Composite quantile regression for the receiver operating characteristic curve

Xiaogang Duan [1], and Xiao-Hua Zhou [2,3]
[1]Department of Statistics, Beijing Normal University, Beijing 100875, China
[2]Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.
[3] Corresponding author: Xiao-Hua Zhou, email: azhou@u.washington.edu

Abstract

The covariate-specific receiver operating characteristic curve is frequently used to evaluate the classification accuracy of a diagnostic test when it is associated with certain covariates. In this paper, we proposed a new procedure for estimating this curve based on a reformulation of the conventional location-scale model as well as the idea of composite quantile regression. Asymptotic normality of the proposed estimators is established, both for the regression parameters and the covariate-specific receiver operating characteristic curve at a fixed false positive point. Simulation results show that the new estimators compare favorably to their main competitors in terms of the standard error. We apply the new procedure to data from the national Alzheimer's coordinating center.

Keywords: Composite quantile regression; Location-scale model; Receiver operating characteristic curve.

1. Introduction

The receiver operating characteristic curve is a useful tool for evaluating the classification ability of a medical diagnostic test when the test result is continuous [12, 14]. This curve plots the sensitivity against 1 minus the specificity of a continuous test by first dichotomizing the test result with a threshold and then varying the threshold over the real line. Mathematically, it equals $S_1\{S_0^{-1}(t)\}$ for $t \in [0, 1]$, where S_1 and S_0 denote the survival functions of the diseased and nondiseased populations, respectively.

In practice, besides the disease status and diagnostic test value, additional covariates, which may influence the accuracy of the test, are often available. The covariate-specific receiver operating characteristic curve, denoted as $R_x(t)$ throughout, provides a way to evaluating the effect of covariates on the test accuracy; here x is a particular value of patient's covariates.

There are mainly two approaches for modeling $R_x(t)$ in the literature. The first is an indirect procedure that has two steps. At step one, a disease-specific regression model is built to capture the relationship between covariates and test results, then model parameters are estimated. At step two, $R_x(t)$ is estimated based on either an empirical or a nonparametric smoothing estimator of the disease-specific survival function using fitted residuals obtained at step one. A location-scale regression model is the most common form for the indirect method and has been widely used in practice. See Pepe [9, 10, 12], Heagerty & Pepe [4], Faraggi [2], Zheng & Heagerty [13], González-Manteiga et al. [3], and Liu & Zhou [8] for examples.

The second approach directly models $R_x(t)$ itself. At the heart of this approach is the specifi-

cation of a model, which directly relates covariates to $R_x(t)$. This method was first proposed by Pepe [9, 11] with a parametric model and then Cai & Pepe [1] extended it semiparametrically by allowing an arbitrary nonparametric baseline function.

Both approaches have advantages and disadvantages. The indirect method is rather flexible, as it does not require the specification of a particular form for any of the disease-specific error distributions. In general, the class of $R_x(t)$ generated by the indirect model is larger than that generated by the direct model. However, one disadvantage of the indirect model is the difficulty with its regression parameter interpretation. For the direct approach, interpreting the regression parameters is easy; yet the existing literature requires a known link function as well as a parametric structure to characterizing covariate's effect. More discussion of these methods may be found in Zhou et al. [14].

In this paper, we focus on the two-step indirect regression method, for which Wedderburn [18]'s quasilikelihood is frequently adopted for estimating the model parameters. This extends the least squares method, and thus inherits all its drawbacks. For example, the quasilikelihood method requires a moment assumption, and behaves badly when the data contain outliers.

Quantile regression has been widely used since the seminal work of Koenker & Bassett [6]. It is attractive not only due to its robustness to non-Gaussian errors, but also because, by considering several quantiles simultaneously, it provides a more complete picture of the conditional distribution of the response. Zou & Yuan [15] further proposed a composite quantile regression technique to combine information across different quantiles in a linear regression model. The asymptotic relative efficiency of the composite quantile regression estimator in relation to the least squares estimator for the regression parameters in a linear model is shown, under general conditions, to be bounded below by 0.7. See Zou & Yuan [15] for details.

In contrast to mean regression techniques, composite quantile regression or even single quantile regression has rarely been used in the analysis of $R_x(t)$. Zheng & Heagerty [13] considered a semiparametric estimator of time dependent receiver operating characteristic curves for longitudinal markers. However, their work, as an extension of Heagerty & Pepe [4], only used the idea of the quantile regression to estimate part of the receiver operating characteristic curve, and employed a quasilikelihood method and spline expansion to estimate the location and scale functions.

Our contributions in this paper are three-fold. First, we introduce the composite quantile regression technique to receiver operating characteristic curve regression under the framework of a location-scale model. This includes general single quantile regression as its special case. Second, we provide a reformulation of the conventional location-scale model, which not only looks much simpler than the old one, but also facilitates the use of composite quantile regression. Third, we work out the asymptotics of the proposed estimators both for the model parameters and the covariate-specific receiver operating characteristic curve at a fixed false positive point. A key difference between our model and that of Zou & Yuan [15] is that we permit the regression function to be nonlinear. Consequently, the criterion function involved in our setup is not necessarily convex in parameters. Thus, Zou & Yuan [15]'s method of proof cannot be used here. Instead, M-estimation ideas are adopted to prove the asymptotic normality of our estimators.

2. Methodology 2.1 Location-scale model

Let T, D and X be the continuous test result, the true disease status and the covariates available for a subject, respectively. Let $D = 1$ denote a diseased subject and 0 a healthy subject. In accordance with convention, we assume that larger values of T are more indicative of the disease.

Consider the location-scale model

$$T = \tilde{\mu}(X, D, \alpha) + D\sigma_1 e_1 + \sigma_0(1 - D)e_0, \tag{1}$$

where $\tilde{\mu}$ is a known function, α is a p-dimensional parameter vector, σ_1 and σ_0 are unknown scale parameters, e_1 and e_0 are disease-specific errors with mean zero and variance one that are both independent of X. Under model (1), the receiver operating characteristic curve given $X = x$ is $R_x(t) = s_1[(\sigma_0/\sigma_1)s_0^{-1}(t) + \sigma_1^{-1}\{\tilde{\mu}(x, 0, \alpha) - \tilde{\mu}(x, 1, \alpha)\}]$, where s_1 and s_0 denote the survival functions of e_1 and e_0, respectively.

Before introducing our estimators, we first provide a new formulation for model (1), which, as a referee pointed out, works here because our primary interest is $R_x(t)$ itself, while the model parameters are nuisance parameters. The basic idea of the new formulation is to reorganize the model so that no explicit assumptions are made on the errors except that they are independent of covariates. This is accomplished by absorbing the scale parameters, and any additive component that is either a constant or depends only on the disease status in the old location formulation, into newly defined errors. The new formulation usually looks simpler than the old one and also facilitates the use of composite quantile regression in estimating the model parameters.

To gain insight into the new formulation, we consider a linear location-scale model with one scalar covariate X, that is, $T = \alpha_1 + \alpha_2 D + \alpha_3 X + \alpha_4 XD + D\sigma_1 e_1 + \sigma_0(1 - D)e_0$. Decompose α_1 as $D\alpha_1 + (1 - D)\alpha_1$ and write $\varepsilon_0 = \alpha_1 + \sigma_0 e_0, \varepsilon_1 = \alpha_1 + \alpha_2 + \sigma_1 e_1$. Hence one can rewrite the preceding linear location-scale model as $T = \alpha_3 X + \alpha_4 XD + D\varepsilon_1 + (1 - D)\varepsilon_0$, or equivalently, $T = \beta_1 X + \beta_2 XD + D\varepsilon_1 + (1 - D)\varepsilon_0$, where $\beta_1 = \alpha_3$ and $\beta_2 = \alpha_4$, and ε_1 and ε_0 are newly defined disease-specific errors. Here, ε_1 and ε_0 are still independent of X. Clearly, the new formulation is much simpler than the old one, because the parameter vector $(\alpha_1, \alpha_2, \sigma_1, \sigma_0)^T$ disappears in the new framework; it is absorbed into the unspecified residuals ε_1 and ε_0.

Generally, model (1) can be reformulated as

$$T = \mu(X, D, \beta) + D\varepsilon_1 + (1 - D)\varepsilon_0, \tag{2}$$

where β is a q-dimensional parameter vector, μ is a known function, and ε_1 and ε_0 are disease-specific errors that are both independent of X and D. Under model (2), it follows that

$$R_x(t) = S_1\{S_0^{-1}(t) + \mu(x, 0, \beta) - \mu(x, 1, \beta)\}$$

for a fixed covariate value $X = x$, where S_0 and S_1 denote the respective survival functions of ε_0 and ε_1. For the preceding linear model, $R_x(t)$ equals $s_1\{(\sigma_0/\sigma_1)s_0^{-1}(t) - (\alpha_2 + \alpha_4 x)/\sigma_1\}$ under the old framework and $S_1\{S_0^{-1}(t) - \beta_2 x\}$ under the new framework, respectively.

2.2 Estimation

Now, we present the estimation method in more detail. Throughout, let K be a fixed positive integer and $\tau_1 < \cdots < \tau_K$ be K fixed points in $(0,1)$. Let $\{(T_i, D_i, X_i) : i = 1, \ldots, n\}$ be n independent copies of (T, D, X) that follows model (2). Further, let β^* be the true value of β, and for $j = 0, 1$, let $b^*_{j1}, \ldots, b^*_{jK}$ be the $\tau_1, \ldots, \tau_K$'th quantiles of ε_j, respectively. Under model (2), the τ_k'th quantile, for $k = 1, \ldots, K$, of T_i given (X_i, D_i) is $\mu(X_i, D_i, \beta^*) + D_i b^*_{1k} + (1 - D_i) b^*_{0k}$. Since the different conditional quantiles share a common parameter β^*, we propose to estimate it by solving the minimization problem

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \sum_{k=1}^{K} \rho_{\tau_k} \left\{ T_i - \mu_i(\beta) - D_i b_{1k} - (1 - D_i) b_{0k} \right\},$$

where $\mu_i(\beta) = \mu(X_i, D_i, \beta)$, $\hat{\theta}_n = (\hat{\beta}_n^T, \hat{b}_{1n}^T, \hat{b}_{0n}^T)^T$, $\theta = (\beta^T, b_1^T, b_0^T)^T \in \Theta = \mathcal{B} \times B_1 \times B_0$, $\mathcal{B} \subset R^q$, $b_j = (b_{j1}, \ldots, b_{jK})^T \in B_j \subset R^K$ for $j = 0, 1$, and $\rho_{\tau_k}(x) = \tau_k x - x I(x \leq 0)$ are the check loss functions, for $k = 1, \ldots, K$.

With $\hat{\beta}_n$, we can define $\hat{\varepsilon}_i = D_i \{T_i - \mu(X_i, 1, \hat{\beta}_n)\} + (1 - D_i)\{T_i - \mu(X_i, 0, \hat{\beta}_n)\}$ for $i = 1, \ldots, n$. Based on the $\hat{\varepsilon}_i$'s, the empirical survival functions for S_1 and S_0 are $\hat{S}_1(\varepsilon) = \{\sum_{i=1}^n D_i\}^{-1} \sum_{i=1}^n D_i I(\hat{\varepsilon}_i > \varepsilon)$ and $\hat{S}_0(\varepsilon) = \{\sum_{i=1}^n (1 - D_i)\}^{-1} \sum_{i=1}^n (1 - D_i) I(\hat{\varepsilon}_i > \varepsilon)$, respectively. Define further $\hat{S}_0^{-1}(t) = \inf\{y, \hat{S}_0(y) < t\}$. The proposed estimator for $R_x(t)$ is

$$\hat{R}_x(t) = \hat{S}_1\{\hat{S}_0^{-1}(t) + \mu(x, 0, \hat{\beta}_n) - \mu(x, 1, \hat{\beta}_n)\}. \tag{3}$$

2.3 Asymptotic results

In this section, we present the main theoretical results. All the technical details are deferred to the Appendix. Throughout the rest of the paper, we write $\pi = \text{pr}(\text{D} = 1)$, $c_{1K} = \sum_{k=1}^K f_1(b^*_{1k})$, $c_{0K} = \sum_{k=1}^K f_0(b^*_{0k})$, and $c_K = 1_K^T \Sigma 1_K$, where 1_K denotes the K-vector of ones, $\Sigma = (\sigma_{kl})_{1 \leq k,l \leq K}$ with $\sigma_{kl} = \tau_k \wedge \tau_l - \tau_k \tau_l$, for $k, l = 1, \ldots, K$, and f_j denotes the probability density function of ε_j, for $j = 0, 1$. Furthermore, we define $\mu_1 = E\{D \dot{\mu}_\beta(X, 1, \beta^*)\}$, $\mu_0 = E\{(1 - D)\dot{\mu}_\beta(X, 0, \beta^*)\}$, $C_1 = E[D\{\dot{\mu}_\beta(X, 1, \beta^*)\}^{\otimes 2}]$ and $C_0 = E[(1 - D)\{\dot{\mu}_\beta(X, 0, \beta^*)\}^{\otimes 2}]$, where $\dot{\mu}_\beta = \partial \mu / \partial \beta$, and $a^{\otimes 2} = aa^T$ for a vector a.

We first summarize the key results. Theorem 1 indicates that the asymptotic covariance matrix of $n^{1/2}(\hat{\beta}_n - \beta^*)$ has a sandwich form. This holds for $n^{1/2}(\hat{\beta}_n^{\text{LS}} - \beta^*)$, as described in Remark 1, which also discusses the asymptotic relative efficiency of the two estimators. By Theorem 2, the limiting distribution of $n^{1/2}\{\hat{R}_x(t) - R_x(t)\}$ depends on $\hat{\beta}_n$ only through the asymptotic covariance matrix of $n^{1/2}(\hat{\beta}_n - \beta^*)$. Inspection of the proof shows that a corresponding result holds for the least squares type estimator. This facilitates an analytic efficiency comparison between these two estimators. Theorem 3 shows that, for some location models, the least squares method and Pepe [10]'s quasilikelihood method produce identical estimates for $R_x(t)$. For a large class of models, the proposed method is a competitive alternative to the quasilikelihood-based method for estimating $R_x(t)$. We assume that the following conditions hold:

Assumption 1 Write $\pi = \text{pr}(\text{D} = 1)$, then $0 < \pi < 1$.

Assumption 2 The parameter space $\Theta = \mathcal{B} \times B_1 \times B_0 \subset R^{q+2K}$ is compact, and the true value $\theta^* = (\beta^*, b_1^*, b_0^*)$ is an interior point of Θ.

Assumption 3 For $j = 0, 1$, the distribution function F_j of ε_j is absolutely continuous, with the continuous densities f_j being uniformly bounded away from 0 and ∞ on each compact subset of the interior of the support of f_j.

Assumption 4 (i) The function $\mu(x, d, \beta)$ is continuous in (x, d) for each β, and is differentiable at β^* for each (x, d), with derivative $\dot{\mu}_\beta(x, d)$ such that the matrix $E\{\dot{\mu}_\beta(X, D)^{\otimes 2}\}$ is positive definite. (ii) For every β_1 and β_2 in $\mathcal{B}$, there exists a measurable function $U(x, d)$ with $E\{U(X, D)^2\} < \infty$, such that $|\mu(x, d, \beta_1) - \mu(x, d, \beta_2)| \leq U(x, d)\|\beta_1 - \beta_2\|$. (iii) There exist a constant k_0 such that for every β_1 and β_2 in $\mathcal{B}$, the inequality $[n^{-1}\sum_{i=1}^n\{\mu(X_i, D_i, \beta_1) - \mu(X_i, D_i, \beta_2)\}^2]^{1/2} \geq k_0\|\beta_1 - \beta_2\|$ holds on a set Ω_n such that $\mathrm{pr}(\Omega_\mathrm{n}) \to 1$ as n tends to infinity. (iv) The class of functions $(y, x, d) \mapsto I\{y - \mu(x, d, \beta) > \omega\}$, for β and ω in some neighborhood of the associated true values, is Donsker.

THEOREM 1 Suppose that Assumptions 1–4 hold. Then the sequence $n^{1/2}(\hat{\beta}_n - \beta^*)$ is asymptotically normal with mean zero and covariance matrix $\Gamma^{-1}\Delta\Gamma^{-1}$, where $\Gamma = c_{1K}(C_1 - \pi^{-1}\mu_1^{\otimes 2}) + c_{0K}\{C_0 - (1-\pi)^{-1}\mu_0^{\otimes 2}\}$ and $\Delta = c_K(C_1 - \pi^{-1}\mu_1^{\otimes 2}) + c_K\{C_0 - (1-\pi)^{-1}\mu_0^{\otimes 2}\}$.

THEOREM 2 Suppose that Assumptions 1–4 hold. Then, for each fixed $t \in (0, 1)$, and a fixed covariate value x, the sequence $n^{1/2}\{\hat{R}_x(t) - R_x(t)\}$ is asymptotically normal with mean zero and variance

$$\pi^{-1}S_1(w^*)\{1 - S_1(w^*)\} + \frac{t(1-t)f_1^2(w^*)}{(1-\pi)f_0^2\{S_0^{-1}(t)\}} + f_1^2(w^*)\nabla_x^T\Gamma^{-1}\Delta\Gamma^{-1}\nabla_x,$$

where $w^* = S_0^{-1}(t) + \mu(x, 0, \beta^*) - \mu(x, 1, \beta^*)$, and $\nabla_x = \{\dot{\mu}_\beta(x, 1, \beta^*) - \dot{\mu}_\beta(x, 0, \beta^*)\} - \{\pi^{-1}\mu_1 - (1-\pi)^{-1}\mu_0\}$.

Remark 1 Suppose that Assumptions 1, 2 and 4 hold, and additionally, the second moments of ε_1 and ε_0 exist. Then, it can be shown that the sequence $n^{1/2}(\hat{\beta}_n^{\mathrm{LS}} - \beta^*)$ is asymptotically normal with mean zero and covariance matrix $\tilde{\Gamma}^{-1}\tilde{\Delta}\tilde{\Gamma}^{-1}$, where $\tilde{\Gamma} = (C_1 - \pi^{-1}\mu_1^{\otimes 2}) + \{C_0 - (1-\pi)^{-1}\mu_0^{\otimes 2}\}$, and $\tilde{\Delta} = \sigma_1^2(C_1 - \pi^{-1}\mu_1^{\otimes 2}) + \sigma_0^2\{C_0 - (1-\pi)^{-1}\mu_0^{\otimes 2}\}$, with $\sigma_1^2 = \mathrm{var}(\varepsilon_1)$ and $\sigma_0^2 = \mathrm{var}(\varepsilon_0)$. In particular, if ε_1 and ε_0 are identically distributed, then $c_{1K} = c_{0K}, \sigma_1^2 = \sigma_0^2$, and the asymptotic relative efficiency of $\hat{\beta}_n$ versus $\hat{\beta}_n^{\mathrm{LS}}$ is $\sigma_1^2 c_{1K}^2/c_K$, which is independent of the specific location model, and is larger than 0.7 for large enough K [15]. This suggests that, for a large class of models, the proposed method provides a safe alternative to the least squares type method in terms of estimating the regression parameters.

THEOREM 3 If $\tilde{\mu}(x, d, \alpha)$ in model (1) can be written as $\alpha_1 + \alpha_2 D + \nu(z^T\alpha_3)$, where $z = (x^T, dx^T)^T$, α_3 is the last $p - 2$ components of α, and ν is a fixed continuously differentiable function, then the least squares type method and the quasilikelihood-based method produce identical estimates for $R_x(t)$ for each x and $t \in (0, 1)$.

References

[1] CAI, T. & PEPE, M. S. (2002). Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *J. Am. Stat. Assoc.* **97**, 1099–1107.

[2] Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician* **52**, 179–192.

[3] González-Manteiga, W., Pardo-Fernández, J. C. & Van Keilegom, I. (2011). ROC curves in nonparametric location-scale regression models. *Scand. J. Statist* **38**, 169–184.

[4] Heagerty, P. J. & Pepe, M. S. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Applied Statistics* **48**, 533–551.

[5] Knight, K. (1998). Limiting distributions for l_1 regression estimators under general conditions. *Ann. Statist.* **26** 755–770.

[6] Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, 33–50.

[7] Koenker, R. (2005). *Quantile Regression.* New York: Cambridge University Press.

[8] Liu, D. P. & Zhou, X. H. (2011). Nonparametric estimation of the covariate-specific ROC curve in presence of ignorable verification bias. *Biometrics* **67**, 906–916.

[9] Pepe, M. S. (1997). A Regression modelling framework for receiver operating characteristic curves in medical diagnostic Testing. *Biometrika* **84**, 595–608.

[10] Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* **54**, 124–135.

[11] Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* **56**, 352–359.

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

[13] Zheng, Y. & Heagerty, P. J. (2004). Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* **5**, 615–632.

[14] Zhou, X. H., Obuchowski, N. A. & McClish, D. K. (2011). *Statistical Methods in Diagnostic Medicine, 2nd ed.* New York: Wiley.

[15] Zou, H. & Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36**, 1108–1126.

[16] van der Vaart, A. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* New York: Springer.

[17] van der Vaart, A. (2000). *Asymptotic Statistics.* New York: Cambridge University Press.

[18] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.