Nonparametric estimation of the auc of an estimated index

Abstract. We describe a nonparametric method of estimating the AUC of an index $\beta' x$ when β is estimaed from the same data, with a focus on nonparametric estimation of the difference of the AUCs of two distinct indices.

1 Introduction

measuring the discrimination of biomarkers using the auc. ΔAUC to measure the difference in discrimination between two markers. often one is based on a subset of covariates of the other marker.

demler 2017 ΔAUC is one of the most widely used measures of discrimination.((should be “difference” in discrimination I think?))

2 Background/setting

Let

$$X_{01}, \ldots, X_{0m} \sim F, IID, X_{11}, \ldots, X_{1n} \sim G, IID$$

Based on this data, analyst obtains coefficients β, γ. May be that one is based on a subset of those coefficients on which the other is based.

The statistic $\Delta \hat{\text{AUC}}$ is

$$\frac{1}{mn}\sum_{i,j}\{\hat{\beta}' X_{0i} < \hat{\beta}' X_{1j}\} - \frac{1}{mn}\sum_{i,j}\{\hat{\gamma}' X_{0i} < \hat{\gamma}' X_{1j}\}$$

An explicit probabilty model may not be specified, and often the two estimation methods imply inconsistent models, e.g., logistic models with nonzero covariates omitted from the reduced model. ((this last example only an issue under the alternative)) Nevertheless inference is sought, particularly 1. whether the difference in the AUCs of the two markers $\hat{\beta}' x$ and $\hat{\gamma}' x$, in some limiting sense ((introduce starred parameters here?)) is nonzero, and if so 2. the magnitude of the difference.

For fixed $\hat{\beta}, \hat{\gamma}$, this statistic may be viewed as a two-sample U-statistic with kernel $(x, y) \mapsto \{\hat{\beta}' x < \hat{\beta}' y\} - \{\hat{\gamma}' x < \hat{\gamma}' y\}$.

Two complications for the basic u-stat theory.

1. under the null of no diff, asy distribution is a weighted combination of chi-squares, weights hard tocompute etc. pepe 2013 helps to resolve the problem of testing the index aucs for significant difference. Let the risk function based on a set of covariates X_0 be denoted $\rho_{X_0}(\cdot) = P(D = 1 \mid \cdot)$. Let $X, X', ...$ They show that the null of equal AUCs of ...,

$$P(\rho_{...}(X_0, X_0') < \rho(X_1, X_1')) = P(\rho_{...}(X_0) < \rho(X_1))$$

holds if and only if the risk functions are equal, $\rho_{X_0,X_0'} = \rho_{X_0}$. Often the estimation procedure is of secondary importance and the goal of testing the null $\Delta\text{AUC} = 0$ is to test the the additional covariates improve discrimination. In this case, the test may be based on the risks instead. Even if an analyst is interested specifically in testing for the difference in AUCs where $\hat{\beta}, \hat{\gamma}$ are obtaiend through a particular procedure, e.g., logistic coefficients. there will often be a monotone link connecting the index to the risk, e.g., the expit function. Since the AUC is invariant to monotone transformations, the risk may still be used to test for a difference. The two models, reduced and full, are compatible under the null, since then $\beta^* = \gamma^*$, provided the coefficient estimation model is correct.

A drawback to this approach is it requires knowing the true risk function. one would be able to obtain the indices β^T and γ^T and compare the discrimination, and use the indices in practice, without knowing the correct risk function. However, unless computing the null distribution of the ΔAUC calls for fewer modeling assumptions, improved efficiency, or something else to recommend it, may as well test risk functions.

literature: heller gives the asy null distribution specifically for betahat estimated by mrc method. [[update: heller doesnt make modeling assumptions on the covariates, only on the estimation of betahat; maybe that is a benefit over testing risk function? ie is there a test for the mrc coefficients that doesnt require modeling covariates?]] demler for lda with gaussian covariates ((no benefit over pepe approach here since parametric assumptions imposed)). recently noted ((cite)) that asy null distribution remains intractable for common estimation methods eg logistic regression.

therefore only consider the alternative here.

2. second issue is that betahat is estimated from the data so that the observations to which the u-statistic is applied are not iid. Non-degenerate u-stat with estimated parameters is typically still normal, though demonstration requires viewing the ustat as a process indexed by random index beta. estimation of the parameter in general affects the asy distribution. address this issue in the remainder‘.

3 Theory

3.1 IID sum

the approach, which we adopt, is to rewrite more complicated estimators as asymptotically equivalent iid sums, amenable to conventional analysis. benefits of approximating by an iid sum: 1. can de-couple the two parts of the difference and just focus on the estimating the auc of a an index estimated (possibly misspeciifed model) from the data. 2. $\hat{\theta}$ is an IID sum, and the sd estimate is the empirical estimator. This is itself an estimate of $\sqrt{n}\hat{\theta}$, not $\hat{\theta}$. So etimated parameters are OK (take taylor expansion) as long as the parameter estimates are consistent and dependence is continuous. of course may affect efficiency of asy convergence. 3. produce routines to reduce auc(beta.hat) to an iid sum, whatever the data is. can apply this separately to the full and reduced data sets, but also applies more generally to a comparison of any correlated aucs with parameters estimated from the data, eg lda versus logistic.

For CDFs F, G on $\mathbb{R}^p$ and vector β define the notation for the AUC of the index $P(\beta' X < \beta' Y), X \sim F, Y \sim G, X \perp\!\!\!\perp Y$,

$$\theta(F, G, \beta) = \int \{\beta' x < \beta' y\} dF(x) dG(y).$$

In this notation, $\Delta\hat{\text{AUC}} = \theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(\hat{F}, \hat{G}, \hat{\gamma})$. We write each term as an IID sum, and later take the difference to represent ΔAUC as an IID sum. Decompose the centered AUC $\theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(F, G, \beta)$ as a sum of two terms, reflecting the two sources of estimation

$$\begin{aligned} & \theta(\hat{F}, \hat{G}, \hat{\beta}) - \theta(F, G, \beta) \\ & = \theta(F + \delta F, G + \delta G, \beta + \delta\beta) - \theta(F, G, \beta + \delta\beta) && (1) \\ & + \theta(F, G, \beta + \delta\beta) - \theta(F, G, \beta) && (2) \end{aligned}$$

Where $\delta F = \hat{F} - F$, etc.

term (1): The function $\theta(\cdot, \cdot, \beta)$ is bilinear,

$$\begin{aligned} & \theta(F + \delta F, G + \delta G, \beta + \delta\beta) - \theta(F, G, \beta + \delta\beta) \\ & = \theta(\delta F, G, \beta + \delta\beta) + \theta(F, \delta G, \beta + \delta\beta) + \theta(\delta F, \delta G, \beta + \delta\beta) \\ & = -\frac{1}{m}\sum_{i=1}^{m}(1 - G(\hat{\beta}' X_{0i}) - \theta(F, G, \hat{\beta})) + \frac{1}{n}\sum_{i=1}^{n}(F(\hat{\beta}' X_{1i}) - \theta(F, G, \hat{\beta})) + o(n^{-1/2}) \end{aligned}$$

The approximation in the third inequality,$\theta(\delta F, \delta G, \beta + \delta\beta) = o(n^{-1/2})$: $P(\sqrt{n}\theta(\delta F, \delta G, \beta + \delta\beta) > \epsilon) \leq P(n^{1/4} \int d|F_n - F|(x) > \sqrt{\epsilon}) + P(n^{1/4} \int d|G_n - G|(x) > \sqrt{\epsilon})$ ...((in case this approach doesnt work, can just cite nolan–pollard approach))

For fixed $\hat{\beta}$, the sums in [ref] are the Hoeffding decomposition of ΔAUC. Same as the first von Mises derivative. For fixed $\hat{\beta}$, represents term (1) as an IID sum to which the CLT may be applied to get the asymptotic distribution of ΔAUC if term (2) were negligible, e.g., if $\hat{\beta}$ were not estimated. The Delong approach in this situation is to estimate F, G using the empirical CDFs $\hat{F}, \hat{G}$, giving rise to the standard Delong statistic for inference on ΔAUC.

term (2): Assume $\sqrt{n}(\hat{\beta} - \beta_0) \to 0$ in probability, $\beta \mapsto \theta(F, G, \beta)$ is differentiable at β_0. Let the function ϕ represent the estimator $\hat{\beta}$ as an IID sum

$$\hat{\beta} - \beta_0 = \sum_{i=1}^{m+n} \phi(X_i) + o(n^{-1/2})$$

i.e., ϕ is an influence function for the $\hat{\beta}$. Then (2) is

$$\begin{aligned}
&\theta(F,G,\beta+\delta\beta)-\theta(F,G,\beta)\\
&=(\hat{\beta}-\beta_0)\frac{\partial}{\partial\beta}\theta(F,G,\beta)+o_P(n^{-1/2})\\
&=\frac{\partial}{\partial\beta}\theta(F,G,\beta)\sum_{i=1}^{m+n}\phi(X_i)+o_P(n^{-1/2}).
\end{aligned}$$

It isn't needed that $\hat{\beta}$ be estimated by a correctly specified model, only that it has some probability limit at the $\sqrt{n}$ rate. The influence function may contain nuisance parameters as long as it depends on them continuously and consistent estimators are available. Though procedure for obtaining the estimate $\hat{\beta}$ and the associated influence function ϕ often involve some parametric assumptions, we still term the procedure described her as non parametric since the model for $\hat{\beta}$ may be misspecified, and the whatever the estimation procedure is it will be known to the analyst, so that an influence function may be chosen, if one exists.

What goes wrong under the null? If the probability limit of $\hat{\beta}$ and $\hat{\gamma}$ are the same, then in term (2) the derivatives are the same. if well specified may be a transofmration of the risk and therfore stationary point, derivatives will both vanish and (2) will be $o(n^{-1/2})$. What about term (1)?

adding the two parts term-wise gives an iid representation of $\theta(\hat{\beta})$:

take the difference with the same representation of another estimator, $\theta(\hat{\gamma})$, to obtain an iid representation of $\Delta\hat{\text{A}}\text{UC}$:

4 Examples

4.1 No effect of beta estimation

In the ordinary course, betahat estimation can be ignored in computing the index of an auc iff term 2 is 0. in the case of ΔAUC = ..., need deriv must ordinarily be 0 at both betahat and gammahat, in which case the usual delong statistic may be used.

term 2 will be 0 in many well-specified models for beta estimation due if the index is monotonically related to the risk function.

proposition:

Given RVs (X_0, D), D binary etc.

1. The roc curve is maximized pointwise over all real functions of X_0 by the likelihood ratio, equivalently, the risk of D based on X_0 ((define this phrase))

2. The auc of a real function of X_0 is maximal iff the function is an increasing function of the risk [at least if distributions assumed continuous] i.e., there is a strictly increasing function $f:\mathbb{R}\to\mathbb{R}$ such that $P(risk(x)<x_0|D=i)=P(f(\beta' x)|D=i)$ for all $x\in\mathbb{R}$ and $i=1,2$.

3. Assuming as above the derivative of the AUC is smooth at β_0, can use delong statistic if index is related to risk by an increasing function.[more precisely: index has the same conditoinal distributions given binary status as an increasing function of the risk ((at least if distributions are continuous))

Proof. 1. is neyman pearson lemma, as pointed out by ((pepe? check if she did it first. swets.)). viewing D as a parameter, the most powerful test of the null $D = 0$ versus the simple alternative $D = 1$ rejects for large values of the likelihood ratio of ((x,g)). Therefore the ROC curve of the likelihood ratio is maximal at each point. Since the ROC curve is the same for incrasing functions of the likelihood, and ((show likelihood is expit of risk)), the same holds of the risk. 2. Though markers not related by increasing function may have the same auc, however since the roc curve of the risk is maximal, an index with the same auc must have the same roc curve, which does imply the index is distributionally equal to an increasing function of the risk.

Example. A prominent example where the index is an increasing function of the risk is the index model for a binary response:

$$P(D = 1) = h(\beta' x), \beta \in \mathbb{R}^p$$

The function h is strictly increasing, such as a probit link, logistic link, identity, etc. In the $p = 1$ case the β cancels and the requirement is simply that the risk be increasing in the sole covariate, i.e., that the covariate or its negation be a risk factor.

However, the ΔAUC consists of 2 aucs, so to apply this example as justification for inference based on the standard Delong statistic requires that both AUC models be well-specified. In the case of comparing a correctly specified full model $P(D = 1) = ...$ to a reduced model ..., the requirement is that the marginalization does not break the model:

$$\int \cdots$$

((maybe cite bridge distribution paper.)) Some examples where this requirement holds are:

1. probit regession with gaussian covariates.

2. fisher lda (homoskedastic with gaussian covariates) (just with gaussian data? conjecture in demler 2012 re elliptic distributions. maybe expand on this example in the mis-specified lda section.) ((lda is collapsible more generally, but risk may not be an increasing function of the index without the gaussian assumption))

((maybe give proof here, or can mix it in with longer example below))

This example was given by Demler 2011 (or 2012?)

3. the logistic view of the last example: logistic regression with a gaussian mixture as covariates in the special case that beta is the lda/malanobis dist beta. This example is almost the same as the homoskedastic LDA example, since [[ref fisher lda display]] the posterior probabilities are given by ... .

4. lda (heteroskedastic) with independent exponential family data mean parameterized

Proof. Let $X_i \mid D = j$ have density $h_i(x_i) \exp(\theta'_{ij} t_i(x_i) - A_{ij})$. If the covariates are independent, the likelihood ratio is then

$$\frac{f(x \mid D = 1)}{f(x \mid D = 0)} \sim \sum_{i=1}^{n} (\theta_{i1} - \theta_{i0})^t x_i$$

((notation: n, also x's.))

If LDA is used to estimate β, then

$$\hat{\beta} \to_p ...$$

and the index at probability limit is .... If 1. the covariates have the same population variances, say $\pi_0 A_0'' + \pi_1 A_1''$, and 2. the parameter θ_{ij} is the mean A_{ij}, then

$$\beta x \sim \sum_{i=1}^{n} (A_{i1}' - A_{i0}') x_i \sim \rho(x)$$

With gaussian data as in 2 but heteroskedastic, or heteroskedastic exponential family as in 2 but not independent, the derivative of θ^* need not be 0. [[ref heteroskedastic lda example below.]]

4.2 must account for beta estimation

Next we describe situations where must account for the β parameter estimation include: misspecification in one of the above situations ((can view ths proposed approach as adding robustness)). or, correct specified but nonzero derivative still.

Example 1. heteroskedastic gaussian lda. Suppose Gaussian linear discriminant analysis is applied to estimate beta but possibly misspecified in that the two classes may not have the same covariance. The model is

$$\begin{aligned} X|D = i &\sim F_i = N_p(\mu_i, \Sigma_i) \\ P(D = 1) &= 1 - P(D = 0) = \pi_1 \end{aligned}$$

The LDA parameter estimation procedure is to base class membership on the sign of $\beta' x$ ((ignore intercepts without loss)), where

$$\begin{aligned} \hat{\beta} &= ... \\ \hat{\mu} &= ... \\ \hat{\Sigma} &= ... \end{aligned}$$

the LDA parameter estimates ((ref above)), which assume a common variance for the two classes, tend in probability to

$$\begin{aligned} \beta^* &= ... \\ \Sigma^* &= ... \end{aligned}$$

The AUC and its derivative at the starred parameters are

$$\begin{aligned} \theta(F, G, \beta^*) &= ... \\ \theta'(F, G, \beta^*) &= ... \end{aligned}$$

The derivative is 0 iff ((eigenvector condition)). In terms of the normal means and variances, this condition is ((...)). 2 examples: 1. independent data, ((...)), 2. proportional covariance matrices. The first is already implied by the general exponential family result [[ref above]] but not the second as the observations are not independent.

show that derivative term can be unbounded. When $\Sigma \approx \sigma^*$ [[not defined yet $\Sigma = \Sigma_0 + \Sigma_1$]], the the influence function is approximately $O(|\Sigma|^{-1/2})$, the root of the inverse Fisher information, and $\theta'(F, G, \beta^*)$ is approximately $O(|\Sigma|^{1/2})$, so that the product, giving the entire non-Delong term, is approximately $O(1)$. Nevertheless it is possible to push the proportion so that the entire non-Delong term $(\Sigma^*)^{-1}\theta'(F, G, \beta^*)$ has large components.

Let

$$\Sigma_0 = ...$$

Then $(\Sigma^*)^{-1}\theta'(F, G, \beta^*) \to \pm\infty\mathbb{1}$ ((write out?)) as $\pi_0 \to 1$ and $\epsilon \to 0$ simultaneously. One therefore expects that under this scenario inference based on the Delong estimator will be faulty, as verified by simulation in ((ref simulation section)). ((possible to give result so that the sign of the derivative term is controlled, so that delong can be forced either to low fpr or low power?))

5 Simulation

6 Discussion

approach extends straightforwardly to other differentiable functions of the data $f(x)$, not just the linear combination.

extends to discrete covariates (modified auc kernel)