Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2021, Vol. 30(7) 1654–1666

DOI: 10.1177/09622802211022377
journals.sagepub.com/home/smm
SAGE

Nested logistic regression models and ΔAUC applications: Change-point analysis

Chun Yin Lee

Abstract

The area under the receiver operating characteristic curve (AUC) is one of the most popular measures for evaluating the performance of a predictive model. In nested models, the change in AUC (ΔAUC) can be a discriminatory measure of whether the newly added predictors provide significant improvement in terms of predictive accuracy. Recently, several authors have shown rigorously that ΔAUC can be degenerate and its asymptotic distribution is no longer normal when the reduced model is true, but it could be the distribution of a linear combination of some χ^2_1 random variables [1,2]. Hence, the normality assumption and existing variance estimate cannot be applied directly for developing a statistical test under the nested models. In this paper, we first provide a brief review on the use of ΔAUC for comparing nested logistic models and the difficulty of retrieving the reference distribution behind. Then, we present a special case of the nested logistic regression models that the newly added predictor to the reduced model contains a change-point in its effects. A new test statistic based on ΔAUC is proposed in this setting. A simple resampling scheme is proposed to approximate the critical values for the test statistic. The inference of the change-point parameter is done via *m*-out-of-*n* bootstrap. Large-scale simulation is conducted to evaluate the finite-sample performance of the ΔAUC test for the change-point model. The proposed method is applied to two real-life datasets for illustration.

Keywords

Area under the receiver operating characteristic curve, change-points, discriminatory measures, *m*-out-of-*n* bootstrap, nested models

1 Introduction

In medical and epidemiological studies, the receiver operating characteristic (ROC) curve serves as a general tool to visualize how well a continuous explanatory variable can predict a binary response. The area under a ROC curve (AUC), on the other hand, is a scalar measure of model discriminatory accuracy which can be easily comprehended by practitioners. In the literature, ΔAUC, the difference between two AUCs, has been frequently used to test for the association between a certain biomarker and the binary response over another biomarker, based on the Mann–Whitney statistic.[1]This is also referred to as "head-to-head" comparison of the two underlying nonnested models.[2] In particular, the DeLong[3] test has become a widely-adopted approach where the test statistic ΔAUC has an asymptotic zero-mean normal distribution under the null hypothesis of no difference in predictive accuracy between the two biomarkers, and the variance estimate has an explicit formulation.

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

Corresponding author:
Chun Yin Lee, Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong.
Email: james-chun-yin.lee@polyu.edu.hk

Recently, extra attention has been paid to implement ΔAUC test to compare a reduced regression model with the full regression model based on the same dataset, where the highly correlated composite risk scores of the two estimated models are used to compute the test statistic instead of the covariate values. It is well known that the DeLong test cannot be directly applied under this kind of nested model setting,[2,4–6] or the test will result in an extremely conservative size under the null hypothesis, and low power if the signal is not strong enough. This is due to the fact that the test statistic based on the nested model degenerates, and it no longer follows a normal distribution under the null hypothesis. Hence, the distribution cannot be fully characterized by the variance parameter. However, the test statistic is still found to be normal with simple formula for the variance estimate under the alternative hypothesis. Several remedies have been proposed in the literature. For instance, Seshan et al.[2] proposed to use a projection-permutation approach in which the newly added predictor variable is decomposed orthogonally in order to simulate the correct reference distribution for the test statistic. Later, Demler et al.[5] proposed a method of injecting a random noise to the reduced and full models, which can help resolving the degeneracy issue and shifting the underlying distribution toward normal, although the power of the ΔAUC test is compromised in this case. Given the complexity of the test statistic under the null hypothesis, a more direct approach, as if it is treated as a proxy for the gold standard in the literature, to achieve the correct reference distribution of the test statistic is via resampling.

Logistic regression models with change-point in the covariate effects can also be considered as a special case of the nested model. Change-point models are particularly useful in clinical studies when the effect of the covariates on the response cannot be assumed to be linear, but it can be thought of being driven by one or more change-points. Change-point models provide favorable flexibility in exploring the nonlinear association between the covariate and response, while it is less susceptible to overfitting issues as compared to modeling with polynomial splines. Moreover, the statistical inference for the change-point parameter is of clinical importance, but it is treated as a nuisance component in nonparameteric smoothing methods. In this paper, we study the logistic regression model where a covariate exhibits its effect on the response only when the covariate value exceeds a certain change-point, which is also known as "threshold" or "threshold limit values" in the literature. This model has broad applicability in medical and epidemiological studies. For demonstration, two medical datasets are explored. In the first example, we study the nonlinear association between the HIV-1 infection rate of infants and the immune response biomarkers of the mother.[7] When the immune response is too low, usually no protective effects would be contributed to lowering the risks of developing HIV-1 disease. Presumably, the immune response is effective, through the indication of a certain unknown threshold value, only when it starts to have a significant impact on the disease outcome. In the second example, we study the nonlinear association between the incidence of chronic bronchitis and average dust concentration in the workplace.[8,9] When the dust concentration is low, it has usually a negligible effect on the risk of incidence. However, beyond a critical level of dust concentration, the risk is assumed to increase with the dust concentration.

Existing methods for testing the presence of change-points in logistic regression are typically based on the maximal score or maximal likelihood ratio (LR) tests.[7,8,10,11] However, to the best of the author's knowledge, no previous work has considered ΔAUC as the test statistic for the presence of change-points. Hence, this paper fills the gaps in developing the ΔAUC test for change-point detection based on the binary regression model, and the method where critical values for the proposed test statistic can be approximated under the appropriate reference distribution.

We structure the paper as follows. In Section 2, we first provide a brief review on the ΔAUC test for ordinary nested logistic regression model. Then, we provide details for the model specification, the proposed test statistic, the resampling approach for hypothesis testing, and the confidence interval estimation of the change-point parameter. In Section 3, the finite-sample performance of the proposed method is studied based on various parameter settings. The proposed method is applied to the two medical datasets in Section 4. Lastly, some concluding remarks are provided in Section 5.

2 The ΔAUC test

2.1 Nested logistic regression models: A review

Let Y_i be the binary medical outcome of the ith individual in a random sample of size n, $i = 1, 2, \ldots, n$. As is customary, we use $n_0 = \sum_{i=1}^{n} I(Y_i = 0)$ and $n_1 = \sum_{i=1}^{n} I(Y_i = 1)$ to denote the total number of nonevents and events, respectively, where $I(.)$ is the usual indicator function. In the presence of a set of explanatory variables, say $\boldsymbol{W}_i = (W_{1i}, \ldots, W_{ip})$, it is often of interest to know whether an additional set of variables, called

$\boldsymbol{V}_i = (V_{1i}, \ldots, V_{iq})$, could significantly increase the predictive accuracy of the original model. That is, for a particular dataset, to compare the model

$$\eta_i = P(Y_i = 1|\boldsymbol{W}_i) = \frac{\exp\{\gamma + \boldsymbol{\beta}^T\boldsymbol{W}_i\}}{1 + \exp\{\gamma + \boldsymbol{\beta}^T\boldsymbol{W}_i\}} \tag{1}$$

over

$$\eta_i^* = P(Y_i = 1|\boldsymbol{W}_i, \boldsymbol{V}_i) = \frac{\exp\{\gamma + \boldsymbol{\beta}^T\boldsymbol{W}_i + \boldsymbol{\alpha}^T\boldsymbol{V}_i\}}{1 + \exp\{\gamma + \boldsymbol{\beta}^T\boldsymbol{W}_i + \boldsymbol{\alpha}^T\boldsymbol{V}_i\}}, \tag{2}$$

where γ is a scalar parameter, and $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are p- and q-dimensional vectors of regression parameters, respectively. Here, one can regard η_i and η_i^* as the event probabilities of the ith individual derived from a transformation of the risk scores, the linear combination of the covariates, in the two models, respectively. For testing the hypotheses

$$H_0 : \boldsymbol{\alpha} = 0 \quad \text{against} \quad H_1 : \boldsymbol{\alpha} \neq 0, \tag{3}$$

One can consider the test statistic, namely ΔA_n, which has the form

$$\Delta A_n = A_n^* - A_n,$$

where $A_n = (n_0 n_1)^{-1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I\left(\hat{\eta}_j^{(1)} \geq \hat{\eta}_i^{(0)}\right)$ and $A_n^* = (n_0 n_1)^{-1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I\left(\tilde{\eta}_j^{*(1)} \geq \tilde{\eta}_i^{*(0)}\right)$ are the Mann–Whitney estimators of the AUC based on the fitted responses of models (1) and (2), commonly computed by the maximum likelihood estimators (MLE) $(\hat{\gamma}, \hat{\boldsymbol{\beta}})$ in model (1) and $(\tilde{\gamma}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}})$ in model (2), respectively. The superscript of η indicates the group (i.e. $Y = 0$ or 1) at which the predicted probabilities are drawn for comparison. Generally, the AUC itself can be regarded as the correct classification probability that a randomly selected event has a higher risk score than a randomly selected nonevent. Thus, a higher AUC value typically indicates a higher predictive accuracy of the model, and hence a higher degree of goodness-of-fit. Therefore, we reject the null hypothesis when we observe a fairly large value of ΔA_n. Essentially, the above H_0 in (3) for $\boldsymbol{V}_i$ having no association with the response is the same as $H_0 : \Delta A_n = 0$.[12]

The distribution of ΔA_n is desired for developing a statistical test based on the hypotheses in (3). For a nonnested model, one can certainly consider the widely used ΔAUC test proposed by DeLong et al. (1988), in which they derived a consistent estimator for the variance of ΔAUC and they showed that the test statistic under the null hypothesis has an asymptotic zero-mean normal distribution. However, several authors[2,4,5] recently showed that such method cannot be applied directly to the nested models because the test statistic can be degenerate. The invalidity arises mainly because the assumptions that $(\hat{\eta}_i, \tilde{\eta}_i^*)$ and $(\hat{\eta}_j, \tilde{\eta}_j^*)$ for $i \neq j$ are mutually independent and that increment and decrement of AUC are equally likely to occur at $\boldsymbol{\alpha} = 0$ are violated.[2] In particular, the asymptotic distribution of ΔA_n is still normal (i.e. the nondegenerate case) under the alternative hypothesis, that is, when at least one of the elements in $\boldsymbol{V}$ is associated with the response. Interestingly, its asymptotic distribution under the null hypothesis (i.e. the degenerate case) can be an infinite sum of weighted χ^2 random variables according to the theory of U-statistics.[5] Nonetheless, Heller et al.[4] considered the test statistic computed via estimating the regression parameters by treating the AUC's as the objective functions, through the use of a kernel function to approximate the nonparametric component of the AUC. They showed that the resulting asymptotic null distribution is a sum of q weighted χ_1^2 random variables. To date, it is remarked that the asymptotic null distribution of the test statistic computed via the most commonly used MLE of binary regression parameters is still considered as mathematically intractable in general. At least, Monte Carlo methods are deemed to be infeasible or otherwise impractical (also see the Appendix).

2.2 Logistic regression models with a change-point in the covariate effects

Change-point models are widely applicable in epidemiological and medical data analyses, where the relationship between the explanatory variable(s) and the response is deemed to be partly linear or nonlinear. In practice,

models with one or two change-points would be considered sufficient in finite-sample settings, as making inference on the models with three or more change-points usually requires a large sample size and a substantially longer computational time. Fong et al.[7] previously studied the statistical tests for the presence of a change-point in covariate effect based on the logistic regression model with an interaction term. They proposed to use the maximal LR tests and a Monte Carlo method to simulate the critical values, which have been proven efficient and consistent. With reference to Section 2.1, it is easy to see that change-point models can also be regarded as a special case of the nested models naturally, where an additional term associated with the change-point parameter is added to the original model to form the full model. This leads us to consider ΔAUC as a statistic for testing the presence of change-point in the covariate effects based on the logistic regression model, while the threshold variable is considered as a continuous surrogate measurement, namely V_1 below. Based on model (1), we are interested to test whether introducing a change-point effect of V_1 would increase the predictive accuracy of the original model significantly. We want to test for the hypotheses

$$H_0 : \alpha_1 = 0 \text{ for all } \delta \in \mathcal{B} \qquad \text{against} \quad H_1 : \alpha_1 \neq 0 \text{ for some } \delta \in \mathcal{B}, \tag{4}$$

in the alternative model

$$\eta_i^{**}(\delta) = P(Y_i = 1 | \boldsymbol{W}_i, V_{1i}) = \frac{\exp\{\gamma + \boldsymbol{\beta}^T \boldsymbol{W}_i + \alpha_1 (V_{1i} - \delta)_+\}}{1 + \exp\{\gamma + \boldsymbol{\beta}^T \boldsymbol{W}_i + \alpha_1 (V_{1i} - \delta)_+\}}, \tag{5}$$

where $a_+ = \max(0, a)$ for constant a, $q = 1$ and δ is an unknown change-point parameter lying in a data-dependent compact support $\mathcal{B}$ that regulates the effect of V_1 on the response. One merit for considering such model specification is that the covariate effect on response changes smoothly, which is considered more practical in real-life data analysis. Also, the model is particularly useful when V_1 is assumed to have negligible effects on the binary outcome given V_1 is relatively low compared to δ, but not for $V_1 > \delta$. In epidemiological or medical studies, δ generally indicates the maximum tolerance level (minimum dosage level) where V_1 could reach before it poses risk (protective effects) to the binary outcome of interest. There are various forms of modeling a change-point in the covariate effects, for example, one can replace $(V_1 - \delta)_+$ by simply a step function $I(V_1 - \delta > 0)$ in (5), or to add V_1 as an element in $\boldsymbol{W}$ in both (1) and (5) for a segmented regression, but these models will not be discussed further in this paper.

Under H_0 in (4), model (5) reduces to model (1) and the corresponding AUC calculation has been discussed in the last subsection. Nonetheless, standard methods for computation of AUC are not applicable to model (5) as δ is a nuisance parameter which is not present under H_0 and it is unknown in nature. Instead, we can first fix the value of δ in $\mathcal{B}$, then calculate the statistic $A_n^{**}(\delta) - A_n$ for each of the prespecified δ's, and select the maximal statistic as the test statistic. The revised ΔAUC test statistic under the change-point problem for H_0 in (4) is

$$\Delta A_n^* = \sup_{\delta \in \mathcal{B}} A_n^{**}(\delta) - A_n$$

where A_n is previously defined and

$$A_n^{**}(\delta) = (n_0 n_1)^{-1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I\left(\tilde{\eta}_j^{**}(\delta)^{(1)} \geq \tilde{\eta}_i^{**}(\delta)^{(0)}\right)$$

is the AUC calculated based on model (5), evaluated at the MLE $\left(\tilde{\gamma}(\delta), \tilde{\boldsymbol{\beta}}(\delta), \tilde{\alpha}_1(\delta)\right)$ for a given δ. As a counterpart, the commonly studied maximal LR statistic takes the form

$$LR_n^* = \sup_{\delta \in \mathcal{B}} LR(\delta) = \sup_{\delta \in \mathcal{B}} \left\{ 2\ell\left(\tilde{\gamma}(\delta), \tilde{\boldsymbol{\beta}}(\delta), \tilde{\alpha}_1(\delta); \delta\right) - 2 \sup_{\gamma, \boldsymbol{\beta}, \alpha_1} \ell(\gamma, \boldsymbol{\beta}, \alpha_1) \right\}$$

which is just two times the difference between the maximum log-likelihood function involved in (5) considering a profile of δ and the usual maximum log-likelihood function involved in (1). Moreover, a natural estimate for δ can

be obtained by considering $\tilde{\delta}_{\Delta AUC} = \text{argmax}_{\delta\in\mathcal{B}} A_n^{**}(\delta) - A_n$ and $\tilde{\delta}_{LR} = \text{argmax}_{\delta\in\mathcal{B}} LR(\delta)$, respectively. In practice, $\mathcal{B}$ includes all distinct values of the observed V_{1i} for $i = 1, \ldots, n$. To avoid edge effects where numerical instability may arise if we evaluate the statistic near the smallest or largest possible values of $\mathcal{B}$, we propose to search over the value of δ in-between the 10th and 90th percentiles of the observed values of the change-point variable V_1. Moreover, we divide the support equally by incorporating a constant grid size d to perform the search for both tests.

2.3 Resampling procedures for ΔA_n^*

The distribution of ΔA_n^* under the null hypothesis of $\boldsymbol{\alpha} = 0$ is proposed to be evaluated based on the resampling methods since its asymptotic distribution is unlikely to be tractable as discussed in Section 2.1, especially in the change-point settings. We adopt the following procedure to resample the data under the null hypothesis:

1. For a given dataset of sample size n, we fit model (1) to obtain the MLE $\hat{\gamma}$ and $\hat{\boldsymbol{\beta}}$. The predicted probability of the ith individual $\hat{\eta}_i$ is calculated for $i = 1, \ldots, n$.
2. Generate a total number of M simulated samples, say $M = 2000$. For the jth simulation, $j = 1, \ldots, M$, produce the simulated response $Y_i^{(j)}$ of the ith individual based on a binary random variable with probability $\hat{\eta}_i$, and $\boldsymbol{W}_i$ is fixed.
3. For testing the hypothesis in (4), calculate $\Delta A_n^{*(j)}$ based on the jth simulated sample, $j = 1, \ldots, M$. With level of significance τ, we reject H_0 if the calculated test statistic is larger than the $100 \times (1-\tau)$th percentile of $(\Delta A_n^{*(1)}, \ldots, \Delta A_n^{*(M)})$.

In the above steps, the critical values for the maximal LR test statistic LR_n^* can also be approximated easily as a by-product of the procedure.

Under H_1, a confidence interval of ΔA_n^* can be calculated based on the empirical percentiles of $(\Delta A_n^{*(1)}, \ldots, \Delta A_n^{*(M)})$, but we simulate the response $Y_i^{(j)}$ using the predicted probability $\tilde{\eta}_i^{**}(\delta)$ based on the MLE of (5) with $\delta = \tilde{\delta}_{\Delta AUC}$ instead of $\hat{\eta}_i$.

2.4 Confidence interval estimation for δ

If the null hypothesis in (4) is rejected, providing a confidence interval for δ is of interest, but the asymptotic distribution of $\tilde{\delta}$ can be complicated. This is a nonstandard problem where the objective functions $A_n^{**}(\delta)$ and $LR(\delta)$ are not differentiable with respect to δ. In general, the classical bootstrap proposed by Efron and Tibshirani[13] produces inconsistent estimators in such problem. Based on the change-point Cox regression model, Lee and Lam[14] shows empirically that the classical bootstrap estimator for the change-point parameter can be inconsistent. For clustered survival data, Deng et al.[15] proposed to use the m-out-of-n bootstrap approach to construct an equal-tailed confidence interval for the change-point parameter and use the method proposed by Bickel and Sakov[16] to select the desired value of m. The m-out-of-n bootstrap approach is defined by sampling m observations with replacement from a dataset of size n, where $m \to \infty$ and $m/n \to 0$. Following the work of Bickel and Sakov[16] and Deng et al.,[15] we propose the following algorithm to select the bootstrap distribution for δ with optimal m, denoted by m^*.

1. Create a sequence of m given by $m_j = [r^j n]$ for $j = 1, 2, \ldots$ and $r \in (0, 1)$, where $[a]$ is the greatest integer less than or equal to a. For each m_j, calculate the empirical bootstrap distribution for the change-point estimator

$$\tilde{F}_{m_j,n}(x) = N^{-1} \sum_{i=1}^{N} I\left(m_j^{1/2} (\tilde{\delta}_{m_j} - \tilde{\delta}) \le x \right)$$

where N is the number of bootstrap samples, $\tilde{\delta}_{m_j}$ is the change-point estimator based the sample of size m_j, and $\tilde{\delta}$ is the change-point estimator based the original sample of size n, evaluated by the metrics $A_n^{**}(\delta)$ or $LR(\delta)$.

2. Let d(.,.) be the Kolmogorov–Smirnov distance and set $m^* = \arg\min_{m_j} d(\tilde{F}_{m_j,n}(x), \tilde{F}_{m_{j+1},n}(x))$. If more than one m_j achieve the minimum, the largest one is chosen.

3. The desired m-out-of-n bootstrap estimator is $\tilde{F}_{m^*,n}(x)$. Denote Q_ψ as the $100 \times \psi$th percentile of the sampled deviation $(\tilde{\delta}_{m^*} - \tilde{\delta})$ from $\tilde{F}_{m^*,n}(x)$. A 95% equal-tailed confidence interval for δ is constructed by $\left(\tilde{\delta} + (m^*/n)^{1/2} Q_{0.025}, \tilde{\delta} + (m^*/n)^{1/2} Q_{0.975}\right)$, where $(m^*/n)^{1/2}$ is an adjustment factor for overestimated variance when $m^* < n$.

There are mainly two distinctions between the above approach and that used in Deng et al.[15] First, we adopt an asymmetric rather than a symmetric confidence interval. Second, the change-point parameter estimate for δ in model (5) has been proven to be $n^{1/2}$-consistent[17] but not n-consistent.

3 Simulation study

We evaluate the finite-sample performance of the proposed test by simulation study. According to the model in (5), we consider the following regression model in the simulation:

$$\pi\{P(Y_i = 1 | W_{1i}, V_{1i})\} = \gamma + \beta_1 W_{1i} + \alpha_1 (V_{1i} - \delta)_+,$$

where $\pi(.)$ is the logit function. We want to test for the null hypothesis of $\alpha_1 = 0$ for all $\delta \in \mathcal{B}$. The random variables W_1 and V_1 are independently generated from the standard normal distribution. In all scenarios that we consider below, the regression parameter β_1 is set to 1, and the proportion of events for the population $E(Y)$ is set to be 0.1, 0.25, or 0.5 via controlling the value of γ. For the calculation of the proposed test statistic, we adopt a grid size $d = 0.1$ in the search of δ and perform resampling $M = 2000$ times for its reference distribution. Table 1 summarizes the performance of the resampling procedures in approximating the critical values for the two maximal tests under the null model with no change-points (i.e. $\alpha_1 = 0$). Based on 1000 replicates, the right-tail empirical percentiles of the test statistics, together with the corresponding averaged critical values, are reported. Three levels of significance are considered, namely $\tau = 0.1$, 0.05, and 0.01. Under the null hypothesis of no change-points, the resampled critical values averages congruent to the empirical percentiles in both tests. In comparison to LR_n^*, a different behavior to ΔA_n^* is that the increase in n will result in the shrinkage of the critical values. This is due to the fact that the test statistic is nonstandardized in nature. Hence, in measuring the dispersion of the resampled critical values, we report the standard-error-to-mean ratio for ΔA_n^* and simply the standard error for LR_n^* in Table 1. Typically, the dispersion measurements of the resampled critical values decrease as the sample size n increases, indicating a more accurate approximation.

Table 2 summarizes the empirical rejection rates in the above three settings under H_0 and six under H_1 where the change-point parameter δ is set to 0. In general, the maximal LR test has very stable and good performance of type I error rate control. The AUC test performs well when $E(Y) = 0.5$ or $E(Y) = 0.25$, but it is noticeably conservative when the prevalence is low, say $E(Y) = 0.1$, with $n < 500$. Under the alternative models (i.e. the last six scenarios of Table 2), it can be seen that both tests provide reasonable power, but the proposed ΔAUC test is generally less powerful than the maximal LR test. Indeed, Seshan et al.[2] have reported that ΔAUC test has lower power than the Wald test based on the ordinary nested logistic regression model detailed in Section 2.1. Since the LR test is asymptotically equivalent to Wald test under the null hypothesis in (3), the presented result for the maximal versions of ΔAUC and LR test statistics here is consistent with their findings.

For illustration purpose, we plot the bivariate distribution of the change-point parameter estimate $\tilde{\delta}$ based on ΔA_n^* and LR_n^* under the null and alternative hypothesis settings, respectively (Figure 1). It is noted that we do not attempt to make inference on the bivariate distribution of the two estimators, but to explore the possible dependency between them, empirically. Under H_0 where the change-point parameter δ does not exist, the estimators scatter around the space quite uniformly, and noticeable dissonance can be observed. A reviewer remarked that the AUC- and LR-based estimators of δ are correlated but can produce opposite results under H_0. Nonetheless, the two estimators cluster around the true change-point parameter value $\delta = 0$ under H_1 with either $\alpha = -1$ or $\alpha = 1$, and they also tend to assemble along the diagonal lines of the graphs. This indicates that, given a dataset from H_1, the change-point parameter estimates provided by the maximal ΔAUC test and LR tests generally "agree" with each other, although they are evaluated based on different metrics. Our empirical analysis shows that the population event proportion $E(Y)$ has no alarming effects on the change-point parameter estimation. Hence, we do not plot the diagrams for scenarios with $E(Y) = 0.25$ and $E(Y) = 0.1$ here.

Table 1. Approximation of the critical values for ΔA_n^* (in the scale of 10^{-3}) and LR_n^* under H_0 evaluated at the $100(1-\tau)$th percentiles.

			Empirical percentiles			Averaged critical values		
$E(Y)$	Method	n	$\tau = 0.1$	$\tau = 0.05$	$\tau = 0.01$	$\tau = 0.1$	$\tau = 0.05$	$\tau = 0.01$
0.5	ΔA_n^*	200	13.72	17.40	26.47	14.26 (0.18)	18.61 (0.18)	28.64 (0.18)
		300	9.61	11.85	18.76	9.43 (0.15)	12.33 (0.15)	19.07 (0.16)
		500	5.87	7.41	11.25	5.62 (0.12)	7.39 (0.12)	11.45 (0.13)
	LR_n^*	200	4.27	5.73	9.17	4.27 (0.15)	5.63 (0.21)	8.84 (0.47)
		300	4.40	5.74	8.76	4.21 (0.15)	5.55 (0.20)	8.70 (0.45)
		500	4.34	5.52	9.73	4.17 (0.14)	5.49 (0.19)	8.61 (0.43)
0.25	ΔA_n^*	200	17.88	22.87	34.40	18.82 (0.24)	24.61 (0.24)	38.01 (0.24)
		300	11.87	15.68	25.77	12.22 (0.19)	16.01 (0.19)	24.85 (0.19)
		500	7.48	9.97	15.83	7.23 (0.14)	9.51 (0.14)	14.80 (0.15)
	LR_n^*	200	4.38	5.66	8.09	4.35 (0.17)	5.75 (0.25)	9.09 (0.51)
		300	4.43	5.85	8.78	4.26 (0.15)	5.63 (0.22)	8.86 (0.46)
		500	4.46	5.98	8.42	4.19 (0.14)	5.54 (0.20)	8.66 (0.45)
0.1	ΔA_n^*	200	37.28	50.27	79.54	40.83 (0.36)	52.83 (0.35)	81.24 (0.35)
		300	23.27	33.48	47.70	26.36 (0.32)	34.56 (0.31)	53.36 (0.31)
		500	14.63	18.96	32.73	14.68 (0.24)	19.43 (0.24)	30.54 (0.24)
	LR_n^*	200	4.82	6.25	9.95	4.89 (0.26)	6.35 (0.27)	9.60 (0.45)
		300	4.48	6.03	9.90	4.62 (0.28)	6.13 (0.32)	9.48 (0.47)
		500	4.29	6.13	10.12	4.34 (0.18)	5.76 (0.25)	9.17 (0.53)

Standard-error-to-mean ratio and standard error of the resampled critical values are presented in the brackets for ΔAUC and LR tests, respectively.

Table 3 reports the coverage probability of the m-out-of-n bootstrap approach for the nominal 95% confidence interval for δ. We set the parameters $\beta_1 = -1$ and $\alpha_1 = 2$ or –2. The sequence of candidates for optimal m are set to be $m_j = [0.75^j n]$ with $j = 1, \ldots, 5$ for $n = 200$ and $j = 1, \ldots, 8$ for $n = 300$, 500, respectively. The number of bootstrap samples is set to be $N = 200$, and the searching grid size is set to be $d = 0.01$. In general, it can be shown that the m-out-of-n bootstrap provides proper coverage to the change-point parameter.

4 Application

4.1 Mother-to-child-transmission dataset

The proposed method for change-point determination, detailed in Section 2.2, is applied to the mother-to-child HIV transmission dataset at which the clinical study aims at exploring the association between immunological biomarkers of the mothers and the risk of transmission of HIV-1 viruses to the new-born babies.[18] The dataset includes 236 HIV-infected mothers ($n = 236$) where 79 of them are transmitters. In addition to the binary response, the dataset contains two explanatory variables. These are a continuous variable named *NAb_SF*162*LS* (treated as V_1) as a measure of neutralization activity of the HIV-1 viruses and the types of birth (treated as W_1) which has been coded as 0 for vaginal birth and 1 for birth via Caesarean section. The effects of *NAb_SF*162*LS* can be thought to be thresholded at a particular value. In general, it is assumed that this covariate has no effect on response if its intensity is lower than a certain threshold value, but it starts to grant a protective effect to the risk of HIV transmission provided that its intensity is large enough. Previously, Fong et al.[19] proposed to use the maximal LR test to analyze the dataset where a Monte Carlo procedure was adopted to sample the corresponding critical values. For illustration, we hereby apply the proposed ΔAUC-based method for change-point determination in comparison to the maximal LR test. As the variable *NAb_SF*162*LS* ranges from 3.9 to 14.1, a grid size of $d = 0.05$ is adopted in the computation of the proposed test statistics and the resampling procedure. In approximating the distribution of the test statistics, the number of resampling M is set to be 10000.

Figure 2 illustrates the profiles of our proposed ΔAUC statistic and the LR statistic. It is observed that the ΔAUC statistic attains its maximal value at $\tilde{\delta}_{\Delta AUC} = 7.58$, while the LR statistic attains its maximum at $\tilde{\delta}_{LR} = 7.33$. The two estimates for the change-point location are very close to each other. The results of the statistical tests can be found in Table 4. The null hypothesis of no change-points is rejected based on both

Table 2. Empirical rejection rates of the proposed ΔAUC test and the maximal likelihood ratio test, evaluated at the $100(1-\tau)$th percentiles, under the null and alternative hypotheses.

			ΔA_n^*			LR_n^*		
			Rejection rates			Rejection rates		
α_1	$E(Y)$	n	$\tau = 0.1$	$\tau = 0.05$	$\tau = 0.01$	$\tau = 0.1$	$\tau = 0.05$	$\tau = 0.01$
0	0.5	200	0.093	0.046	0.008	0.099	0.054	0.009
		300	0.106	0.052	0.009	0.111	0.055	0.009
		500	0.108	0.052	0.008	0.110	0.052	0.013
0	0.25	200	0.087	0.042	0.007	0.103	0.047	0.008
		300	0.091	0.047	0.006	0.110	0.055	0.009
		500	0.110	0.059	0.008	0.114	0.057	0.007
0	0.1	200	0.076	0.032	0.003	0.099	0.047	0.011
		300	0.081	0.033	0.005	0.094	0.047	0.014
		500	0.097	0.049	0.009	0.103	0.057	0.015
−1	0.5	200	0.936	0.889	0.712	0.954	0.919	0.764
		300	0.982	0.970	0.902	0.990	0.977	0.936
		500	1.000	0.998	0.993	1.000	1.000	0.998
−1	0.25	200	0.818	0.700	0.421	0.854	0.748	0.519
		300	0.940	0.886	0.694	0.959	0.921	0.760
		500	0.993	0.981	0.931	0.999	0.988	0.961
−1	0.1	200	0.494	0.332	0.102	0.589	0.422	0.209
		300	0.663	0.521	0.236	0.698	0.576	0.346
		500	0.846	0.748	0.437	0.892	0.817	0.545
1	0.5	200	0.942	0.880	0.713	0.953	0.903	0.775
		300	0.983	0.970	0.908	0.992	0.982	0.941
		500	0.999	0.999	0.996	0.999	0.999	0.998
1	0.25	200	0.877	0.819	0.591	0.930	0.868	0.702
		300	0.973	0.946	0.845	0.988	0.972	0.909
		500	1.000	1.000	0.987	1.000	1.000	0.997
1	0.1	200	0.643	0.506	0.237	0.749	0.639	0.419
		300	0.828	0.723	0.481	0.901	0.845	0.653
		500	0.963	0.931	0.791	0.986	0.969	0.905

Change-point parameter $\delta = 0$ for $\alpha_1 \neq 0$.

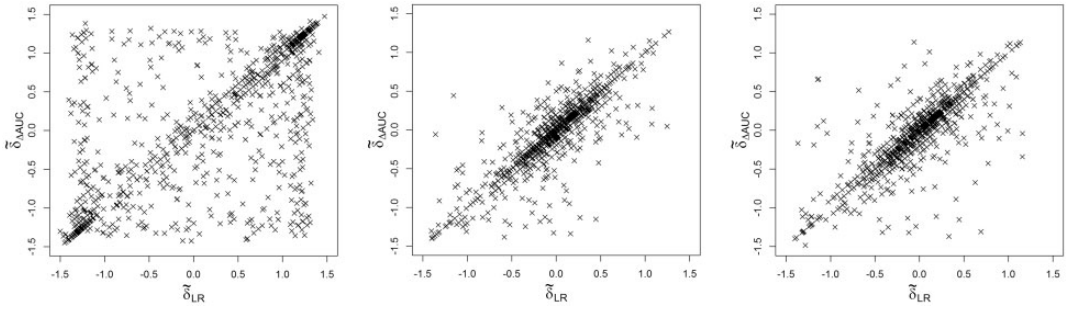

Figure 1. Empirical distributions of $\tilde{\delta}$ based on ΔA_n^* and LR_n^* for scenarios with $E(Y) = 0.5$, $n = 500$; left panel: $\alpha_1 = 0$, middle panel: $\alpha_1 = 1$, and right panel: $\alpha_1 = -1$.
AUC: area under the receiver operating characteristic curve; LR: likelihood ratio.

tests with p-values less than 0.01 from the resampling procedures. A significant incremental amount of predictive accuracy (0.1182, in terms of AUC) can be gained when we consider a change-point model in this application. A resampling 95% confidence interval for the AUC improvement is $(0.0510, 0.1902)$. Based on $N = 1000$ bootstrap samples with $d = 0.01, m_j = [0.75^j n]$ for $j = 1, \ldots, 5$, the m-out-of-n 95% confidence intervals for the change-point parameter δ are $(6.43, 9.34)$ and $(6.03, 10.47)$ based on the ΔAUC and LR tests, respectively. For illustrative

Table 3. Coverage probability of the nominal 95% confidence interval for δ based on the *m*-out-of-*n* bootstrap under the alternative hypothesis.

		ΔA_n^*			LR_n^*		
α_1	$E(Y)$	$n = 200$	$n = 300$	$n = 500$	$n = 200$	$n = 300$	$n = 500$
–2	0.5	0.967	0.965	0.976	0.942	0.961	0.973
–2	0.25	0.953	0.974	0.965	0.955	0.937	0.954
–2	0.1	0.957	0.928	0.935	0.932	0.943	0.937
2	0.5	0.970	0.966	0.975	0.947	0.963	0.970
2	0.25	0.972	0.967	0.962	0.940	0.948	0.943
2	0.1	0.965	0.936	0.930	0.931	0.926	0.923

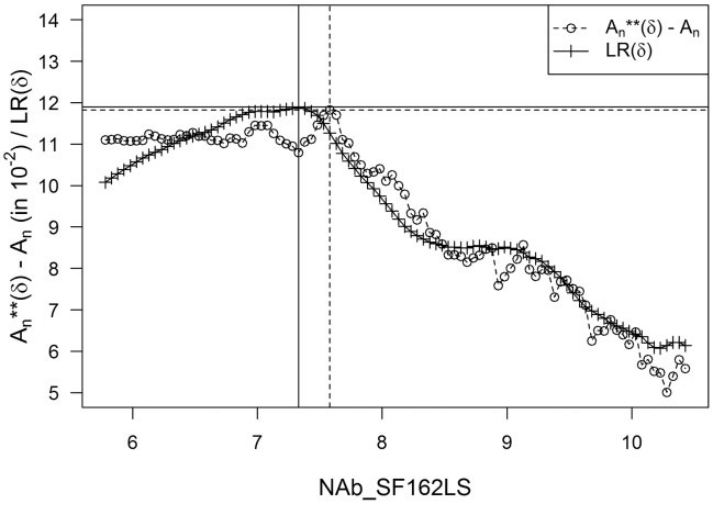

Figure 2. The plots of $A_n^{**}(\delta) - A_n$ (in the scale of 10^{-2}) and $LR(\delta)$ for the mother-to-child HIV transmission dataset at which $\tilde{\delta} = 0.758$ based on the proposed test statistic ΔA_n^* and $\tilde{\delta} = 0.733$ based on the maximal likelihood ratio test statistic LR_n^*.
LR: likelihood ratio.

Table 4. Test results for the mother-to-child HIV transmission dataset.

			Critical values via resampling			
Test	Statistic	$\tilde{\delta}$	$\tau = 0.1$	$\tau = 0.05$	$\tau = 0.01$	*p*-value
ΔA_n^*	0.1182	7.58	0.0629	0.0730	0.0959	0.0011
LR_n^*	11.8947	7.33	4.4374	5.8430	9.3634	0.0023

purpose, the logistic regression model conditioned on $\delta = 7.58$ is fitted, and it is observed that the thresholding effect associated with $(NAb\_SF162LS - \delta)_+$ is significantly negative ($\tilde{\alpha}_1 = -0.4104$) according to its standard error ($s.e.(\tilde{\alpha}_1) = 0.1306$). For $NAb\_SF162LS > 7.58$, the odd ratio is 0.6633 when there is a unit increase of $NAb\_SF162LS$.

4.2 Chronic bronchitis dataset

The second dataset studied the occurrence of chronic bronchitis of $n = 1256$ Munich workers which was collected between 1966 and 1977. The data were analyzed by Ulm[8] and Küchenhoff and Carroll.[9] The association between the risk of developing bronchitis disease and three explanatory variables, namely average dust concentration (temporarily denoted as variable Z) in the industrial workplace, duration of exposure (W_1), and smoking

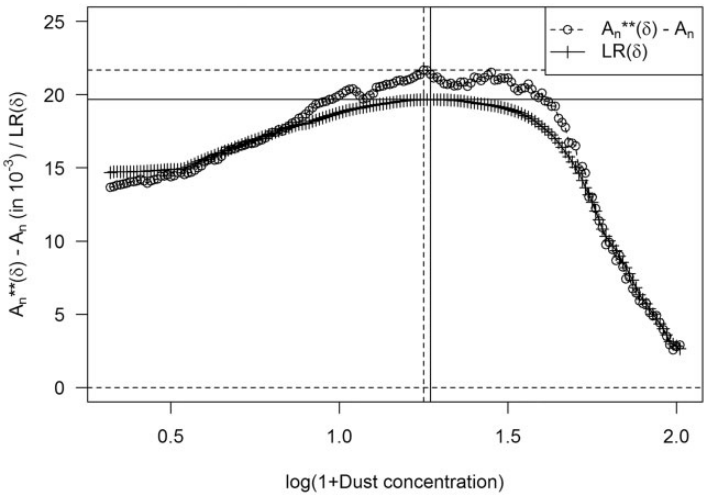

Figure 3. The plots of $A_n^{**}(\delta) - A_n$ (in the scale of 10^{-3}) and $LR(\delta)$ for the bronchitis dataset at which $\tilde{\delta} = 1.25$ based on the proposed test statistic ΔA_n^* and $\tilde{\delta} = 1.27$ based on the maximal likelihood ratio test statistic LR_n^*.
LR: likelihood ratio.

Table 5. Test results for the bronchitis dataset.

			Critical values via resampling			
Test	Statistic	$\tilde{\delta}$	$\tau = 0.1$	$\tau = 0.05$	$\tau = 0.01$	*p*-value
ΔA_n^*	0.0217	1.25	0.0054	0.0069	0.0112	<0.0001
LR_n^*	19.6634	1.27	4.5397	6.0803	8.7192	<0.0001

status (W_2), is explored. Following Küchenhoff and Carroll,[9] we adopt a monotonic transformation to the average dust concentration such that the change-point variable of interest becomes $V_1 = \log(1 + Z)$ which ranges from 0.182 to 3.219. The effects of V_1 on the chance of having chronic bronchitis are commonly modeled by a change-point parameter δ which is termed as "threshold limit value" in occupational medicine. In particular, low dust concentration (i.e. low value of V_1) has no effect on the risk of workers, but the risk will increase subsequently if the dust concentration is found to be higher than a particular unknown level δ.

In order to test for the presence of a change-point, we apply the proposed method to the dataset with $M = 10{,}000$ and $d = 0.01$. From Figure 3, the change-point estimates of the two maximal tests, namely 1.25 for ΔA_n^* and 1.27 for LR_n^*, are very close to each other. In Table 5, the two tests reject the null hypothesis of no change-points with *p*-values < 0.001. The AUC improvement is 0.0217 with a resampling 95% confidence interval $(0.0070, 0.0437)$. Based on $N = 1000$ bootstrap samples with $d = 0.002, m_j = [0.75^j n]$ for $j = 1, \ldots, 8$, the *m*-out-of-*n* 95% confidence intervals for δ are $(0.70, 1.68)$ and $(0.66, 1.66)$ based on the ΔAUC and LR tests, respectively. Similar to previous example, we fit the model conditioned on $\delta = 1.25$, and the estimated coefficient for $(V_1 - \delta)_+$ is 0.8261 with standard error 0.1850. This may indicate that, when the average dust concentration at workplace is higher than $\exp(1.25) - 1 = 2.49$ mg/cm^3, it starts to attribute to the risk of developing bronchitis of the Munich workers.

After all, the performance of the AUC and LR tests in the two applications is very similar to each other in terms of change-point detection and change-point location estimates.

5 Discussion

As areas under the ROC curves provide summary measures for discriminatory accuracy of the underlying predictive models, it is natural to consider the pairwise difference of the AUC for model comparison. This paper tries to provide a brief review on the use of ΔAUC for comparing nested binary regression model and explore the potential difficulties in obtaining the reference distribution of the test statistic under the null hypothesis. In

particular, it is shown that the null distribution, when using the commonly used MLE of binary regression parameters to compute the statistic, could be complicated, and Monte Carlo methods cannot be applied directly in practice. Hence, the most popular or reliable way for providing a correct reference distribution will certainly lie on some data permutation or perturbation methods. Nevertheless, the efficiency of the latter methods may compromise the use of ΔAUC for the hypothesis testing purposes, especially in the large n or large p problems. On the contrary, the LR test or Wald test is very efficient, and the results can be obtained easily from standard software outputs. Even in logistic models with a change-point, the maximal LR test can be shown to have an asymptotic distribution represented by score vectors, and it can be achieved easily via the simulation of multivariate normal random variables.[7] Despite the above, prevalence of conducting a ΔAUC test remains high in the field of biostatistics for comparing nested logistic models.

Change-point analysis for binary outcomes is one of the most important research topics in medical research for risk prediction, since the linear assumption in the covariate effects may not be appropriate on some occasions. We study a special case of the nested logistic regression model where the influence of a newly added covariate is regulated by an unknown change-point parameter δ under the alternative model. A statistical test based on ΔAUC is proposed for the presence of a change-point in the effects of a newly added covariate, provided that there are already p explanatory variables that are linearly associated with the outcome of interest. A simple grid search method is adopted to compute the proposed test statistic and the estimator of the change-point location. A reviewer remarked that it is also important to consider alternative methods such as the gradient descent method when the clinically meaningful threshold is small but the support of the biomarker is large. A resampling method is proposed to obtain the approximated critical values for the test statistic, that it does not attempt to restore the normality of the test statistic but to retrieve the exact null distribution. It is found that the proposed test is valid and feasible. Simulation study shows that the newly proposed test works well under the null and alternative hypotheses in finite-sample size settings.

Previously, Pepe et al.[12] showed that the test for the association between a new predictor variable and the response is equivalent to the AUC test under the nested logistic models. Hence, it is often recommended for practitioners to avoid testing for AUC improvement but to conduct the generally most powerful test for association (e.g. LR test) in evaluating the performance of a new predictor variable. In the change-point settings, we also show empirically that ΔA_n^* has similar, or otherwise slightly inferior power than LR_n^* under the alternatives. The natural change-point estimators $\tilde{\delta}$ derived from the two maximal tests share similar patterns, and they are able to capture the true location most of the time under H_1. It suffices to say that the ΔAUC test could only provide minimal amount of additional information, that is, the incremental value of the predictive accuracy for the change-point model against the reduced linear model. In this case, it is more clinically important to construct the confidence intervals for the AUC improvement and the change-point location.

Recently, there is an increased popularity of the analysis using ROC curves to manipulate the performance of risk prediction in survival models.[20] The idea is to consider a series of time-dependent ROC curves developed based on the individual counting processes such that a scalar measure for the overall model predictive accuracy can also be achieved by considering the integral of AUC(t) with respect to, say, the distribution of time.[21] Since the censored data also contains a binary response and continuous explanatory variables as the structure, it is conjectured that the proposed method can be extended to accommodate survival outcomes in the near future.

Acknowledgements

The author is grateful to the editor and two reviewers for their valuable comments and suggestions that greatly improved the quality of the paper.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Chun Yin Lee https://orcid.org/0000-0002-7207-2519

References

1. Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**: 29–36.
2. Seshan VE, Gönen M and Begg CB. Comparing ROC curves derived from regression models. *Stat Med* 2013; **32**: 1483–1493.
3. DeLong ER, DeLong DM and Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**: 837–845.
4. Heller G, Seshan VE, Moskowitz CS, et al. Inference for the difference in the area under the ROC curve derived from nested binary regression models. *Biostatistics* 2016; **18**: 260–274.
5. Demler OV, Pencina MJ, Cook NR, et al. Asymptotic distribution of ΔAUC, NRIs, and IDI based on theory of U-statistics. *Stat Med* 2017; **36**: 3334–3360.
6. Demler OV, Pencina MJ and D'Agostino Sr RB. Misuse of Delong test to compare AUCs for nested models. *Stat Med* 2012; **31**: 2577–2587.
7. Fong Y, Di C and Permar S. Change point testing in logistic regression models with interaction term. *Stat Med* 2015; **34**: 1483–1494.
8. Ulm K. A statistical method for assessing a threshold in epidemiological studies. *Stat Med* 1991; **10**: 341–349.
9. Küchenhoff H and Carroll RJ. Segmented regression with errors in predictors: semiparametric and parametric methods. *Stat Med* 1997; **16**: 169–188.
10. Pastor-Barriuso R, Guallar E and Coresh J. Transition models for change-point estimation in logistic regression. *Stat Med* 2003; **22**: 1141–1162.
11. Lee S, Seo MH and Shin Y. Testing for threshold effects in regression models. *J Am Stat Assoc* 2011; **106**: 220–231.
12. Pepe MS, Kerr KF, Longton G, et al. Testing for improvement in prediction model performance. *Stat Med* 2013; **32**: 1467–1482.
13. Efron B and Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci* 1986; **1**: 54–75.
14. Lee CY and Lam KF. Survival analysis with change-points in covariate effects. *Stat Methods Med Res* 2020; **29**: 3235–3248.
15. Deng Y, Zeng D, Zhao J, et al. Proportional hazards model with a change point for clustered event data. *Biometrics* 2017; **73**: 835–845.
16. Bickel PJ and Sakov A. On the choice of *m* in the *m* out of *n* bootstrap and confidence bounds for extrema. *Stat Sin* 2008; **18**: 967–985.
17. Fong Y, Di C, Huang Y, et al. Model robust inference for continuous threshold regression models. *Biometrics* 2017; **73**: 452–462.
18. Permar SR, Fong Y, Vandergrift N, et al. Maternal HIV-1 envelope-specific antibody responses and reduced risk of perinatal transmission. *J Clin Invest* 2015; **125**: 2702–2706.
19. Fong Y, Huang Y, Gilbert PB, et al. Chngpt: threshold regression model estimation and inference. *BMC Bioinform* 2017; **18**: 454–460.
20. Heagerty PJ and Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005; **61**: 92–105.
21. Lambert J and Chevret S. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Stat Methods Med Res* 2016; **25**: 2088–2102.
22. Cox DR and Hinkley DV. *Theoretical statistics*. Boca Raton, FL: Chapman and Hall/CRC, 1979.

Appendix

Here, we show briefly that the test statistic based on AUC computed via the MLE could be complicated. Consider just the simple case for $q = 1$ in model (2) and let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha_1)$ be the regression parameters without the intercept term γ. Denote the MLE for $\boldsymbol{\beta}$ in model (1) be $\hat{\boldsymbol{\beta}}$ and the MLE for $\boldsymbol{\theta}$ in model (2) be $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{\alpha_1})$.

Since $\exp(x)/\{1 + \exp(x)\}$ is an increasing function, it suffices to write

$$A_n(\hat{\boldsymbol{\beta}}) = (n_0 n_1)^{-1} I\Big[\hat{\boldsymbol{\beta}}^T(\boldsymbol{W}_j^{(1)} - \boldsymbol{W}_i^{(0)}) \geq 0\Big],$$

$$A_n^*(\tilde{\boldsymbol{\theta}}) = (n_0 n_1)^{-1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I\Big[\tilde{\boldsymbol{\beta}}^T(\boldsymbol{W}_j^{(1)} - \boldsymbol{W}_i^{(0)}) + \tilde{\alpha_1}(V_{1j} - V_{1i}) \geq 0\Big].$$

Assume that we can apply some twice differentiable smooth functions to approximate $I(.)$ based on some tuning parameters. Denote $\boldsymbol{G}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{G}_{\boldsymbol{\beta}}^T & G_{\alpha_1} \end{pmatrix}^T = \frac{\partial}{\partial \boldsymbol{\theta}} A_n^*(\boldsymbol{\theta})$ and $\Sigma(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{G}(\boldsymbol{\theta}) = \begin{pmatrix} \Sigma_{\boldsymbol{\beta}\boldsymbol{\beta}} & \Sigma_{\boldsymbol{\beta}\alpha_1} \\ \Sigma_{\alpha_1\boldsymbol{\beta}} & \Sigma_{\alpha_1\alpha_1} \end{pmatrix}$. Under H_0, $\alpha_{10} =$

0, $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, 0)$ and $A_n^*(\boldsymbol{\theta}_0) = A_n(\boldsymbol{\beta}_0)$. Taylor series expansions of $A_n(\hat{\boldsymbol{\beta}})$ around $\boldsymbol{\beta}_0$ and $A_n^*(\tilde{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$ yields the difference

$$\begin{aligned} A_n^*(\tilde{\boldsymbol{\theta}}) - A_n(\hat{\boldsymbol{\beta}}) = {} & \boldsymbol{G}(\boldsymbol{\theta}_0)^T(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \boldsymbol{G}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ & + \frac{1}{2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\Sigma(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \frac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\Sigma_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(1). \end{aligned}$$

According to the proof of theorem 1 in Heller et al.,[4] and see also p. 308 of Cox and Hinkley,[22] we can relate $\hat{\boldsymbol{\beta}}$ with $\tilde{\boldsymbol{\theta}}$ by

$$(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) = -\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{I}_{\boldsymbol{\beta}\alpha_1}(\boldsymbol{\theta}_0)\tilde{\alpha_1},$$

where $\boldsymbol{I}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \boldsymbol{I}_{\boldsymbol{\beta}\alpha_1} \\ \boldsymbol{I}_{\alpha_1\boldsymbol{\beta}} & I_{\alpha_1\alpha_1} \end{pmatrix}$ is the information matrix derived from the negative second-order derivative of the log-likelihood of model (2). It immediately follows that

$$\begin{aligned} T_1 &= \boldsymbol{G}(\boldsymbol{\theta}_0)^T(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \boldsymbol{G}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &= \boldsymbol{G}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^T(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) + G_{\alpha_1}(\boldsymbol{\theta}_0)\tilde{\alpha_1} \\ &= \left[G_{\alpha_1}(\boldsymbol{\theta}_0) - G_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^T\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{I}_{\boldsymbol{\beta}\alpha_1}(\boldsymbol{\theta}_0)\right]\tilde{\alpha_1}. \end{aligned}$$

Hence, $n^{1/2}T_1$ is a zero-mean Gaussian variable. Similarly, we can show that

$$\begin{aligned} T_2 &= \frac{1}{2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\Sigma(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \frac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\Sigma_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &= \left\{\Sigma_{\alpha_1\boldsymbol{\beta}}(\boldsymbol{\theta}_0) - \boldsymbol{I}_{\boldsymbol{\beta}\alpha_1}(\boldsymbol{\theta}_0)^T\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{\theta}_0)\Sigma_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_0)\right\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\tilde{\alpha_1} \\ &\quad + \frac{1}{2}\left\{\Sigma_{\alpha_1\alpha_1}(\boldsymbol{\theta}_0) + \boldsymbol{I}_{\boldsymbol{\beta}\alpha_1}(\boldsymbol{\theta}_0)^T\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{\theta}_0)\Sigma_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_0)\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{I}_{\boldsymbol{\beta}\alpha_1}(\boldsymbol{\theta}_0)\right\}\tilde{\alpha_1}^2, \end{aligned}$$

in which the asymptotic distribution of the first term in $2nT_2$ is unable to be resolved easily, but the second term is clearly associated with a properly scaled χ_1^2 distribution.