that $\hat{\beta}_{ML} = (const.) \times \hat{\beta}_{OLS} + o_p(n^{-1/2})$.

[5] If the only goal is ROC analysis, one could simply set the predictive index as $\hat{g}(X_i) = X_i'\hat{\beta}$ in both cases because neither the constant $\hat{\alpha}$ nor the increasing function $\Lambda(\cdot)$ affects the sample ROC curve.

3.1 Intuition and analytical examples

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

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

Let $e\text{AUC}_X$ denote the area under the former and $e\text{AUC}_{-X}$ the area under the latter ROC curve. It is not hard to show that the two curves are symmetric about the point $(1/2, 1/2)$ so that $e\text{AUC}_X = 1 - e\text{AUC}_{-X}$, i.e., $e\text{AUC}_X > 1/2 \Leftrightarrow e\text{AUC}_{-X} < 1/2$. The asymptotic null distribution (4) applies to $e\text{AUC}_X$ as well as $e\text{AUC}_{-X}$.

Given $g(x) = \alpha + \beta x$, the decision rule $1(g(X_i) > c)$ is clearly equivalent to $1(X_i > c)$ for any $\beta > 0$ and to $1(-X_i > c)$ for any $\beta < 0$. Now suppose that we let an OLS regression of Y on X "pick" the value of α and β as described above.[6] The regression coefficients $\hat{\alpha}$ and $\hat{\beta}$ are given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\bar{Y}(1-\bar{Y})(\bar{X}_1 - \bar{X}_0)}{\widehat{Var}(X)} \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}, \tag{5}$$

where $\bar{Y}$ is the full sample mean of Y, and $\bar{X}_j$, $j = 0, 1$ is the sample mean of X in the $Y = j$ subsample. We then apply the estimated index $\hat{g}(X_i) = \hat{\alpha} + \hat{\beta}X_i$ to obtain point forecasts $\hat{Y}_i(c)$ as in (1), and construct the empirical ROC curve by varying c.

As a result of this procedure, the rule $1(\hat{g}(X_i) > c)$ will be a random "mixture" between Rule$(+X)$ and Rule$(-X)$ no matter how small $\hat{\beta}$ is in absolute value.[7] Let $e\text{AUC}_M$ denote the area under the associated ROC curve (the subscript M stands for mixture or model). Clearly,

$$e\text{AUC}_M = (e\text{AUC}_X) \cdot 1(\hat{\beta} > 0) + (e\text{AUC}_{-X}) \cdot 1(\hat{\beta} < 0). \tag{6}$$

While $e\text{AUC}_X$ and $e\text{AUC}_{-X}$ are asymptotically normal, this is generally not true for $e\text{AUC}_M$. The reason is that the sign of $\hat{\beta}$ is correlated with the in-sample classification perfor-

[6] Of course, when X is a scalar predictor, there is no practical reason to run such a preliminary regression. We consider this setting because it offers the simplest way to explain why the asymptotic normality of eAUC fails with pre-estimation.

[7] By the law of the iterated logarithm, the probability of the event $\hat{\beta} = 0$ is asymptotically zero even as $\hat{\beta}$ converges to zero in probability under the null. We will therefore ignore this event.

mance of Rule($+X$) versus Rule($-X$). That is, $\hat{\beta}$ is likely to be positive when $e\text{AUC}_X > 1/2$ and negative when $e\text{AUC}_{-X} > 1/2$. More specifically, if X and Y are independent, Rule($+X$) and Rule($-X$) are equally useless in the population, but in finite samples one of the rules will still slightly outperform the other just by random variation. While the absolute value of $\hat{\beta}$ approaches zero, the sign of $\hat{\beta}$ will still adjust to this difference, and will correspond to the better of the two rules. Therefore, $e\text{AUC}_M$ is unlikely to fall below 1/2, i.e., the distribution of $e\text{AUC}_M$ cannot be symmetric around 1/2.

We formally illustrate the intuition outlined above in two special cases: when X is binary and when X is uniform over [0,1].

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

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

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 $\hat{F}_X = \bar{X}_0$, the second formula for $\hat{\beta}$ in display (5) implies $sign(\hat{\beta}) = sign(e\text{AUC}_X - 1/2)$. In other words, the events $\hat{\beta} > 0$ 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, one can start from equation (6) to show that

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

Therefore, by (4),

$$\frac{\sqrt{n}(e\text{AUC}_M - 1/2)}{\sqrt{\frac{Bn}{n_0 n1}}} = \frac{\sqrt{n}|e\text{AUC}_X - 1/2|}{\sqrt{\frac{Bn}{n_0 n1}}} \overset{a}{\sim} |N(0, 1)|.$$

Thus, a one-sided t-test of H_0 : AUC = 1/2 vs. H_1 : AUC > 1/2 based naively on (4) is twice as likely to reject the null as the chosen nominal size. ∎

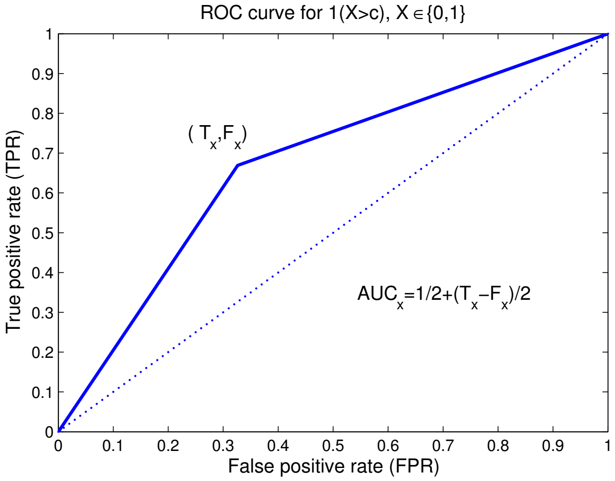

Figure 2: ROC curve with binary predictor

Example 4 If X is a continuous random variable in Rule(X), $e\mathrm{AUC}_X$ is given by the U-statistic $\hat{U}$, defined in (3), with $g(x) = x$. The general influence function representation of $\hat{U}$ is developed, for example, in Lehmann (1999, Ch. 6). Under the additional assumption that X is uniform $[0, 1]$, and X is independent of Y, this representation implies

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

Then, by the second expression for $\hat{\beta}$ under (5), the sign of $\hat{\beta}$ coincides with the sign of $e\mathrm{AUC}_X - 1/2$ with probability approaching one, i.e., the events $\hat{\beta} > 0$ and $e\mathrm{AUC}_X > 1/2$ are again perfectly correlated, albeit asymptotically. By the same argument as in Example 3, a one-sided t-test of H_0 : AUC $= 1/2$ based naively on (4) has asymptotic size twice the chosen nominal size. ∎

Examples 3 and 4 are special in that the correlation between the events $\hat{\beta} > 0$ and $e\mathrm{AUC}_X > 1/2$ 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 $e\mathrm{AUC}_M < 1/2$. Thus, the limit distribution of $\sqrt{n}(e\mathrm{AUC}_M - 1/2)$ is not always the absolute value of a mean zero normal, but it is not normal either.

The discussion so far has concentrated on scalar predictors X. Nevertheless, the intuition about the first stage regression of Y on X creating spurious matches between the outcomes and the predicted values carries over to the case when X is a vector. The following general result provides further insight.

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

(i) $\hat{\beta} = \bar{Y}(1-\bar{Y})\hat{M}^{-1}(\bar{X}_1 - \bar{X}_0)$, *where* $\hat{M} = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$.

(ii) $(\bar{X}_1 - \bar{X}_0)'\hat{\beta} > 0$.

Part (i) of Lemma 1 follows from standard expressions for the OLS estimator and some algebra. It generalizes the formula given for the slope coefficient in equation (5). Part (ii) is a consequence of $\hat{M}$ being positive definite.

Part (ii) of Lemma 1 implies that OLS will always 'choose' the slope coefficients so that they satisfy $\bar{X}_1'\hat{\beta} > \bar{X}_0'\hat{\beta}$. Using the notation introduced before equation (3), this generally means that there are more $(X_{0,i}, X_{1,j})$ pairs chosen from the two subsamples with

$$\hat{g}(X_{1,j}) = \hat{\alpha} + X_{1,j}'\hat{\beta} > \hat{g}(X_{0,i}) = \hat{\alpha} + X_{0,i}'\hat{\beta}$$

than the other way around (unless the distribution of X is very outlier prone and the sample is small). But based on (3), this is precisely the same as saying that the empirical AUC is greater than 1/2 (at least with high probability). Thus, the overfitting problem persists for multivariate X, and as we will shortly see, using (4) to test H_0 : AUC $= 1/2$ can causes more severe size distortion as the dimension of X increases.

3.2 Monte Carlo evidence

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 fat tailed distributions, and cases where the components of X are correlated. For each specification of X, $\tau \equiv P(Y = 1)$ is fixed at two different levels, $\tau = 0.5$ and $\tau = 0.85$.

We draw 10,000 random samples of size $n = 100$, $n = 500$ and 5000 from the distribution of (X, Y). 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. We compute eAUC and then we 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 distribution stated in (4). Actual rejection rates over the 10,000 Monte Carlo repetitions are presented in Table 1.

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 $\hat{Y}_i(c) = 1(\hat{g}(X_i) > c)$ 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 related to the distribution of X. When X is a uniform[0,1] scalar, Example 4 shows 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 X is χ^2_1, there is roughly a 20% chance that $e\text{AUC}_M < 1/2$ even in large samples, yet actual size is still twice the nominal size. This suggests that while $\hat{\beta}$ and $e\text{AUC}_X - 1/2$ can have opposite signs, this does not happen when $e\text{AUC}_X - 1/2$ is sufficiently large in absolute value.

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 normal X are close to the bounded support case. These differences are likely due to 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.

4 Theoretical results for Bernoulli predictors

The general characterization of the asymptotic null distribution of eAUC in the presence of pre-estimated model parameters is a very challenging problem. We will present analytical results in a special case—when X is a vector of Bernoulli predictors.

Table 1: Simulated rejection rates of H_0 : AUC $= 1/2$ vs. H_0 : AUC $> 1/2$ under independence of X and Y based on traditional asymptotics

	$n = 100$				$n = 500$				$n = 5000$			
	$\alpha = 5\%$		$\alpha = 10\%$		$\alpha = 5\%$		$\alpha = 10\%$		$\alpha = 5\%$		$\alpha = 10\%$	
dim(X)=1 τ :	0.5	0.85	0.5	0.85	0.5	0.85	0.5	0.85	0.5	0.85	0.5	0.85
$X \sim N(0,1)$	0.102	0.102	0.202	0.206	0.097	0.099	0.199	0.199	0.097	0.099	0.200	0.198
$X \sim U[0,1]$	0.102	0.102	0.205	0.205	0.102	0.099	0.201	0.202	0.105	0.096	0.200	0.200
$X \sim Bern(.5)$	0.090	0.096	0.194	0.204	0.094	0.095	0.191	0.189	0.100	0.096	0.195	0.199
$X \sim \beta(2,1)$	0.097	0.103	0.202	0.207	0.104	0.102	0.195	0.203	0.099	0.097	0.205	0.206
$X \sim \chi^2_1$	0.101	0.102	0.198	0.203	0.100	0.102	0.197	0.199	0.096	0.102	0.192	0.198
$X \sim e^{N(0,1)}$	0.099	0.104	0.201	0.201	0.096	0.098	0.185	0.190	0.096	0.100	0.190	0.192
dim(X)=2												
X_i ∼iid $N(0,1)$	0.257	0.259	0.435	0.439	0.254	0.255	0.429	0.430	0.249	0.254	0.426	0.430
X_i ∼iid $U[0,1]$	0.263	0.262	0.444	0.447	0.258	0.257	0.441	0.441	0.258	0.258	0.431	0.440
X_i ∼iid $Bern(.5)$	0.235	0.233	0.426	0.423	0.237	0.244	0.412	0.429	0.223	0.235	0.413	0.421
X_i ∼iid $\beta(2,1)$	0.258	0.249	0.440	0.426	0.251	0.255	0.434	0.435	0.260	0.258	0.435	0.430
X_i ∼iid χ^2_1	0.220	0.223	0.373	0.385	0.221	0.210	0.367	0.362	0.207	0.210	0.362	0.363
X_i ∼iid $e^{N(0,1)}$	0.204	0.212	0.348	0.362	0.200	0.197	0.345	0.340	0.188	0.192	0.325	0.330
dim(X)=3												
X_i ∼iid $N(0,1)$	0.431	0.432	0.640	0.639	0.426	0.425	0.633	0.632	0.419	0.418	0.625	0.626
X_i ∼iid $U[0,1]$	0.441	0.443	0.650	0.653	0.433	0.434	0.644	0.646	0.432	0.426	0.641	0.643
X_i ∼iid $Bern(.5)$	0.407	0.404	0.628	0.630	0.418	0.407	0.637	0.627	0.412	0.636	0.421	0.630
X_i ∼iid $\beta(2,1)$	0.434	0.438	0.640	0.649	0.426	0.432	0.635	0.637	0.430	0.425	0.631	0.632
X_i ∼iid χ^2_1	0.372	0.378	0.551	0.573	0.360	0.354	0.540	0.540	0.352	0.350	0.536	0.536
X_i ∼iid $e^{N(0,1)}$	0.346	0.341	0.522	0.531	0.309	0.319	0.479	0.497	0.295	0.299	0.466	0.464
dim(X)=3												
$X_i = \sum_{s=1}^{i} Z_s$	0.435	0.432	0.640	0.638	0.429	0.426	0.635	0.634	0.428	0.432	0.630	0.639
$X_i = \sum_{s=1}^{i} U_s$	0.442	0.447	0.650	0.655	0.437	0.434	0.644	0.649	0.437	0.428	0.642	0.637
$X_i = \sum_{s=1}^{i} B_s$	0.408	0.400	0.621	0.631	0.415	0.419	0.631	0.641	0.428	0.428	0.639	0.643
$X_i = \sum_{s=1}^{i} \beta_s$	0.436	0.444	0.639	0.646	0.428	0.425	0.639	0.631	0.430	0.424	0.640	0.636
$X_i = \sum_{s=1}^{i} K_s$	0.359	0.367	0.550	0.555	0.356	0.353	0.542	0.539	0.347	0.356	0.527	0.545
$X_i = \sum_{s=1}^{i} L_s$	0.332	0.345	0.511	0.530	0.313	0.304	0.484	0.481	0.299	0.295	0.460	0.470
dim(X)=10												
X_i ∼iid $N(0,1)$	0.983	0.982	0.998	0.996	0.981	0.979	0.996	0.995	0.977	0.978	0.996	0.997
X_i ∼iid $U[0,1]$	0.983	0.983	0.997	0.997	0.981	0.983	0.997	0.997	0.982	0.983	0.997	0.996
X_i ∼iid $e^{N(0,1)}$	0.962	0.950	0.990	0.986	0.925	0.925	0.974	0.976	0.899	0.905	0.962	0.963

Note: α is the nominal significance level and $\tau = P(Y = 1)$. Z_s, U_s, B_s, β_s, K_s and L_s denote iid $N(0,1)$, uniform$[0,1]$, Bernoulli(0.5), $\beta(2,1)$, χ^2_1 and lognormal$(0,1)$ random variables, respectively.

4.1 The setup and some geometric arguments

Let X be a $d \times 1$ vector of Bernoulli predictors with variance-covariance matrix Σ_X, assumed to be non-singular, but not necessarily diagonal. The support of X is given by the vertices of the unit cube in $\mathbb{R}^d$, i.e., $S := \textit{support}(X) = \{0, 1\}^d$.[8] We want to predict Y based on a linear combination of the components X, i.e., we set $g(x) = x'b$ for some coefficient vector $b \in \mathbb{R}^d$, and consider decision rules of the form

$$\hat{Y}(c) = 1(X'b > c). \tag{7}$$

We denote the area under the corresponding population ROC curve as AUC_b, and the area under the corresponding empirical ROC curve as $e\text{AUC}_b$. The notation emphasizes the dependence of this quantity on the given value of b.

Our ultimate goal is to characterize the distribution of $e\text{AUC}_b$ not for fixed values of b, but rather when b is replaced by the vector of slope coefficients from an OLS regression of Y on X, estimated over the same random sample from which the empirical ROC curve is subsequently constructed. Nevertheless, to be able to treat this problem, we will first study the geometry of the ROC curve for fixed values of b.

Thinking of x as a vector of continuous variables, the equation $x'b = c$ defines a hyperplane in $\mathbb{R}^d$ for any given value of b and c. This hyperplane divides the support of X, the set $S = \{0, 1\}^d$, into two subsets—the set of points above the plane and the set of points below the plane. More formally, let

$$S_b^+(c) = \{s \in S : s'b > c\}$$

be the set of points in S above the plane. These are precisely those values of X for which a positive outcome ($Y = 1$) is predicted; hence the '+' superscript. As c varies, the $x'b = c$ hyperplane shifts up and down in a parallel fashion. For very large values of c, the set $S_b^+(c)$ is empty, and then it gradually expands as c decreases, until it becomes equal to S. Most values of b will possess the property that the points $s \in S$ enter $S_b^+(c)$ one at a time. More formally:

[8] More generally, the support could be a strict subset of S. The full support assumption is solely for convenience; all results stated in this note go through without it.

Definition 1 *We say that the point $b \in \mathbb{R}^d$ possesses the* gradual increase *(GI) property if for any given c with $S_b^+(c) \neq S$, there exists $c' < c$ such that $S_b^+(c') \setminus S_b^+(c)$ is a singleton.*

It is intuitively clear, and not hard to show formally, that this property holds for all $b \in \mathbb{R}^d$ except of a set of Lebesgue measure zero.[9]

For $s \in S$, let $P(s) = \mathbb{P}(X = s \mid Y = 1)$ and $Q(s) = \mathbb{P}(X = s \mid Y = 0)$. The true and false positive rates associated with a given value of c can be computed, respectively, as

$$T(c) = \sum_{s \in S_b^+(c)} P(s) \text{ and } F(c) = \sum_{s \in S_b^+(c)} Q(s). \tag{8}$$

Given $b \in \mathbb{R}^d$ with the GI property, let $s_{(1)}, s_{(2)}, \ldots, s_{(K)}$, $K := 2^d$, denote the unique order in which the points of S enter the set $S_b^+(c)$ as c decreases. (We will also say that b induces the ordering $s_{(1)}, s_{(2)}, \ldots, s_{(K)}$ on S.) Assuming $P(s)$ and $Q(s)$ have full support[10], each additional point contributes a positive amount to the sums in (8); if the drop in c is not large enough for a new point to enter, then $T(c)$ and $F(c)$ stay constant. Hence, the pair $(F(c), T(c))$ takes on $K + 1$ different values as c decreases; denote these, in order, by $(F_0, T_0), (F_1, T_1), \ldots, (F_{K-1}, T_{K-1}), (F_K, T_K)$, where $F_0 = T_0 = 0$ and $F_K = T_K = 1$. Let $P_k = P(s_{(k)})$ and $Q_k = Q(s_{(k)})$, $k = 1, \ldots, K$. With these definitions we can write

$$F_k = Q_1 + \ldots + Q_k \text{ and } T_k = P_1 + \ldots + P_k, \quad k = 0, 1, \ldots, K, \tag{9}$$

where the empty sum is interpreted as zero. Lemma 3 in Appendix A.1 states a closed form formula for AUC_b in terms of F_k and T_k.

Given a random sample $\{(X_i, Y_i)\}_{i=1}^n$, let $\hat{P}(s)$ and $\hat{Q}(s)$ denote the empirical versions of the measures P and Q, respectively; e.g., $\hat{P}(s) = \sum_{i=1}^n 1(X_i = s)Y_i / \sum_{i=1}^n Y_i$. The empirical ROC curve can be constructed exactly as described above, with $\hat{P}$ replacing P and $\hat{Q}$ replacing Q. It follows that for values of b with the GI property, the distribution of the random variable $e\text{AUC}_b$ depends on b only through the ordering it induces on S. This means that if b and b' induce the same ordering, then necessarily $e\text{AUC}_b \overset{d}{=} e\text{AUC}_{b'}$.

[9] A simple counterexample to the GI property is the zero vector. If $b = 0$, then the set $S_b^+(c)$ jumps from being the empty set to S as c drops from positive to negative.

[10] This assumption is not essential.

Finally, consider replacing the fixed value of b with a random point; specifically, the vector of slope coefficients from an OLS regression of Y on X and a constant, computed over the same sample. That is, we use the decision rule $\hat{Y}(c) = 1(X_i'\hat{\beta} > c)$, where $\hat{\beta}$ is computed as in Lemma 1. (Not including the constant α in the decision rule is clearly inconsequential for constructing the ROC curve.) We will denote the corresponding empirical AUC as $e\mathrm{AUC}_{\hat{\beta}}$.

4.2 Theoretical results

We will now characterize the asymptotic null distribution of $e\mathrm{AUC}_{\hat{\beta}}$. To this end, we first establish the joint distribution of $e\mathrm{AUC}_b$ and $\hat{\beta}$ under the null on independence for any *fixed* value of b.

Given $b \in \mathbb{R}^d$ with the GI property, let $s_{(1)}, \dots, s_{(K)}$ denote the ordering of S induced by b, and let $s_{(k),j} \in \{0,1\}$ be the jth component of $s_{(k)}$, $j \in \{1, \dots, d\}$. Define the $(K-1)\times(K-1)$ matrix V as

$$V(k,l) = \begin{cases} \mathbb{P}(X = s_{(k)})[1 - \mathbb{P}(X = s_{(k)})] & \text{if} \quad k = l \\ -\mathbb{P}(X = s_{(k)})\mathbb{P}(X = s_{(l)}) & \text{if} \quad k \neq l \end{cases} \tag{10}$$

$k, l = 1, \dots, K-1$. In addition, let the functions $g_0 : \mathbb{R}^{2(K-1)} \to \mathbb{R}$ and $g_1 : \mathbb{R}^{2(K-1)} \to \mathbb{R}^d$ be defined as

$$\begin{aligned} g_0(P_1, \dots, P_{K-1}, Q_1, \dots, Q_{K-1}) &:= \frac{1}{2} + \frac{1}{2}\sum_{k=1}^{K-1}(F_{k+1}T_k - F_kT_{k+1}) \\ g_1(P_1, \dots, P_{K-1}, Q_1, \dots, Q_{K-1}) &:= \tau(1-\tau)\Sigma_X^{-1}\begin{pmatrix} \sum_{k=1}^K P_k s_{(k),1} - \sum_{k=1}^K Q_k s_{(k),1} \\ \vdots \\ \sum_{k=1}^K P_j s_{(k),d} - \sum_{k=1}^K Q_k s_{(k),d} \end{pmatrix}, \end{aligned}$$

where the dependence of F_k and T_k on $P_1, \dots, P_{K-1}, Q_1, \dots, Q_{K-1}$ is as described in equation (9). The definitions of g_0 and g_1 come from the fact that

$$g_0(P_1, \dots, P_{K-1}, Q_1, \dots, Q_{K-1}) = AUC_b$$

by Lemma 3 in Appendix A.1, and

$$g_1(P_1, \dots, P_{K-1}, Q_1, \dots, Q_{K-1}) = \operatorname*{p\,lim}_{n\to\infty} \hat{\beta}$$

by Lemma 5 in Appendix A.1. Finally, under the null, V is the variance-covariance matrix of the random vectors $(\hat{P}_1, \dots, \hat{P}_{K-1})$ as well as $(\hat{Q}_1, \dots, \hat{Q}_{K-1})$; see Lemma 4 in Appendix A.1.

Set $g = \binom{g_0}{g_1}$. We state the following result.

Proposition 1 *Let $b \in \mathbb{R}^d$ possess the GI property. If X is independent of Y, the asymptotic joint distribution of $eAUC_b$ and $\hat{\beta}$ is given by*

$$\sqrt{n} \begin{pmatrix} eAUC_b - 1/2 \\ \hat{\beta} \end{pmatrix} \to_d N \left(0_{(1+d)\times 1}, \nabla g \begin{pmatrix} \frac{1}{\tau} V & 0 \\ 0 & \frac{1}{1-\tau} V \end{pmatrix} \nabla g' \right), \tag{11}$$

where ∇g is the $(1+d) \times 2(K-1)$ matrix with rows given by the gradients of the components of g, evaluated at $(P_1, \dots, P_{K-1}, Q_1, \dots, Q_{K-1})$ and $\tau = \mathbb{P}(Y = 1)$.

Remarks

1. Proposition 1, to our knowledge, is completely new in the literature.

2. The choice of b is arbitrary (as long as the GI condition holds), and does not necessarily coincide with the probability limit of $\hat{\beta}$, which is zero under the null. This is an important feature that will allow us to account for the estimation effect in the distribution of $eAUC_{\hat{\beta}}$.

3. The proof of Proposition 1 employs the multivariate delta method and is given in Appendix A.2.

We will now employ Proposition 1 to characterize the asymptotic distribution of $e\text{AUC}_{\hat{\beta}}$. Consider all possible $K!$ orderings (permutations) of the $K = 2^d$ points in S, and enumerate these orderings as $\ell = 1, 2, \dots, K!$.[11] Define

$$O_\ell = \{b \in \mathbb{R}^d : b \text{ has the GI property and } b \text{ induces the ordering } \ell\}, \quad \ell = 1, \dots, K!$$

[11] For example, for $d = 2$, $s_{(1)} = (1, 1)$, $s_{(2)} = (1, 0)$, $s_{(3)} = (0, 1)$, $s_{(4)} = (0, 0)$ is one of 4!=24 possible orderings; one could take this as $\ell = 1$, etc.

The sets O_ℓ are mutually exclusive; in fact, many of the O_ℓ are necessarily empty.[12] Furthermore, $\cup O_\ell$ covers the entire space $\mathbb{R}^d$ save for a set of Lebesgue measure zero (the set of points without the GI property). Clearly, $\hat{\beta} \in O_\ell$ iff $\sqrt{n}\hat{\beta} \in O_\ell$. As $\sqrt{n}\hat{\beta}$ is asymptotically normally distributed, $\mathbb{P}[\sqrt{n}\hat{\beta} \in \cup O_\ell] = \mathbb{P}[\hat{\beta} \in \cup O_\ell] \overset{a}{=} 1$, where $\overset{a}{=}$ denotes asymptotic equality (i.e., the limit as $n \to \infty$).

From each nonempty O_ℓ one can choose a fixed representative element b_ℓ. As argued in the previous section, the distribution of $e\text{AUC}_{\hat{\beta}}$ conditional on $\{\hat{\beta} \in O_\ell\}$ is then the same as the distribution of $e\text{AUC}_{b_\ell}$ conditional on $\{\hat{\beta} \in O_\ell\}$, because $\hat{\beta}$ and b_ℓ induce the same ordering on S. However, the latter conditional distribution is pinned down by the *joint* distribution of $e\text{AUC}_{b_\ell}$ and $\hat{\beta}$, which we have already characterized in Proposition 1. Using the law of total probability, we can then state the following general result:

Proposition 2 *The asymptotic null distribution of $eAUC_{\hat{\beta}}$ can be decomposed as:*

$$\mathbb{P}[\sqrt{n}(eAUC_{\hat{\beta}} - 1/2) \leq z] \overset{a}{=} \sum_{O_\ell \neq \emptyset} \mathbb{P}[\sqrt{n}(eAUC_{b_\ell} - 1/2) \leq z \mid \hat{\beta} \in O_\ell] \mathbb{P}(\hat{\beta} \in O_\ell), \tag{12}$$

where b_ℓ is a fixed representative element of O_ℓ for O_ℓ nonempty, and the conditional probabilities can be calculated from the joint distribution stated in Proposition 1.

Remarks

1. Terms k and ℓ in sum (12) can be combined if the distribution of $e\text{AUC}_{b_k}$ given $\hat{\beta} \in O_k$ is the same as the distribution of $e\text{AUC}_{b_\ell}$ given $\hat{\beta} \in O_\ell$. This occurs, for example, when the ordering induced by b_ℓ is the reverse of the ordering induced by b_k. We will show this formally in Appendix A.2.

2. An example of sum (12) being reduced to a single term is when Y and the components of X are mutually independent Bernoulli(0.5) random variables. By symmetry, $\hat{\beta} \in O_\ell$ then has the same probability for all nonempty O_ℓ, and the distribution of $e\text{AUC}_{b_\ell}$ given $\hat{\beta} \in O_\ell$ is the same for all such ℓ. Therefore, the sum in (12) reduces to a single term $\mathbb{P}[\sqrt{n}(eAUC_{b_\ell} - 1/2) \leq z \mid \hat{\beta} \in O_\ell]$.

[12] For example, if $d = 2$, there is no b with the GI property for which $s_{(1)} = (1, 1)$ and $s_{(2)} = (0, 0)$. A simple graph of S with some lines $x'b$ makes this clear.

In sum, Proposition 1 and 2 jointly characterize the asymptotic null distribution of $e\text{AUC}_{\hat{\beta}}$. This characterization is however rather intrinsic. To obtain a more explicit expression for the asymptotic distribution of $e\text{AUC}_{\hat{\beta}}$, one needs to (i) obtain an explicit expression for ∇g in equation (11), and (ii) find conditions on the components of $\hat{\beta}$ that determine which ordering it induces on S. These tasks become increasingly cumbersome as the dimensionality of X increases. Here we state the general result for $d = 2$ only. The following lemma deals with task (i).

Lemma 2 *Suppose that $d = 2$. Given any ordering $s_{(1)}, \ldots, s_{(4)}$ of S, the matrix ∇g is given by:*

$$\begin{aligned}\nabla g \quad &= \quad \begin{pmatrix} 1/2 & 0_{(1\times 2)} \\ 0_{(2\times 1)} & \tau(1-\tau)\Sigma_X^{-1} \end{pmatrix} \\ &\times \quad \begin{pmatrix} 1+Q_2+Q_3 & 1-Q_1+Q_3 & 1-Q_1-Q_2 & -(1+P_2+P_3) & -(1-P_1+P_3) & -(1-P_1-P_2) \\ s_{(1),1}-s_{(4),1} & s_{(2),1}-s_{(4),1} & s_{(3),1}-s_{(4),1} & s_{(4),1}-s_{(1),1} & s_{(4),1}-s_{(2),1} & s_{(4),1}-s_{(3),1} \\ s_{(1),2}-s_{(4),2} & s_{(2),2}-s_{(4),2} & s_{(3),2}-s_{(4),2} & s_{(4),2}-s_{(1),2} & s_{(4),2}-s_{(2),2} & s_{(4),2}-s_{(3),2} \end{pmatrix},\end{aligned}$$

where $s_{(1),j}$, $j = 1, 2$ denotes the jth component of $s_{(1)}$, etc.

The next result describes explicitly the asymptotic null distribution of $eAUC_{\hat{\beta}}$ in the bivariate case.

Proposition 3 (The bivariate case) *For $d = 2$ consider the following orderings of S:*

Ordering 1: $s_{(1)} = (1, 1)$, $s_{(2)} = (1, 0)$, $s_{(3)} = (0, 1)$, $s_{(4)} = (0, 0)$;

Ordering 2: $s_{(1)} = (1, 1)$, $s_{(2)} = (0, 1)$, $s_{(3)} = (1, 0)$, $s_{(4)} = (0, 0)$;

Ordering 3: $s_{(1)} = (1, 0)$, $s_{(2)} = (1, 1)$, $s_{(3)} = (0, 0)$, $s_{(4)} = (0, 1)$;

Ordering 4: $s_{(1)} = (1, 0)$, $s_{(2)} = (0, 0)$, $s_{(3)} = (1, 1)$, $s_{(4)} = (0, 1)$.

For each ordering $\ell = 1, \ldots, 4$, define the 3×3 matrix V_ℓ as in (10); the matrix ∇g_ℓ as in Lemma 2, and the matrix V_ℓ^* as the asymptotic variance matrix in (11). Let

$$(A_0, A_1, A_2) \sim N(0, V_1^*), \quad (B_0, B_1, B_2) \sim N(0, V_2^*)$$
$$(C_0, C_1, C_2) \sim N(0, V_3^*), \quad (D_0, D_1, D_2) \sim N(0, V_4^*)$$

be four independent jointly normal random vectors. Then, under the assumption that X and Y are independent,

$$
\begin{aligned}
&\mathbb{P}\left[\sqrt{n}(eAUC_{\hat{\beta}} - 1/2) \leq z\right] \\
&\quad \overset{a}{=} \mathbb{P}\left[A_0 \leq z \middle| A_1 > A_2 > 0\right] \times 2\mathbb{P}\Big[A_1 > A_2 > 0\Big] \\
&\quad + \mathbb{P}\left[B_0 \leq z \middle| B_2 > B_1 > 0\right] \times 2\mathbb{P}\Big[B_2 > B_1 > 0\Big] \\
&\quad + \mathbb{P}\left[C_0 \leq z \middle| C_1 > 0 > C_2, C_1 > |C_2|\right] \times 2\mathbb{P}\Big[C_1 > 0 > C_2, C_1 > |C_2|\Big] \\
&\quad + \mathbb{P}\left[D_0 \leq z \middle| D_1 > 0 > D_2, D_1 < |D_2|\right] \times 2\mathbb{P}\Big[D_1 > 0 > D_2, D_1 < |D_2|\Big],
\end{aligned}
$$

for any $z \in \mathbb{R}$.

Remarks

1. If X_1, X_2 and Y are jointly independent Bernoulli(.5) random variables, then $V_\ell^* = V^*$ for $\ell = 1, \ldots, 4$, where the matrix V^* is given by

   $$\begin{pmatrix} 5/16 & 1/2 & 1/4 \\ 1/2 & 1 & 0 \\ 1/4 & 0 & 1 \end{pmatrix}.$$

   In this case the formula for the limit distribution simplifies to $\mathbb{P}\left[A_0 \leq z \middle| A_1 > A_2 > 0\right]$.

2. Based on Proposition 3, it is straightforward to simulate the asymptotic null distribution of $\sqrt{n}(e\text{AUC}_{\hat{\beta}} - 1/2)$. First, one draws a very large number M of independent observations (three-dimensional vectors) from each of the four trivariate normal distributions stated in the theorem. Second, one discards those observations from each sample whose second and third components do not satisfy the condition stated in the corresponding conditional probability. Finally, one combines the remaining observations and extracts the empirical distribution of the first component. This empirical distribution approximates the distribution of $\sqrt{n}(e\text{AUC}_{\hat{\beta}} - 1/2)$, and one can compute quantiles, etc., as needed.

5 Monte Carlo simulations for Bernoulli predictors

We will conduct simulations to investigate two sets of questions. First, we want to explore the shape of the asymptotic distribution stated in Proposition 3, and assess how accurately it approximates the exact finite sample distribution of $e\text{AUC}_{\hat{\beta}}$ under the null. Second, we want to show that when applicable, these theoretical results facilitate more accurate inference about the null hypothesis $H_0 : \text{AUC} = 1/2$ than out-of-sample testing strategies. In particular, while out-of-sample tests restore the validity of the traditional distribution theory given in (4), they are subject to loss of power, because part of the data is held out for evaluation.

5.1 Approximating the finite sample distribution of eAUC$_{\hat{\beta}}$

We consider four data generating processes (DGPs); in each case $Y \in \{0, 1\}$ is independent of $X = (X_1, X_2)$. We vary the probability $\tau = \mathbb{P}(Y = 1)$ and the joint distribution from which (X_1, X_2) is drawn. More specifically:

DGP 1: $\tau = 0.5$; X_1 and X_2 are independent Bernoulli(0.5) random variables.

DGP 2: $\tau = 0.8$; X_1 and X_2 are independent Bernoulli(0.5) random variables.

DGP 3: $\tau = 0.5$; X_1 and X_2 are Bernoulli random variables with joint distribution

$$\mathbb{P}[X = (1, 1)] = .6; \mathbb{P}[X = (1, 0)] = .05; \mathbb{P}[X = (0, 1)] = .1; \mathbb{P}[X = (0, 0)] = .25.$$

DGP 4: $\tau = 0.8$; X_1 and X_2 are Bernoulli random variables with the same joint distribution as in DGP 3.

We draw samples of size $n = 60, 120$ and 1000 from each DGP, regress Y on X_1, X_2 and a constant, use the fitted values as a predictive index to construct the empirical ROC curve, and compute the empirical AUC. The procedure is repeated over 100,000 Monte Carlo iterations to approximate the actual distribution of $\sqrt{n}(eAUC_{\hat{\beta}} - 1/2)$ for $n = 60, 120, 1000$. The theoretical asymptotic distribution ($n = \infty$) is simulated as described in Remark 2 after Proposition 3 with $M = 10$ million. Table 2 compares various quantiles of these distributions

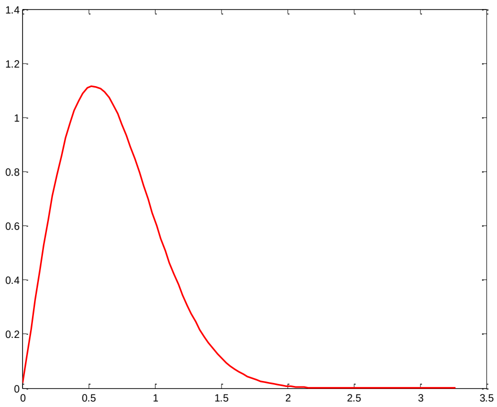

Figure 3: The asymptotic distribution of $\sqrt{n}(eAUC_{\hat{\beta}} - 1/2)$ for DGP 1

under the different DGPs. Figure 3 shows the (estimated) p.d.f. of the theoretical asymptotic distribution for DGP 1.

As seen in Figure 3, the asymptotic distribution of $\sqrt{n}(eAUC_{\hat{\beta}} - 1/2)$ is clearly non-normal; rather, it is markedly skewed to the right, which is a reflection of in-sample over-fitting. Table 2 further shows that the theoretical asymptotic distribution gives a very good approximation to the actual finite sample distribution already for $n = 60$, though the approximation is slightly poorer when the marginal distribution of Y is skewed. While a mathematical proof needs no further validation, these simulations show that the result stated in Proposition 3 is sound.

5.2 In-sample vs. out-of-sample inference about AUC

The Monte Carlo results in Table 1 on the size distortion of the traditional in-sample test of H_0 : AUC $= 1/2$ make a compelling case for out-of-sample inference, i.e., estimating and evaluating the model over separate data points. Given that these observations are independent, inference based on (4) is asymptotically valid. However, this way of eliminating the size distortion is not costless—by splitting the data set into a training sample and an evaluation sample, one loses some power relative to in-sample tests that make use of all available data. Though applicable only in limited situations, our analytic results make it

Table 2: Asymptotic approximation to the null distribution of $eAUC_{\hat{\beta}}$

	Distribution of $\sqrt{n}(eAUC_{\hat{\beta}} - 1/2)$			
	Actual			Asy.
Percentile	$n = 60$	$n = 120$	$n = 1000$	distribution
DGP 1	X_1 and X_2 independent, $\tau = 0.5$			
99th	1.660	1.651	1.649	**1.653**
95th	1.343	1.334	1.335	**1.332**
90th	1.177	1.170	1.170	**1.167**
75th	0.917	0.910	0.905	**0.905**
50th	0.648	0.645	0.641	**0.640**
25th	0.416	0.414	0.410	**0.412**
5th	0.169	0.177	0.172	**0.174**
DGP 2	X_1 and X_2 independent, $\tau = 0.8$			
99th	2.153	2.090	2.082	**2.066**
95th	1.730	1.690	1.670	**1.665**
90th	1.509	1.483	1.464	**1.459**
75th	1.164	1.149	1.136	**1.131**
50th	0.820	0.812	0.803	**0.799**
25th	0.528	0.522	0.517	**0.515**
5th	0.218	0.219	0.218	**0.217**
DGP 3	X_1 and X_2 dependent, $\tau = 0.5$			
99th	1.466	1.463	1.467	**1.466**
95th	1.183	1.175	1.177	**1.174**
90th	1.036	1.027	1.029	**1.026**
75th	0.804	0.795	0.793	**0.791**
50th	0.565	0.558	0.555	**0.555**
25th	0.359	0.356	0.353	**0.353**
5th	0.142	0.142	0.137	**0.139**
DGP 4	X_1 and X_2 dependent, $\tau = 0.8$			
99th	1.910	1.869	1.828	**1.832**
95th	1.519	1.485	1.465	**1.468**
90th	1.325	1.296	1.280	**1.282**
75th	1.014	1.002	0.987	**0.988**
50th	0.710	0.700	0.692	**0.693**
25th	0.450	0.446	0.442	**0.442**
5th	0.177	0.177	0.171	**0.173**

Note: In all cases, Y is independent of (X_1, X_2). Actual small sample distributions are based on 100,000 Monte Carlo simulations. The asymptotic distribution is constructed from 10 million draws from the distribution described in Proposition 3.

possible to achieve size control despite estimating and evaluating the predictive model over the same sample, and to gain power at the same time by using all available data for inference.

We present further simulations to demonstrate these points. The DGP employed in this exercise is of the following form:

DGP(β): $Y = 1$ with probability $p(X) = \Lambda(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$ and $Y = 0$ with probability $1 - p(X)$, where X_1 and X_2 are independent Bernoulli(0.5) predictors, $\Lambda(\cdot)$ is the logistic c.d.f., and $\beta = (\beta_0, \beta_1, \beta_2)$ are coefficients.

We consider various settings for the coefficient vector β. For simplicity, we fix $\beta_0 = \beta_2 = 0$ and vary the value of β_1 over the set $\{0, 0.1, 0.2, \ldots, 1\}$. Thus, for $\beta_1 = 0$ the null hypothesis H_0 : AUC $= 1/2$ is true (the DGP coincides with DGP 1), and it is false otherwise. Given a random sample of size n generated from DGP(β), we test H_0 in various ways:

(i) In-sample inference using the full sample of size n. We run an OLS regression of Y on $X = (X_1,\ X_2)$ and a constant using the full sample, compute the fitted values $\hat{g}(X_i) = \hat{\beta}_0 + \hat{\beta} X_{1i} + \hat{\beta}_2 X_{2i}$ for each observation $i = 1, \ldots, n$, construct the in-sample ROC curve, and compute $eAUC_{\hat{\beta}}$. We then test H_0 : AUC $= 1/2$ using the asymptotic distribution stated in Proposition 3.[13]

(ii) Split sample inference. We divide the n observations into two subsamples: for a fraction $q \in (0, 1)$, observations 1 through nq are used as a training sample, i.e., to estimate the regression of Y on $X = (X_1, X_2)$. The remaining $(1 - q)n$ observations are used for evaluation, i.e., we calculate $\hat{g}(X_i)$ for each $i = nq + 1, nq + 2, \ldots, n$, using the coefficients estimated over the training sample. We then construct the ROC curve for the evaluation sample, compute the empirical AUC, and test H_0 : AUC $= 1/2$ using the traditional Mann-Whitney distribution theory (4) with sample size $(1 - q)n$.

(iii) F-fold cross validation. This is a more sophisticated version of the split sample procedure. The full sample of size n is divided into F equal parts; in each step of the

[13] To mimic a practical application as closely as possible, we do not simulate the critical values ahead of time. Instead we estimate $V_1^*, \ldots, V_4^*$ in each Monte Carlo cycle from the data at hand and then use these estimates to simulate the distribution as described in Remark 2 after in Proposition 3.

procedure one part is held out for evaluation, and the regression of Y on X is estimated with the rest of the data. The estimated model is used to compute a predictive index for the held out outcomes. The procedure is repeated with each of the F parts being held out in turn. We thus obtain a predictive index value $\hat{g}(X_i)$ for each observation $i = 1, \ldots, n$, and compute the empirical AUC the usual way. Finally, we test $H_0 : \text{AUC} = 1/2$ using the Mann-Whitney distribution theory with sample size n.

Table 3 shows rejection rates of H_0 over 10,000 repetitions of these experiments for $n = 150$, 300, 600; $q = 1/2$, 2/3, 4/5; $F = 2$, 3, 5, and all possible values of β_1. The first column ($\beta_1 = 0$) shows the size of the tests while the rest of the columns ($\beta_1 = 0.1, \ldots, 1$) show power.

We make a number of observations. First, the in-sample test and the out-of-sample tests based on a single split are very accurately sized, while many-fold cross validation can be somewhat too conservative. Second, the in-sample test outperforms all other methods in terms of power, regardless of the value of β_1. The uniformly second best method is 2-fold cross validation, i.e., using one half of the sample for estimation, the other for validation, and then exchanging the roles. Third, the power advantage of the fully in-sample test even relative to 2-fold cross validation can be sizable; for $n = 300$ and $\beta_1 = 0.7$ it is as large as 12 percentage points. Figure 4 shows the power gain for this sample size over all values of β_1. In sum, the results demonstrate that there are non-negligible power gains to be had from in-sample testing of the hypothesis $H_0 : \text{AUC} = 1/2$ using all available data, at least for small to moderately large sample sizes.

6 Illustrative applications with Bernoulli predictors

To show the difference theoretical results such as those stated in Proposition 3 can make in a practical setting, we employ a classic data set on personal loans originally analyzed by Fahrmeir and Hamerle (1984). Other illustrative uses of the same data include Fahrmeir and Tutz (1994) and Lieli and White (2010). The data set contains information on 1,000 loan holders, all clients of a commercial bank in Southern Germany. The dependent variable

Table 3: Rejection probabilities of various tests of H_0 : AUC $= 1/2$

β_1:	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Method						$n = 150$					
In-sample	**0.051**	**0.054**	**0.078**	**0.113**	**0.176**	**0.251**	**0.344**	**0.436**	**0.539**	**0.654**	**0.735**
50-50 split	0.050	0.056	0.070	0.093	0.130	0.170	0.223	0.295	0.360	0.441	0.506
2-fold cv	0.049	0.055	0.074	0.099	0.151	0.215	0.280	0.372	0.455	0.559	0.648
67-33 split	0.047	0.051	0.060	0.078	0.116	0.137	0.186	0.228	0.278	0.330	0.401
3-fold cv	0.034	0.035	0.050	0.077	0.120	0.163	0.233	0.294	0.395	0.484	0.577
80-20 split	0.046	0.049	0.060	0.071	0.089	0.106	0.144	0.171	0.211	0.247	0.266
5-fold cv	0.020	0.024	0.029	0.053	0.079	0.119	0.177	0.240	0.321	0.415	0.480
						$n = 300$					
In-sample	**0.050**	**0.066**	**0.105**	**0.197**	**0.310**	**0.460**	**0.615**	**0.746**	**0.859**	**0.920**	**0.965**
50-50 split	0.049	0.065	0.086	0.150	0.212	0.305	0.423	0.524	0.646	0.741	0.806
2-fold cv	0.050	0.067	0.095	0.177	0.269	0.379	0.524	0.623	0.777	0.854	0.912
67-33 split	0.047	0.058	0.081	0.127	0.169	0.251	0.327	0.407	0.507	0.603	0.677
3-fold cv	0.033	0.045	0.071	0.136	0.200	0.328	0.457	0.585	0.715	0.825	0.903
80-20 split	0.045	0.054	0.070	0.096	0.129	0.186	0.237	0.276	0.347	0.430	0.476
5-fold cv	0.020	0.025	0.040	0.091	0.150	0.251	0.374	0.493	0.619	0.751	0.830
						$n = 600$					
In-sample	**0.052**	**0.080**	**0.174**	**0.349**	**0.563**	**0.768**	**0.904**	**0.969**	**0.992**	**0.999**	**1.000**
50-50 split	0.053	0.073	0.134	0.241	0.382	0.550	0.707	0.827	0.907	0.952	0.990
2-fold cv	0.054	0.073	0.155	0.296	0.479	0.674	0.830	0.924	0.973	0.992	0.997
67-33 split	0.049	0.063	0.110	0.191	0.304	0.439	0.573	0.689	0.790	0.864	0.920
3-fold cv	0.036	0.049	0.115	0.238	0.403	0.620	0.789	0.901	0.962	0.992	0.996
80-20 split	0.050	0.064	0.100	0.153	0.223	0.305	0.406	0.495	0.579	0.674	0.743
5-fold cv	0.021	0.035	0.081	0.178	0.313	0.498	0.707	0.829	0.926	0.970	0.992

Note: The figures are simulated rejection probabilities of the null hypothesis H_0 : AUC $= 1/2$ using various tests. The null hypothesis is true for $\beta_1 = 0$ and false otherwise. The in-sample test is based on the asymptotic distribution stated in Proposition 3, while the out-of-sample tests use (4). In all cases, and the nominal size is 5% and the number of Monte Carlo repetitions is 10,000.

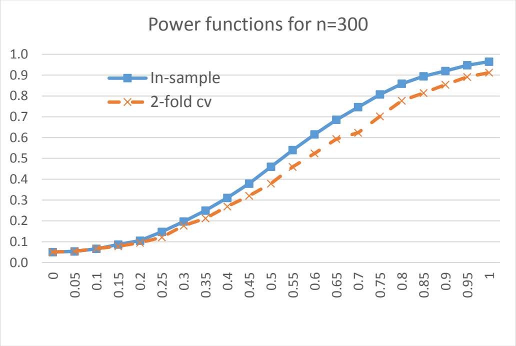

Figure 4: The power of the in-sample test vs. 2-fold cv as a function of β_1 for $n = 300$

is a binary dummy that takes the value one if the client is in good standing (i.e., the loan is being repaid as specified by the contract) and is zero otherwise. The data set also contains a large number of covariates describing the borrower's socioeconomic status and the loan contract, which can be taken as potential predictors of borrower standing.[14]

Our goal here not to conduct a systematic search for a model with high predictive ability or profitability; for such exercises see Fahrmeir and Tutz (1994) and Lieli and White (2010). Rather, we will examine some small, strictly illustrative models that fit the theoretical framework in Section 4, and show how taking pre-estimation into account can affect inference about the predictive ability of the model as measured by the empirical AUC.

Example 5 A potentially useful predictor of borrower status in the data set is monthly installment payment X as a fraction of income. The variable is recorded on a categorical scale: $X = 1$ if the fraction is above 35%; $X = 2$ between 35% and 25%; $X = 3$ between 25% and 20%, and $X = 4$ below 20%. We convert this variable into two dummies: an indicator of installment rates in excess of 35% ($irate\_hi$) and an indicator of installment rates below 20% ($irate\_lo$). Thus, the baseline category is the union of categories 2 and 3. The regression of the dependent variable ($status$) on the two dummies and a constant is given by:

[14] There are 700 individuals in the data set in good standing and 300 in bad standing. These numbers are a result of stratified sampling and do not accurately reflect the unconditional probability of default. We simply ignore this issue here and proceed as if the data were a random sample. (The estimated ROC curve and AUC are still consistent under stratified sampling but inference about AUC assumes a random sample.)

$$\widehat{status} = 0.7242 + \underset{(0.0456)}{0.0258} \times irate\_hi - \underset{(0.0313)}{0.0583} \times irate\_lo$$

As the results indicate, the linear relationship between installment rate and borrower standing is, somewhat surprisingly, not very strong or intuitive. Nevertheless, the empirical AUC associated with the predictive index from the regression is 0.5418 with a standard error of 0.0181, as given by the denominator in (4), under the null of independence. Thus, a naive comparison of the t-statistic (0.5418-0.5)/0.0181=2.31 with the usual 5% critical value of 1.645 would lead one to reject $H_0 : \text{AUC} = 1/2$ versus $H_1 : \text{AUC} > 1/2$ very convincingly.

On the other hand, the simulated 5% critical value for the statistic $\sqrt{n}(e\text{AUC}_{\hat{\beta}} - 1/2)$ taking pre-estimation into account is 1.321, which yields a critical AUC value of $0.5 + 1.321/\sqrt{1000} = 0.5418$. The realized value of the empirical AUC is almost exactly the same, indicating that rejecting the null at the 5% level is not a straightforward decision if the estimation effect is to be accounted for. The evidence that the model would classify the outcome better than unbalanced coinflips is not overly strong. ∎

Example 6 The original data set also contains the following two dummies: (i) an indicator showing whether the borrower is a foreign 'guest worker' (*fworker*); and (ii) an indicator showing whether the borrower has a phone line registered under their name (*phone*). The regression of borrower status on the two indicators, over the full sample of observations, is given by:

$$\widehat{status} = 0.6762 + \underset{(0.0768)}{0.2071} \times fworker + \underset{(0.0295)}{0.0400} \times phone$$

The empirical AUC associated with the estimated model is 0.5436, but part of this figure is due to in-sample overfitting. To evaluate classification performance without this effect, one can conduct out-of-sample inference, i.e., partition the full data set into a training sample of, say, size 500 and an evaluation sample of size 500. As demonstrated in Section 5.2, the traditional asymptotic normality result (4) applies to the empirical AUC computed over the evaluation sample. A problem with testing $H_0 : \text{AUC} = 1/2$ this way is that if power is low, the result may depend on the exact sample split used. Indeed, if one conducts a one-sided 5% test based on (4) using 1000 randomly chosen 50-50 sample splits, then H_0 is rejected about 51% of the time. Conducting 2-fold cross validation raises the rejection rate to 66%, but the overall conclusion is still ambiguous.

Our theory allows for the proper handling of the pre-estimation effect without the need to split the sample. In the situation at hand, the simulated 5% critical value for $\sqrt{n}(e\text{AUC}_{\hat{\beta}} - 1/2)$ is equal to 1.201, translating into a $0.5 + 1.201/\sqrt{1000} = 0.5380$ critical value for AUC itself. As the full sample empirical AUC exceeds this value, we conclude that the model classifies better than pure chance (at the 5% significance level). ∎

7 Conclusion

In this paper 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 analytical examples and Monte Carlo simulations to show that if the sample ROC curve is constructed from a model pre-estimated on the same data set, then, under the null: (i) the empirical AUC does not follow the normal limit distribution derived from conventional asymptotic theory; (ii) the traditional test of $H_0 : \text{AUC} = 1/2$ is severely oversized.

We have also stated completely novel analytical results on the asymptotic null distribution of the empirical AUC constructed from a first stage linear regression model estimated over the same sample. These results assume binary regressors. While this is admittedly restrictive, we think that our results are of considerable theoretical interest. They provide a rare analytic characterization of overfitting, and constitute a first step toward a more general theory.

As for the practical relevance of the asymptotic results, we have shown that they can be used to conduct a fully in-sample test of the null hypothesis $H_0 : \text{AUC} = 1/2$, and hence help avoid power losses entailed by out-of-sample evaluation strategies. These results motivate further research into the asymptotic distribution of the empirical AUC with pre-estimation. However, the general problem appears very challenging.

There are several related questions not addressed in this paper. While we characterize the asymptotic distribution of the empirical AUC analytically, critical values still need to be obtained by simulation. Simulation based inference methods can also be applied without exact knowledge of the asymptotic distribution, but care is required in the present setting. In one of the working papers on which this article is based, Hsu and Lieli (2015) show that the standard bootstrap, based on resampling from the empirical *joint* distribution of (X, Y), is inconsistent for the asymptotic null distribution of the empirical AUC with pre-estimation. Nevertheless, if resampling is conducted explicitly under the null, i.e., one resamples from the marginal distribution of the outcome and the joint distribution of the predictors independently, then one can compute valid bootstrap p-values in the standard way (see the Tibshirani, Hall and Wilson (1992) exchange on bootstrap hypothesis tests). Comparing the power properties of fully in-sample analytical tests with the latter bootstrap procedure is a practically relevant exercise that is out of the scope of the present paper.

Finally, one could focus on establishing a connection between the joint significance of the coefficients in the first stage regression and the empirical AUC being significantly different from 1/2. Demler et al. (2011) provide a result of this type under joint normality of the predictors and a parametric estimator of AUC.

A. Appendix: Proofs

A.1 Three useful lemmas

We first collect three lemmas we will use in subsequent proofs.

Consider an ROC curve constructed from a predictive index that can take on a finite number of values. We state a formula for computing the area under such an ROC curve.

Lemma 3 *Let $(F_0, T_0), (F_1, T_1), \ldots, (F_{K-1}, T_{K-1}), (F_K, T_K)$ be a set of points in the unit square $[0,1]\times[0,1]$ such that $0 = F_0 \leq F_1 \leq \ldots \leq F_{K-1} \leq F_K = 1$ and $0 = T_0 \leq T_1 \leq \ldots \leq T_{K-1} \leq T_K = 1$. Construct a curve by connecting these points by a straight line. The area under the curve can be computed as*

$$AUC = \sum_{k=1}^{K} \frac{T_k + T_{k-1}}{2}(F_k - F_{k-1}) = \frac{1}{2} + \frac{1}{2}\sum_{k=1}^{K-1}(F_{k+1}T_k - F_k T_{k+1}).$$

Proof: The first formula is obtained by breaking up AUC into neighboring trapezoids and adding up their area. The second equality is algebra. The formulas stated in Lemma 3 are valid even if some of the points (F_k, T_k) are not distinct. ∎

The next lemma states the joint distribution of the random variables $\hat{P}(s)$, $s \in S$.

Lemma 4 *Let $s_1, \ldots, s_M$ be $M \leq 2^d$ distinct points from S. The random variables*

$$\sqrt{n}[\hat{P}(s_1) - P(s_1)], \ldots, \sqrt{n}[\hat{P}(s_M) - P(s_M)] \tag{13}$$

are asymptotically jointly normal with mean zero and variance-covariance matrix V_P such that the (i,j) element of V_P is given by

$$V_P(k,l) = \begin{cases} P(s_k)[1 - P(s_k)]/\tau & \text{if} \quad k = l \\ -P(s_k)P(s_l)/\tau & \text{if} \quad k \neq l \end{cases}$$

$k, l = 1, \ldots, M$.

Proof: It is easy to show that $\hat{P}(s)$ is asymptotically linear with influence function representation:

$$\sqrt{n}[\hat{P}(s) - P(s)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{1(X_i = s)Y_i}{\tau} - P(s) - \frac{P(s)}{\tau}(Y_i - \tau)\right] + o_p(1)$$

The result then follows from the multivariate central limit theorem for i.i.d. random vectors and direct calculation of the covariance between the influence functions of $\hat{P}(s)$ and $\hat{P}(s')$. ∎

Remark By symmetry, an analogous result holds for the random variables

$$\sqrt{n}[\hat{Q}(s_1) - Q(s_1)], \ldots, \sqrt{n}[\hat{Q}(s_M) - Q(s_M)]; \tag{14}$$

the only difference is that the asymptotic variance V_Q is constructed from $Q(s)$ instead of $P(s)$ and τ is replaced by $1-\tau$. As $\hat{P}(s)$ and $\hat{Q}(s')$ are computed over non-overlapping subsamples, these statistics are independent for any $s, s' \in S$ and sample size n. The asymptotic distribution of all the random variables in (13) and (14) is therefore jointly normal with mean zero and a block-diagonal variance-covariance matrix with the diagonal blocks given by V_P and V_Q.

The last lemma concerns the slope coefficients in a regression model with a binary dependent variable.

Lemma 5 (a) *Let β denote the $d \times 1$ vector of slope coefficients from the linear projection of Y on X and a constant. Then:*

$$\begin{aligned}\beta &= \tau(1-\tau)\Sigma_X^{-1}[E(X \mid Y=1) - E(X \mid Y=0)] \\ &= \tau(1-\tau)\Sigma_X^{-1}\begin{pmatrix}\mathbb{P}(X_1 = 1 \mid Y=1) - \mathbb{P}(X_1 = 1 \mid Y=0) \\ \vdots \\ \mathbb{P}(X_d = 1 \mid Y=1) - \mathbb{P}(X_d = 1 \mid Y=0)\end{pmatrix} \qquad (15)\end{aligned}$$

(b) *Let $\hat{\beta}$ denote the $d \times 1$ vector of slope coefficients from an OLS regression of Y on X and a constant. If X and Y are independent, then $\hat{\beta} = \tilde{\beta} + o_p(n^{-1/2})$, where*

$$\tilde{\beta} = \tau(1-\tau)\Sigma_X^{-1}\begin{pmatrix}\sum_{s \in S:\, s_1 = 1}\hat{P}(s) - \sum_{s \in S:\, s_1 = 1}\hat{Q}(s) \\ \vdots \\ \sum_{s \in S:\, s_d = 1}\hat{P}(s) - \sum_{s \in S:\, s_d = 1}\hat{Q}(s)\end{pmatrix} \qquad (16)$$

with s_j denoting the jth component of s.

Proof: Part (a): The result follows from straightforward manipulations of standard linear projection formulas, taking into account that $Y \in \{0,1\}$. Part (b): $\hat{\beta}$ can be written as the sample analog of equation (15); hence, $\hat{\beta} - \tilde{\beta} = [\hat{\tau}(1-\hat{\tau})\hat{\Sigma}_X^{-1} - \tau(1-\tau)\Sigma_X^{-1}] \times$ [the matrix in equation (16)]. Under independence of X and Y, the matrix in the first term is $o_p(1)$, and the latter is $O_p(n^{-1/2})$. Therefore, $\hat{\beta} - \tilde{\beta} = o_p(1)O_p(n^{-1/2}) = o_p(n^{-1/2})$. ▮

A.2 Proofs of the results stated in the text

Proposition 1 We will apply the multivariate delta method to derive the asymptotic joint distribution of eAUC$_b$ and $\hat{\beta}$ for $b \in \mathbb{R}^d$ with the GI property. We can use Lemma 3 and equation (9) to write AUC$_b$ = $g_0(P_1, \ldots, P_{K-1}, Q_1, \ldots, Q_{K-1})$, where the definition of the function g_0 is stated just before Proposition 1. Clearly, the sample analog area, eAUC$_b$, is given by

$$eAUC_b = g_0(\hat{P}_1, \ldots, \hat{P}_{K-1}, \hat{Q}_1, \ldots, \hat{Q}_{K-1}),$$

where $\hat{P}_k = \hat{P}(s_{(k)})$ and $\hat{Q}_k = \hat{Q}(s_{(k)})$.

Let β denote the $d \times 1$ vector of slope coefficients from the linear projection of Y on X and a constant. By Lemma 5 part (a), we can write $\beta = g_1(P_1, \ldots, P_{K-1}, Q_1, \ldots, Q_{K-1})$, where the definition of the function g_1 is stated just before Proposition 1. We can further write

$$\tilde{\beta} = g_1(\hat{P}_1, \ldots, \hat{P}_{K-1}, \hat{Q}_1, \ldots, \hat{Q}_{K-1}),$$

where $\tilde{\beta}$ is defined in Lemma 5 part (b).

By Lemma 4 and the subsequent remark,

$$\sqrt{n} \begin{pmatrix} \hat{P}_1 - P_1 \\ \vdots \\ \hat{P}_{K-1} - P_{K-1} \\ \hat{Q}_1 - Q_1 \\ \vdots \\ \hat{Q}_{K-1} - Q_{K-1} \end{pmatrix} \rightarrow_d N\left(0_{2(K-1)\times 1}, \begin{pmatrix} V_P & 0 \\ 0 & V_Q \end{pmatrix}\right), \tag{17}$$

where the $(K-1) \times (K-1)$ matrices V_P and V_Q are constructed from the points $s_{(1)}, s_{(2)}, \ldots, s_{(K-1)}$ as in Lemma 4. Let $g = (g_0, g_1)'$, a map from $\mathbb{R}^{2(K-1)}$ to $\mathbb{R}^{1+d}$. We can write

$$\begin{aligned} &\sqrt{n} \begin{pmatrix} eAUC_b - AUC_b \\ \tilde{\beta} - \beta \end{pmatrix} \\ &= \sqrt{n}[g(\hat{P}_1, \ldots, \hat{P}_{K-1}, \hat{Q}_1, \ldots, \hat{Q}_{K-1}) - g(P_1, \ldots, P_{K-1}, Q_1, \ldots, Q_{K-1})]. \end{aligned}$$

Then, by equation (17) and the multivariate delta method (e.g., DasGupta 2008, Thm. 3.7),

$$\sqrt{n} \begin{pmatrix} eAUC_b - AUC_b \\ \tilde{\beta} - \beta \end{pmatrix} \rightarrow_d N\left(0_{(1+d)\times 1}, \nabla g \begin{pmatrix} V_P & 0 \\ 0 & V_Q \end{pmatrix} \nabla g'\right), \tag{18}$$

where ∇g is the $(1+d) \times 2(K-1)$ matrix with rows given by the gradients of the components of g, evaluated at the point $(P_1, \ldots, P_{K-1}, Q_1, \ldots, Q_{K-1})$. Result (11) follows by imposing the null ($\beta = 0$, $\tau V_P = (1-\tau)V_Q = V$), and observing that by Lemma 5(b), $\hat{\beta} = \tilde{\beta} + o_p(n^{-1/2})$ under the null. Note: Result (18) is valid without imposing independence; nevertheless, it is only under the null that $\hat{\beta}$ and $\tilde{\beta}$ are asymptotically equivalent. ∎

Proposition 2 The proof is given in the text; see the paragraph preceding Proposition 2 and the remark following it. ∎

Lemma 2 We will show how to calculate $\nabla g = \binom{\nabla g_0}{\nabla g_1}$ in general.

Fix $k \in \{1, \ldots, K-1\}$. Equation (9) shows that AUC$_b$ depends on P_k only through $T_k, \ldots, T_{K-1}$, so by the chain rule

$$\frac{\partial AUC_b}{\partial P_k} = \frac{\partial AUC_b}{\partial T_k}\frac{\partial T_k}{\partial P_k} + \frac{\partial AUC_b}{\partial T_{k+1}}\frac{\partial T_{k+1}}{\partial P_k} + \ldots + \frac{\partial AUC_b}{\partial T_{K-1}}\frac{\partial T_{K-1}}{\partial P_k},\ k = 1, \ldots, K-1.$$

Equation (9) also shows that $\partial T_j/\partial P_k = 1$ for $j \geq k$ so that

$$\frac{\partial AUC_b}{\partial P_k} = \frac{\partial AUC_b}{\partial T_k} + \frac{\partial AUC_b}{\partial T_{k+1}} + \ldots + \frac{\partial AUC_b}{\partial T_{K-1}},\ k = 1, \ldots, K-1.$$

Using the definition of g_0, it is straightforward to verify that

$$\frac{\partial AUC_b}{\partial T_k} = \frac{1}{2}(F_{k+1} - F_{k-1}),$$

yielding

$$\frac{\partial AUC_b}{\partial P_k} = \frac{1}{2}\sum_{j=k}^{K-1}(F_{j+1} - F_{j-1}), \quad k = 1, \ldots, K-1.$$

A similar argument shows that

$$\frac{\partial AUC_b}{\partial Q_k} = -\frac{1}{2}\sum_{j=k}^{K-1}(T_{j+1} - T_{j-1}), \quad k = 1, \ldots, K-1.$$

Arranging these partial derivatives in a row vector in the appropriate order and substituting equation (9) gives ∇g_0.

Turning to ∇g_1, first observe that $\tau(1-\tau)\Sigma_X^{-1}$ does not depend on $P_1, \ldots, P_{K-1}, Q_1, \ldots, Q_{K-1}$. Therefore, by the linearity of the derivative operator,

$$\nabla g_1 = \tau(1-\tau)\Sigma_X^{-1}\begin{pmatrix} \nabla\left[\sum_{j=1}^{K} P_j s_{(j),1} - \sum_{j=1}^{K} Q_j s_{(j),1}\right] \\ \vdots \\ \nabla\left[\sum_{j=1}^{K} P_j s_{(j),d} - \sum_{j=1}^{K} Q_j s_{(j),d}\right] \end{pmatrix}.$$

Let $k \in \{1, \ldots, K-1\}$. It is clear that, say,

$$\frac{\partial}{\partial P_k}\left[\sum_{j=1}^{K} P_j s_{(j),1} - \sum_{j=1}^{K} Q_j s_{(j),1}\right] = s_{(k),1} - s_{(K),1},$$

because $P_K = 1 - P_1 - \ldots - P_{K-1}$. The rest of the derivatives are computed similarly.

Specializing to the case $d = 2$ $(K = 4)$ gives

$$\nabla g_0 = \frac{1}{2}[1 + Q_2 + Q_3, 1 - Q_1 + Q_3, 1 - Q_1 - Q_2, \\ -(1 + P_2 + P_3), -(1 - P_1 + P_3), -(1 - P_1 - P_2)]$$

and

$$\begin{aligned} \nabla g_1 &= \tau(1-\tau)\Sigma_X^{-1} \\ &\times \begin{pmatrix} s_{(1),1} - s_{(4),1} & s_{(2),1} - s_{(4),1} & s_{(3),1} - s_{(4),1} & s_{(4),1} - s_{(1),1} & s_{(4),1} - s_{(2),1} & s_{(4),1} - s_{(3),1} \\ s_{(1),2} - s_{(4),2} & s_{(2),2} - s_{(4),2} & s_{(3),2} - s_{(4),2} & s_{(4),2} - s_{(1),2} & s_{(4),2} - s_{(2),2} & s_{(4),2} - s_{(3),2} \end{pmatrix}. \end{aligned}$$

Stacking these matrices gives the formula stated in Lemma 2(a).

Specializing to the case $d = 3$ $(K = 8)$, and using the ordering given in part (b) of Lemma 2 gives the formula stated in Appendix B. ▮

Proposition 3 Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ denote the vector of slope coefficients from a linear regression of Y on X_1, X_2, and suppose that $\hat{\beta}_1 > \hat{\beta}_2 > 0$. Clearly, the index $X'\hat{\beta}$ attains its largest value when $X = (1,1)$ so that $s_{(1)} = (1,1)$. The second largest value is attained for $X = (1,0)$ so that $s_{(2)} = (1,0)$. Similarly, $s_{(3)} = (0,1)$ and $s_{(4)} = (0,0)$. In short, the condition $\hat{\beta}_1 > \hat{\beta}_2 > 0$ completely determines the ordering $s_{(1)}, \ldots, s_{(4)}$, and gives Ordering 1 in particular. Therefore, the conditional distribution of $\text{eAUC}_{\hat{\beta}}$ given $\hat{\beta}_1 > \hat{\beta}_2 > 0$ is the same as the conditional distribution of eAUC_b given $\hat{\beta}_1 > \hat{\beta}_2 > 0$ for, say, $b = (2,1)$. We can use Proposition 1 to characterize the latter distribution under the null.

Specifically, define V_1^* as in Proposition 3, let $(A_0, A_1, A_2) \sim N(0, V_1^*)$, and take, say, $b = (2,1)$. Proposition 1 then implies

$$\begin{aligned}
&\mathbb{P}[\sqrt{n}(eAUC_{\hat{\beta}} - 1/2) \leq z \mid \hat{\beta}_1 > \hat{\beta}_2 > 0] \\
&= \mathbb{P}[\sqrt{n}(eAUC_b - 1/2) \leq z \mid \hat{\beta}_1 > \hat{\beta}_2 > 0] \overset{a}{=} \mathbb{P}[A_0 \leq z \mid A_1 > A_2 > 0].
\end{aligned}$$

In addition to the case considered above, call it Case 1, there are potentially seven more, mutually exclusive and exhaustive, cases to consider. The complete list of cases is:

Case 1: $\hat{\beta}_1 > \hat{\beta}_2 > 0$, order $s_{(1)} = (1,1)$, $s_{(2)} = (1,0)$, $s_{(3)} = (0,1)$, $s_{(4)} = (0,0)$;

Case 2: $\hat{\beta}_2 > \hat{\beta}_1 > 0$, order $s_{(1)} = (1,1)$, $s_{(2)} = (0,1)$, $s_{(3)} = (1,0)$, $s_{(4)} = (0,0)$;

Case 3: $\hat{\beta}_1 > 0 > \hat{\beta}_2$, $\hat{\beta}_1 > |\hat{\beta}_2|$, order $s_{(1)} = (1,0)$, $s_{(2)} = (1,1)$, $s_{(3)} = (0,0)$, $s_{(4)} = (0,1)$;

Case 4: $\hat{\beta}_1 > 0 > \hat{\beta}_2$, $\hat{\beta}_1 < |\hat{\beta}_2|$, order $s_{(1)} = (1,0)$, $s_{(2)} = (0,0)$, $s_{(3)} = (1,1)$, $s_{(4)} = (0,1)$;

Case 5: $\hat{\beta}_2 > 0 > \hat{\beta}_1$, $\hat{\beta}_2 > |\hat{\beta}_1|$, order $s_{(1)} = (0,1)$, $s_{(2)} = (1,1)$, $s_{(3)} = (0,0)$, $s_{(4)} = (1,0)$;

Case 6: $\hat{\beta}_2 > 0 > \hat{\beta}_1$, $\hat{\beta}_2 < |\hat{\beta}_1|$, order $s_{(1)} = (0,1)$, $s_{(2)} = (0,0)$, $s_{(3)} = (1,1)$, $s_{(4)} = (1,0)$;

Case 7: $0 > \hat{\beta}_1 > \hat{\beta}_2$, order $s_{(1)} = (0,0)$, $s_{(2)} = (1,0)$, $s_{(3)} = (0,1)$, $s_{(4)} = (1,1)$;

Case 8: $0 > \hat{\beta}_2 > \hat{\beta}_1$, order $s_{(1)} = (0,0)$, $s_{(2)} = (0,1)$, $s_{(3)} = (1,0)$, $s_{(4)} = (1,1)$.

Any other ordering of S is unfeasible in that there is no b with the GI property that generates it.

Note that one needs to distinguish between Case 1 and 2 because X_1 and X_2 are correlated Bernoulli random variables and are not generally exchangeable.[15] Furthermore, one needs to distinguish between, say, Case 3 and Case 4 because the condition $\hat{\beta}_1 > 0 > \hat{\beta}_2$ does not uniquely pin down the ordering of the points in S. Nevertheless, we will argue that it is still sufficient to consider the first four cases; in particular,

Case 1 $\Leftrightarrow$ Case 8, Case 2 $\Leftrightarrow$ Case 7, Case 3 $\Leftrightarrow$ Case 6, Case 4 $\Leftrightarrow$ Case 5,

in the sense that the conditional distribution of $\text{eAUC}_{\hat{\beta}}$ given $\{\hat{\beta} \in \text{Case } i\}$ does not differ across equivalent cases. This will allow us to combine the corresponding orderings in Proposition 2.

[15] Exchanging X_1 and X_2 generally changes the asymptotic joint distribution stated in (11).

To see the stated equivalences, let $Y' = 1 - Y$ and consider replacing Y with Y'. Quantities computed using the transformed data are denoted by a prime superscript. Take, say, Case 1 and Case 8. Clearly, $\hat{\beta} \in$Case 1 iff $\hat{\beta}' \in$Case 8 as $\hat{\beta}' = -\hat{\beta}$. We further observe the following facts:

Fact 1: $eAUC_b'$=$eAUC_{-b}$ for any b. This follows as the decision rules $\hat{Y} = 1(X'b > c)$ and $\hat{Y}' = 1(-X'b > c)$ induce the same ROC curve.

Fact 2: (eAUC$_b$, $\hat{\beta}$)$\overset{a}{\sim}$(eAUC$_b'$, $\hat{\beta}'$) for any fixed b with $b_1 > b_2 > 0$. The asymptotic null distribution given in equation (11) depends on the joint distribution of (Y, X) only through the joint distribution of X and $\tau = \mathbb{P}(Y = 1)$. The relabeling of the outcome interchanges the role of τ and $1 - \tau$ in the definition of V_1^*. It is easy to check that this reversal does not change the matrix V_1^* itself, hence the limit distribution.

By Fact 2, $(eAUC_b' \big| \hat{\beta}' \in \text{Case } 1) \overset{a}{\sim} (eAUC_b \big| \hat{\beta} \in \text{Case } 1)$ for any b with $b_1 > b_2 > 0$. By the relationship between $\hat{\beta}'$ and $\hat{\beta}$ and Fact 1, $(eAUC_{-b} \big| \hat{\beta} \in \text{Case } 8) \overset{a}{\sim} (eAUC_b \big| \hat{\beta} \in \text{Case } 1)$. Equivalently, $(eAUC_{\hat{\beta}} \big| \hat{\beta} \in \text{Case } 8) \overset{a}{\sim} (eAUC_{\hat{\beta}} \big| \hat{\beta} \in \text{Case } 1)$, which is what we wanted to show. It is also clear that Case 1 and 8 have the same probability, because the asymptotic distribution of $\hat{\beta}$ is symmetric about the origin, even if X_1 and X_2 are correlated. The remaining equivalencies can be argued similarly.

We can now combine Cases 1 through 8 using Proposition 2, taking the stated equivalencies into account. This gives the limit distribution stated in Proposition 3. ▮

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] Fahrmeir, L. and A. Hamerle (1984): *Multivariate statistische Verfahren.* Berlin: De Gruyter.

[8] Fahrmeir, L. and G. Tutz (1994): "Multivariate Statistical Modeling Based on Generalized Linear Models."

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

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

[11] Hsu, Y-C. and R.P. Lieli (2015): "Using the Area Under an Estimated ROC Curve to Test the Adequacy of Binary Predictors." Working paper.

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

Lieli, R.P. and H. White (2010): "The Construction of Empirical Credit Scoring Models Based on Maximization Principles." *Journal of Econometrics* 157: 110-119.

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

[14] Schularik, M. and A.M. Taylor (2012): "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870-2008." *American Economic Review* 102: 1029-1061.

[15] Tibshirani, R., P. Hall and S.R. Wilson (1992): "Bootstrap Hypothesis Testing." *Biometrics*, 48, pp. 969-970.