The Null Distribution of the Empirical AUC for Classifiers with Estimated Parameters: a Special Case

Robert P. Lieli* Yu-Chin Hsu†

June 13, 2016

Abstract

We study the distribution of the area under an empirical receiver operating characteristic (ROC) curve constructed from a first stage regression model with parameters estimated on the same data set. We provide a general, but somewhat intrinsic, characterization of the limit distribution of this area, denoted AUC, when the regressors are Bernoulli random variables jointly independent of the outcome. Using the general theory, we further analyze the limit distribution in the two regressor case. It is non-normal and right-skewed. Though the theory applies, explicit expressions for the limit distribution are cumbersome to write down for a larger number of regressors. We provide a trivariate example as further illustration.

Keywords: binary classification, ROC curve, area under the ROC curve, overfitting, hypothesis testing, model selection

*Department of Economics, Central European University, Nador u. 11, Budapest, H-1051. Email: lielir@ceu.edu

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

1 Introduction

There are well-known classic results in the literature describing the asymptotic distribution of the area under a sample ROC curve (Bamber 1975, DeLong et al. 1988). The setting in which these results are derived assumes that the predictor used in constructing the ROC curve is "raw data", i.e., a random variable with a fixed distribution that does not change with the sample size. In many applications there are, however, a number of potential predictors, and it is often desirable to combine them into a single predictive index. It is natural to employ a data-dependent method for this purpose, such as a linear or logit regression model. If the ROC curve is then constructed from the same data that was used to estimate the predictive index, it is no longer obvious that the results cited above continue to apply. Indeed, Demler et al. (2012) and Hsu and Lieli (2015) document the failure of standard inference about AUC in the presence of estimated parameters when one tests nested model specifications against each other or when the null is that the predictors are jointly uninformative about the outcome. The latter hypothesis, which is the null hypothesis considered in this paper, implies that the population ROC curve coincides with the main diagonal of the unit square and AUC is therefore 1/2.

Our contribution is to characterize the null distribution of the empirical AUC in the special case where the predictive index is a linear combination of Bernoulli random variables with weights estimated from the same data set by ordinary least squares (OLS). Under the null of independence, this procedure is asymptotically equivalent to constructing the predictive index by linear discriminant analysis or a logit regression. While our approach, in principle, can be applied to any number of Bernoulli regressors, it will be clear that writing down the asymptotic distribution explicitly becomes very cumbersome as the number of variables increases. We therefore state explicit results completely generally only in the two dimensional case, and examine a special case in three dimensions as further illustration.

Our results make it obvious that the normal limiting distribution presented, for example, in Bamber (1975) cannot be applied to test the null hypothesis that $AUC = 1/2$ when model coefficients are estimated in-sample. Nevertheless, there are ways to get around this inference problem without relying on the results of this paper. First, if the predictive model is

estimated on a training sample, and is tested on an independent evaluation sample, then the validity of "classic" inference is restored. Still, splitting the sample entails a certain amount of power loss, and in smaller samples results might be sensitive to the exact sample split used (Airola et al., 2010, consider various cross-validation approaches to deal with this problem). Second, while we characterize the asymptotic distribution explicitly, critical values still need to be obtained by simulation. However, simulation based inference methods can also be applied without exact knowledge of the asymptotic distribution, but care is required. Of note, Hsu and Lieli (2015) show that the standard bootstrap, based on resampling from the empirical joint distribution of the data, also fails in this context. 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 p-values in the standard way (see the Tibshirani, Hall and Wilson, 1992 exchange on bootstrap hypothesis tests). Finally, one can focus on simply testing the joint significance of the coefficients in the first stage regression, though a one-to-one correspondence with AUC is proven only under joint normality (Demler et al. 2011). In any case, it is still useful to know how first stage estimation affects the asymptotic distribution of the empirical AUC itself.

These practical roundabouts notwithstanding, we think that our results are of considerable theoretical interest. They provide a rare explicit characterization of the consequences of overfitting, i.e., estimating and evaluating a predictor on the same sample, albeit in a (very) special case. The general problem is hard; further partial results are provided by Hsu and Lieli (2015) for continuous regressors.

The rest of the paper is organized as follows. We introduce the formal setup and study its geometry in Section 2. We state the general as well as the specific results in Section 3. Section 4 is a Monte Carlo illustration. Section 5 concludes. Technical details of the proofs are provided in the Appendix.

2 The setup and some geometric arguments

Let X be a $d \times 1$ vector of Bernoulli random variables defined on some probability space equipped with the measure $\mathbb{P}$. Let Σ_X denote the variance-covariance matrix of 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., $S := support(X) = \{0,1\}^d$.[1] In addition, let Y denote another Bernoulli(τ) random variable. We are concerned with predicting the outcome Y based on a linear combination of the covariates X. More specifically, given $b \in \mathbb{R}^d$ we consider index-based prediction (decision) rules of the form

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

where c is a scalar cutoff. As c varies 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 receiver operating characteristic (ROC) curve the in the unit square. The area under this curve is denoted AUC_b (the notation emphasizes the dependence of this quantity on the value of b). Given a random sample of observations $\{(X_i, Y_i)\}_{i=1}^n$, the empirical ROC curve is constructed from the sample analog quantities $(\hat{F}(c), \hat{T}(c))$, where

$$\hat{F}(c) = \sum_{i=1}^n 1(X_i'b > c)(1 - Y_i) \Big/ \sum_{i=1}^n (1 - Y_i) \;\text{ and }\; \hat{T}(c) = \sum_{i=1}^n 1(X_i'b > c)Y_i \Big/ \sum_{i=1}^n Y_i.$$

The area under the empirical ROC curve is denoted eAUC_b.

In practice, the coefficient b is often chosen in a data-dependent way. In particular, let $\hat{\beta}$ denote the vector of slope coefficients from an OLS regression of Y on X and a constant, estimated over the same random sample from which the empirical ROC curve is subsequently constructed. The objective of this paper is to analyze the limit distribution of $\text{eAUC}_{\hat{\beta}}$, the area under the sample ROC curve associated with the decision rule $1(X'\hat{\beta} > c)$, given the null hypothesis that X and Y are independent. Because of the presence of the estimator

[1] 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.

$\hat{\beta}$, standard asymptotic normality results derived from U-statistics theory do not apply. Instead, we will characterize the limit distribution with the help of the following geometric argument.

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 observations on X for which a positive outcome ($Y = 1$) is predicted; hence the '+' superscript. As c varies, the 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 enter $S_b^+(c)$ one at a time. More formally, this means that 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. For brevity, we will say that such values of b possess the *gradual increase* (GI) property. It is intuitively clear, and not hard to show formally, that this property holds for all but a (Lebesgue) measure zero set of points in $\mathbb{R}^d$.[2]

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

$$T(c) = \sum_{s \in S_b^+(c)} P(s) \quad \text{and} \quad F(c) = \sum_{s \in S_b^+(c)} Q(s). \tag{2}$$

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[3], each additional point contributes a positive amount to the sums in (2); 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

[2] A simple counterexample is the point $b = 0$; in this case the set $S_b^+(c)$ jumps from being the empty set to S as c decreases.

[3] This assumption is not essential.

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

$$F_k = Q_1 + \ldots + Q_k \text{ and } T_k = P_1 + \ldots + P_k, \quad k = 0, 1, \ldots, K, \tag{3}$$

where the empty sum is interpreted as zero. By convention, the points (F_k, T_k) are connected by straight line segments to obtain the ROC curve (see, e.g., Fawcett 2004), and AUC_b is computed accordingly (see Lemma 2 in Appendix A.1).

Similarly, 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 $eAUC_b$ depends on b only through the unique ordering it induces on S.

3 Main results

Our first result establishes the joint distribution of $eAUC_b$ and $\hat{\beta}$ under the null. Given $b \in \mathbb{R}^d$ with the GI property, let $s_{(1)}, \ldots, s_{(K)}$ denote the ordering on S induced by b. Define the $(K-1) \times (K-1)$ matrix V as

$$V(k,l) = \begin{cases} \mathbb{P}(X = s_{(k)})[1 - \mathbb{P}(X = s_{(k)})] & \text{if} \quad k = l \\ -\mathbb{P}(X = s_{(k)})\mathbb{P}(X = s_{(l)}) & \text{if} \quad k \neq l \end{cases} \tag{4}$$

$k, l = 1, \ldots, K-1$. 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

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

where $s_{(k),1}$ denotes the first component of $s_{(k)}$, etc., and the dependence of F_k and T_k on $P_1, \ldots, P_{K-1}, Q_1, \ldots, Q_{K-1}$ is as described in equation (3). 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*

$$\sqrt{n}\begin{pmatrix} eAUC_b - 1/2 \\ \hat{\beta} \end{pmatrix} \to_d N\left(0_{(1+d)\times 1}, \nabla g \begin{pmatrix} \frac{1}{\tau}V & 0 \\ 0 & \frac{1}{1-\tau}V \end{pmatrix} \nabla g'\right), \tag{5}$$

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

Remarks

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

2. Proposition 1 'decouples' b from 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. As suggested by the form of the variance matrix, Proposition 1 is obtained by the multivariate delta method. The proof is given in Appendix A.2.

4. Under the null hypothesis $P_k = Q_k = \mathbb{P}(X = s_{(k)})$.

We will now employ Proposition 1 to characterize the asymptotic distribution of $eAUC_{\hat{\beta}}$. Given some enumeration of all the $K!$ permutations of the points in S, let

$$O_\ell = \{b \in \mathbb{R}^d : b \text{ has the GI property and } b \text{ induces the ordering } \ell\}.$$

The sets O_ℓ are mutually exclusive; in fact, many of the O_ℓ are necessarily empty.[4] Clearly, $\hat{\beta} \in O_\ell$ iff $\sqrt{n}\hat{\beta} \in O_\ell$. As $\sqrt{n}\hat{\beta}$ is normally distributed in large samples, and the set of points without the GI property has Lebesgue measure zero in $\mathbb{R}^d$, $\mathbb{P}[\sqrt{n}\hat{\beta} \in \cup_\ell O_\ell] = \mathbb{P}[\hat{\beta} \in \cup_\ell O_\ell] \stackrel{a}{=}$

[4]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.

1, where $\stackrel{a}{=}$ denotes asymptotic equality (i.e., the limit as $n \to \infty$). That is, the events $\{\hat{\beta} \in O_\ell\}$ form a partition of the underlying sample space up to an event with asymptotic probability zero. The following result is then a consequence of the law of total probability.

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

$$\mathbb{P}[\sqrt{n}(eAUC_{\hat{\beta}} - 1/2) \leq z] \stackrel{a}{=} \sum_{O_\ell \neq \emptyset} \mathbb{P}[\sqrt{n}(eAUC_{b_\ell} - 1/2) \leq z \mid \hat{\beta} \in O_\ell] P(\hat{\beta} \in O_\ell), \qquad (6)$$

where b_ℓ is a fixed representative element of O_ℓ for O_ℓ nonempty.

Remarks

1. Proposition 2 makes use of the observation that the distribution of $eAUC_{\hat{\beta}}$ depends on $\hat{\beta}$ only through the ordering it induces on S; see Section 2. Thus, given $\hat{\beta} \in O_\ell$, one can replace $\hat{\beta}$ in $eAUC_{\hat{\beta}}$ with a pre-chosen fixed point $b_\ell \in O_\ell$ without changing the conditional distribution.

2. It is the *relative* size of the components of $\hat{\beta}$ that determine the ordering induced by $\hat{\beta}$; for example, in the two regressor case the ordering is determined by $\hat{\beta}_1/\hat{\beta}_2$. This is the reason why $\hat{\beta}$, while converging to zero under the null, will still induce a unique ordering with (asymptotic) probability one.

3. Terms k and ℓ in sum (6) can be combined if the distribution of $eAUC_{b_k}$ given $\hat{\beta} \in O_k$ is the same as the distribution of $eAUC_{b_\ell}$ 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.

4. An example of sum (6) 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 $eAUC_{b_\ell}$ given $\hat{\beta} \in O_\ell$ is the same for all such ℓ. Therefore, the sum in (6) reduces to $\mathbb{P}[\sqrt{n}(eAUC_{b_\ell} - 1/2) \leq z \mid \hat{\beta} \in O_\ell]$.

As Proposition 1 determines the joint distribution of $\hat{\beta}$ and $eAUC_b$, it also implicitly determines the conditional distribution of $eAUC_{b_\ell}$ given $\hat{\beta} \in O_\ell$ in (6). Thus, Proposition 1 and 2 jointly characterize the asymptotic null distribution of $eAUC_{\hat{\beta}}$. To make this characterization explicit, one needs to (i) obtain an explicit expression for ∇g in equation (5), and (ii) find a set of conditions on the components of $\hat{\beta}$ that uniquely and exhaustively determine the possible orderings of S induced by $\hat{\beta}$. These tasks become increasingly cumbersome as the dimensionality of X increases. We state the general result for $d = 2$ and the perfectly symmetric case for $d = 3$. The following lemma deals with task (i).

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

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

(b) Suppose that $d = 3$. Given the ordering

$$s_{(1)} = (1,1,1), s_{(2)} = (1,1,0), s_{(3)} = (1,0,1), s_{(4)} = (1,0,0)$$
$$s_{(5)} = (0,1,1), s_{(6)} = (0,1,0), s_{(7)} = (0,0,1), s_{(8)} = (0,0,0),$$

the matrix ∇g is given by the formula stated in Appendix B.

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 (4); the matrix ∇g_ℓ as in Lemma 1(a), and the matrix V_ℓ^* as the asymptotic variance matrix in (5). 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,

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

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

Remark 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 $\mathbb{P}\left[A_0 \leq z \middle| A_1 > A_2 > 0\right]$.

The final result treats the perfectly symmetric trivariate case.

Proposition 4 (A trivariate special case) *Let $d = 3$ and suppose that X_1, X_2, X_3 and Y are jointly independent Bernoulli(.5) random variables. For the ordering*

$$s_{(1)} = (1,1,1), s_{(2)} = (1,1,0), s_{(3)} = (1,0,1), s_{(4)} = (1,0,0)$$
$$s_{(5)} = (0,1,1), s_{(6)} = (0,1,0), s_{(7)} = (0,0,1), s_{(8)} = (0,0,0),$$

define the matrix V as in equation (4), ∇g as in Lemma 1(b), and the matrix V^* as the asymptotic variance matrix in (5). Let $(A_0, A_1, A_2, A_3) \sim N(0, V^*)$. Then:

$$\mathbb{P}\left[\sqrt{n}(eAUC_{\hat{\beta}} - 1/2) \leq z\right] \stackrel{a}{=} \mathbb{P}\left[A_0 \leq z \middle| A_1 > A_2 > A_3 > 0, A_1 > A_2 + A_3\right]$$

for any $z \in \mathbb{R}$ *with* V^* *given by*

$$V^* = \begin{pmatrix} 21/64 & 1/2 & 1/4 & 1/8 \\ 1/2 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/8 & 0 & 0 & 1 \end{pmatrix}.$$

4 Illustration by Monte Carlo simulations

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 to predict the outcome Y in-sample, and compute the resulting empirical AUC. The procedure is repeated over 100,000 Monte Carlo iterations to approximate the actual distribution of $\sqrt{n}(eAUC_{\hat{\beta}} - 1/2)$. The theoretical asymptotic distribution is simulated by drawing 10 million observations from the multivariate normal distributions described in Proposition 3, constructing the appropriate conditional distributions, and mixing them together. Table 1 displays various quantiles of the simulated asymptotic distributions in bold, and Figure 1 shows a kernel density estimate of the p.d.f. for DGP 1.

Table 1: 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.

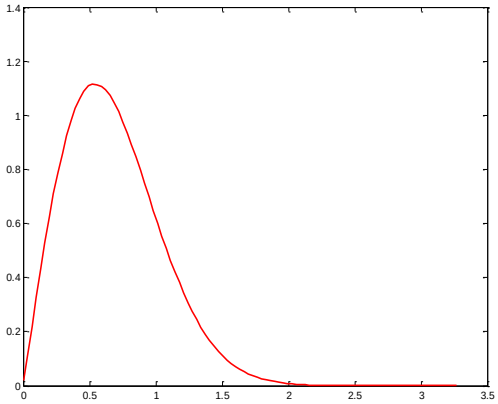

Figure 1: The asymptotic distribution of $\sqrt{n}(eAUC_{\hat{\beta}} - 1/2)$ for DGP 1

As seen in Figure 1, 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 overfitting. Table 1 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 these simulations do not have powers of validation comparable to an analytic proof, they do suggest that the result stated in Proposition 3 is sound.

5 Conclusion

We have stated analytical results on the asymptotic distribution of the empirical AUC constructed from a linear regression model estimated over the same sample. The results assume binary regressors. Under the null hypothesis that these variables are independent of the outcome to be predicted, in-sample overfitting causes the limit distribution to deviate from normality. Thus, naive applications of asymptotic normality results derived from U-statistics theory to test $H_0 : AUC = 1/2$ are flawed. While the setup considered here is admittedly rather special, our results still contribute to a better understanding of in-sample overfitting and constitute a first step toward a more general theory.

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 2 *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 2 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 3 *Let* $s_1, \ldots, s_M$ *be* $M \leq 2^d$ *distinct points from* S. *The random variables*

$$\sqrt{n}[\hat{P}(s_1) - P(s_1)], \ldots, \sqrt{n}[\hat{P}(s_M) - P(s_M)] \tag{7}$$

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

$$\sqrt{n}[\hat{Q}(s_1) - Q(s_1)], \ldots, \sqrt{n}[\hat{Q}(s_M) - Q(s_M)]; \tag{8}$$

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 (7) and (8) 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 4 *(a) Let β denote the $d \times 1$ vector of slope coefficients from the linear projection of Y on X and a constant. Then:*

$$\begin{aligned} \beta &= \tau(1-\tau)\Sigma_X^{-1}[E(X \mid Y=1) - E(X \mid Y=0)] \\ &= \tau(1-\tau)\Sigma_X^{-1} \begin{pmatrix} \mathbb{P}(X_1=1 \mid Y=1) - \mathbb{P}(X_1=1 \mid Y=0) \\ \vdots \\ \mathbb{P}(X_d=1 \mid Y=1) - \mathbb{P}(X_d=1 \mid Y=0) \end{pmatrix} \end{aligned} \tag{9}$$

(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

$$\tilde{\beta} = \tau(1-\tau)\Sigma_X^{-1} \begin{pmatrix} \sum_{s \in S:\, s_1=1} \hat{P}(s) - \sum_{s \in S:\, s_1=1} \hat{Q}(s) \\ \vdots \\ \sum_{s \in S:\, s_d=1} \hat{P}(s) - \sum_{s \in S:\, s_d=1} \hat{Q}(s) \end{pmatrix} \tag{10}$$

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 (9); hence, $\hat{\beta} - \tilde{\beta} = [\hat{\tau}(1-\hat{\tau})\hat{\Sigma}_X^{-1} - \tau(1-\tau)\Sigma_X^{-1}] \times$ [the difference between the two matrices in equations (9) and (10)]. Under independence of X and Y, the former matrix 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 2 and equation (3) to write $\text{AUC}_b = g_0(P_1, \ldots, P_{K-1}, Q_1, \ldots, Q_{K-1})$, 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 4 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 4 part (b).

By Lemma 3 and the subsequent remark,

$$\sqrt{n}\begin{pmatrix} \hat{P}_1 - P_1 \\ \vdots \\ \hat{P}_{K-1} - P_{K-1} \\ \hat{Q}_1 - Q_1 \\ \vdots \\ \hat{Q}_{K-1} - Q_{K-1} \end{pmatrix} \to_d N\left(0_{2(K-1)\times 1}, \begin{pmatrix} V_P & 0 \\ 0 & V_Q \end{pmatrix}\right), \tag{11}$$

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 3. 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 (11) and the multivariate delta method (e.g., DasGupta 2008, Thm. 3.7),

$$\sqrt{n}\begin{pmatrix} eAUC_b - AUC_b \\ \tilde{\beta} - \beta \end{pmatrix} \to_d N\left(0_{(1+d)\times 1}, \nabla g \begin{pmatrix} V_P & 0 \\ 0 & V_Q \end{pmatrix} \nabla g'\right), \tag{12}$$

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 (5) follows by imposing the null ($\beta = 0$, $\tau V_P = (1-\tau)V_Q = V$), and observing that by Lemma 4(b), $\hat{\beta} = \tilde{\beta} + o_p(n^{-1/2})$ under the null. Note: Result (12) 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 1 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 (3) 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 (3) 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 (3) 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 1(a).

Specializing to the case $d = 3$ ($K = 8$), and using the ordering given in part (b) of Lemma 1 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

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

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.[5] 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,

$$\text{Case 1} \Leftrightarrow \text{Case 8}, \quad \text{Case 2} \Leftrightarrow \text{Case 7}, \quad \text{Case 3} \Leftrightarrow \text{Case 6}, \quad \text{Case 4} \Leftrightarrow \text{Case 5},$$

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.

[5] Exchanging X_1 and X_2 generally changes the asymptotic joint distribution stated in (5).

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: $(\text{eAUC}_b,\ \hat{\beta}) \overset{a}{\sim} (\text{eAUC}'_b,\ \hat{\beta}')$ for any fixed b with $b_1 > b_2 > 0$. The asymptotic null distribution given in equation (5) 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. ∎

Proposition 4 Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ denote the vector of slope coefficients from a linear regression of Y on X_1, X_2, X_3 and a constant, and suppose that

$$\hat{\beta}_1 > \hat{\beta}_2 > \hat{\beta}_3 > 0 \text{ and } \hat{\beta}_1 > \hat{\beta}_2 + \hat{\beta}_3.$$

It is easy to check that these conditions uniquely induce the ordering $s_{(1)} = (1,1,1)$, $s_{(2)} = (1,1,0)$, $s_{(3)} = (1,0,1)$, $s_{(4)} = (1,0,0)$, $s_{(5)} = (0,1,1)$, $s_{(6)} = (0,1,0)$, $s_{(7)} = (0,0,1)$, $s_{(8)} = (0,0,0)$, which is the ordering stated in Proposition 4. The formula for the asymptotic distribution now follows from Proposition 1 and remark 4 after Proposition 2. ∎

B. Appendix: The formula for ∇g in the symmetric trivariate case

We define the 4×7 matrices G_Q and G_P the following way. Let

$$G_Q = \begin{pmatrix} 1+\sum_{i=1}^{7} Q_i & 1-Q_1+\sum_{i=3}^{7} Q_i & 1-Q_1-Q_2+\sum_{i=4}^{7} Q_i & 1-(\sum_{i=1}^{3} Q_i)+\sum_{i=5}^{7} Q_i & 1-(\sum_{i=1}^{4} Q_i)+Q_6+Q_7 & 1-(\sum_{i=1}^{5} Q_i)+Q_7 & 1-\sum_{i=1}^{7} Q_i \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

and

$$G_P = \begin{pmatrix} 1+\sum_{i=1}^{7} P_i & 1-P_1+\sum_{i=3}^{7} P_i & 1-P_1-P_2+\sum_{i=4}^{7} P_i & 1-(\sum_{i=1}^{3} P_i)+\sum_{i=5}^{7} P_i & 1-(\sum_{i=1}^{4} P_i)+P_6+P_7 & 1-(\sum_{i=1}^{5} P_i)+P_7 & 1-\sum_{i=1}^{7} P_i \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Then:

$$\nabla g = \begin{pmatrix} 1/2 & 0_{(1\times 3)} \\ 0_{(3\times 1)} & \tau(1-\tau)\Sigma_X^{-1} \end{pmatrix} (G_Q \mid -G_P).$$

References

[1] Airola, A., T. Pahikkala, W. Waegeman, B. De Baets, T. Salakoski (2010): "A comparison of AUC estimators in small-sample studies." *Journal of Machine Learning Research: Workshop and Conference Proceedings* 8: 3-13.

[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] DasGupta, A. (2008): *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer.

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

[8] 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.

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