Published in final edited form as:
J Am Stat Assoc. 2019 ; 114(528): 1815–1825. doi:10.1080/01621459.2018.1515079.

Prediction Accuracy Measures for a Nonlinear Model and for Right-Censored Time-to-Event Data

Gang Li[a], **Xiaoyan Wang**[b]

[a]Departments of Biostatistics and Biomathematics, University of California, Los Angeles, CA

[b]Division of General Internal Medicine and Health Services Research, University of California, Los Angeles, CA

Abstract

This article develops a pair of new prediction summary measures for a nonlinear prediction function with right-censored time-to-event data. The first measure, defined as the proportion of explained variance by a linearly corrected prediction function, quantifies the potential predictive power of the nonlinear prediction function. The second measure, defined as the proportion of explained prediction error by its corrected prediction function, gauges the closeness of the prediction function to its corrected version and serves as a supplementary measure to indicate (by a value less than 1) whether the correction is needed to fulfill its potential predictive power and quantify how much prediction error reduction can be realized with the correction. The two measures together provide a complete summary of the predictive accuracy of the nonlinear prediction function. We motivate these measures by first establishing a variance decomposition and a prediction error decomposition at the population level and then deriving uncensored and censored sample versions of these decompositions. We note that for the least square prediction function under the linear model with no censoring, the first measure reduces to the classical coefficient of determination and the second measure degenerates to 1. We show that the sample measures are consistent estimators of their population counterparts and conduct extensive simulations to investigate their finite sample properties. A real data illustration is provided using the PBC data. Supplementary materials for this article are available online. An R package PAmeasures has been developed and made available via the CRAN R library. Supplementary materials for this article are available online.

Keywords

Censoring; Coefficient of determination; Cox's proportional hazards model; Explained prediction error; Explained variance

CONTACT Gang Li vli@ucla.edu Departments of Biostatistics and Biomathematics, University of California, Los Angeles, CA 90095-1772.

Supplementary Materials
Appendix: Lemmas, proofs of the theorems, and additional simulation results.

1. Introduction

In this article, we study prediction accuracy measures for a nonlinear prediction function based on a possibly misspecified nonlinear model with right-censored time-to-event data. By far, the most commonly used prediction accuracy measure for a linear model is the R^2 statistic, or coefficient of determination. Let Y be a real-valued random variable and X be a vector of p real-valued explanatory random variables or covariates. Assume that one observes a random sample $(Y_1, X_1),...,(Y_n, X_n)$ from the distribution of (Y, X). The R^2 statistic is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}, \quad (1)$$

where $\hat{Y}_i = a + b^T X_i$ is the least squares predicted value for subject i. The R^2 statistic has the straightforward interpretation as the proportion of variation of Y, which is explained by the least squares prediction function due to the following variance decomposition:

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2. \quad (2)$$

$$\text{total variation} = \text{explained variation} + \text{unexplained variation}$$

Despite its popularity in linear regression, the above R^2 statistic is not readily applicable to a nonlinear model since the decomposition (2) no longer holds. In the past decades, much efforts have been devoted to extending the R^2 statistic to nonlinear models. Among others, the pseudo, R^2 statistics for a nonlinear model include likelihood-based measures (Goodman 1971; McFadden et al. 1973; Maddala 1986; CoxandSnell 1989; Magee 1990; Nagelkerke 1991), information-based measures (McFadden et al. 1973; Kent 1983), ranking-based measures (Harrell et al. 1982), variation-based measures (Theil 1970; Efron 1978; Haberman 1982; Hilden 1991; Cox and Wermuth 1992; Ash and Shwartz 1999), and the multiple correlation coefficient measure (Mittlböck and Schemper 1996; Zheng and Agresti 2000). However, none of the existing pseudo R^2 measures are motivated directly from a variance decomposition and none have received the same widespread acceptance as the classical R^2 for linear regression. Interested readers are referred to Zheng and Agresti (2000) for an excellent survey of existing pseudo R^2 measures and further references.

In this article, we first develop a pair of new prediction accuracy measures for a nonlinear prediction function of an individual response. We begin with defining population prediction accuracy measures. Based on a variance decomposition, we define a ρ^2 measure as the proportion of the explained variance of Y by a corrected prediction function, which is shown to coincide with the squared multiple correlation coefficient between the response and the predicted response. Because it describes the proportion of the explained variance by its corrected prediction function, ρ^2 measures the *potential* predictive power of the predictive function and thus by itself is not sufficient to summarize the prediction accuracy of the prediction function. As a remedy, we introduce another parameter, λ^2, defined as the proportion of prediction error explained by the corrected prediction function based on a prediction error decomposition, to measure how close the prediction function is to its corrected version and quantifies how much prediction error reduction is realized with the correction. The two parameters capture complementary information regarding the predictive accuracy of the prediction function and provide a complete summary of its predictive power. We further develop sample versions of the variance and prediction error decompositions based on uncensored data, define the corresponding sample prediction accuracy measures, namely R^2 and L^2, and establish their asymptotic properties. It is worth noting that for the least squares prediction function under the linear model, L^2 always degenerates to 1 and thus only R^2 is needed to describe its predictive accuracy.

We further extend the proposed prediction accuracy measures to event time models with right-censored time-to-event data. Note that even for the linear model, it is not clear how to extend the R^2 statistic defined in (1) to right-censored data since some of the Y values are not observed. A variety of pseudo R^2 measures and other loss functions have been proposed for event time models with right-censored data (Kent and O'Quigley 1988; Korn and Simon 1990; Graf et al. 1999; Schemper and Henderson 2000; Royston and Sauerbrei 2004; O'Quigley, Xu, and Stare 2005; Stare, Perme, and Henderson 2011). For example, the EV option in the SAS PHREG procedure gives a generalized R^2 measure proposed by Schemper and Henderson (2000) for Cox's (1972) proportional hazards model. Amore recent proposal by Stare, Perme, and Henderson (2011) uses explained rank information, which is applicable to a wide range of event time models. Stare, Perme, and Henderson (2011) also gave a thorough literature review of prediction accuracy measures for event time models. We highlight that for linear regression, none of the existing pseudo R^2 measures for right-censored data reduce to the classical R^2 statistic in the absence of censoring. Moreover, under a correctly specified model, they do not converge to the nonparametric population R^2 value $\rho_{NP}^2 \equiv \mathrm{var}(E(Y\mid X))/\mathrm{var}(Y)$, the proportion of the explained variance by $E(Y\mid X)$, as the sample size grows large. Finally, as illustrated in Section 4 (Table 1), the pseudo R^2 measures of Schemper and Henderson (2000) and Stare, Perme, and Henderson(2011) may fail to distinguish between Cox's models with the same regression coefficients but different baseline hazards: they could remain constant for different Cox's models as ρ_{NP}^2 varies from 0 to 1. In Section 3, we propose right-censored sample versions of R^2 and L^2 by deriving a variance decomposition and a prediction error decomposition for right-censored data and show that they are consistent estimators of the population parameters ρ^2 and λ^2. The proposed measures possess multiple appealing properties that most existing pseudo R^2 measures do not have. First, for the linear model with no censoring, our R^2 statistic reduces to the classical coefficient of determination and L^2 degenerates to 1. Second, when the prediction function is the conditional mean response based on a correctly specified model, our R^2 statistic is a consistent estimate of the nonparametric coefficient of determination ρ_{NP}^2, and L^2 converges to 1 as the sample size grows to infinity. Third, our method is applicable to a variety of event time models including the Cox proportional hazards model, accelerated failure time models, additive risk models, threshold regression model, proportional odds model, and transformation models, with time-independent covariates and independently right-censored data. Fourth, our measures are defined without requiring the prediction model to be correctly specified. Finally, the proposed R^2 statistic provides a natural benchmark to compare the potential prediction power between prediction models that are not necessarily nested or correctly specified as illustrated in the real data example in Section 5.

The rest of the article is organized as follows. In Section 2.1, we define a pair of population prediction accuracy measures for a nonlinear prediction function by deriving a variance decomposition and a prediction error decomposition. Sample measures based on independent and identically distributed complete data are then proposed and studied in Section 2.2. Section 3 extends these measures to event time models with right-censored data. Section 4 presents simulations to illustrate potential weaknesses of some existing pseudo R^2 proposals for right-censored data and investigate the finite sample performance of the proposed measures. A real data example is given in Section 5. Further remarks are provided in Section 6. Additional lemmas, all theoretical proofs, and more numerical results are collected in the supplementary materials.

2. Prediction Accuracy Measures for a Nonlinear Model

Denote by $F(y|x) = P(Y \le y|X = x)$ the true conditional distribution function of Y given $X = x$. Consider a regression model of Y on X described by a family of conditional distribution functions $\mathscr{M} = \{F_\theta(y|x) : \theta \in \Theta\}$, where θ is either finite or infinite dimensional. For example, $F_\theta(y|x) = \Phi(y - \alpha - \beta^T x)/\sigma)$ for the linear regression model with a normal $N(0, \sigma^2)$ error, where $\theta = (\alpha, \beta^T, \sigma^2)$ and Φ is the standard normal cumulative distribution function. For the Cox (1972) proportional hazards model, $F_\theta(y|x) = 1 - \{1 - F_0(y)\}^{\exp(\beta^T x)}$ where $\theta = (\beta, F_0)$ consists of a finite dimensional regression parameter β and an infinite dimensional unknown baseline distribution function F_0. Model $\mathscr{M}$ is said to be misspecified if it does not include the true conditional distribution function $F(y|x)$ as a member.

For any $\theta \in \Theta$, let $m_\theta(X)$ be a prediction function of Y obtained as a functional of $F_\theta(\cdot|X)$. Common examples of $m_\theta(X)$ include the conditional mean response $m_\theta(x) = \int y dF_\theta(y|x)$ and the conditional median response $m_\theta(x) = F_\theta^{-1}(0.5|x)$. Assume that $\hat{\theta}$ is a sample statistic that converges to a limit θ^* as n grows to ∞. As discussed in the supplementary materials (Appendix A.1) θ^* is typically the true parameter value under a correctly specified model, and the parameter value that minimizes the Kullback-Leibler information criterion under a misspecified model.

Below, we first develop population prediction accuracy measures for $m_{\theta^*}(X)$, which can be regarded as the asymptotic accuracy measures for the predictive power of $m_{\hat{\theta}}(x)$. Their sample versions for $m_{\hat{\theta}}(X)$ will then be derived in a similar fashion.

2.1. Population Prediction Accuracy Measures

For any p-variate function $P(x)$, define $\text{MSPE}(P(X)) = E\{Y - P(X)\}^2$. as the *mean squared prediction error* of $P(X)$ for predicting Y. In general, it would be desirable for a prediction function $P(X)$ of Y to possess at least the following properties: (i) $E\{P(X)\} = \mu_Y$ and (ii) MSPE $(P(X)) \le \text{MSPE}(\mu_Y)$, where $\mu_Y = E(Y)$ is the best prediction among all constant (non-informative) predictions of Y as measured by MSPE. However, such minimal requirements are not always met by $m_{\theta*}(X)$ when the model $\mathscr{M}$ is possibly misspecified or when the prediction is not based on the conditional mean response. Below, we introduce a linearly corrected prediction function, which always meets the above requirements (i) and (ii) and is pivotal to assessing the predictive accuracy of $m_{\theta*}(X)$.

Definition 2.1.—The linearly corrected prediction function of $m_{\theta*}(X)$ is defined as

$$m_{\theta*}^{(c)}(X) = \tilde{a} + \tilde{b} m_{\theta*}(X) = \mu_Y + \frac{\text{cov}(Y, m_{\theta*}(X))}{\text{var}(m_{\theta*}(X))}[m_{\theta*}(X) - E\{m_{\theta*}(X)\}], \quad (3)$$

where $(\tilde{a}, \tilde{b}) = \text{argmin}_{\alpha, \beta} E\{Y - (\alpha + \beta m_{\theta*}(X))\}^2$.

It is easy to see that $m_{\theta*}^{(c)}(X)$ satisfies the above mentioned requirements (i) and (ii) and that $MPSE\left(m_{\theta*}^{(c)}(X)\right) \le MPSE(m_{\theta*}(X))$. More importantly, by Lemma A.1 (supplementary materials), $m_{\theta*}^{(c)}(X)$ facilitates the following variance and prediction error decompositions:

$$\text{var}(Y) = E\left\{m_{\theta*}^{(c)}(X) - \mu_Y\right\}^2 + E\left\{Y - m_{\theta*}^{(c)}(X)\right\}^2, \quad (4)$$

and

$$\text{MSPE}(m_{\theta*}(X)) = E\left\{Y - m_{\theta*}^{(c)}(X)\right\}^2 + E\left\{m_{\theta*}^{(c)}(X) - m_{\theta*}(X)\right\}^2, \quad (5)$$

which lead to the following prediction accuracy measures for $m_{\theta*}(X)$.

Definition 2.2.—Define

$$\rho_{m_{\theta*}}^2 = 1 - \frac{E\left\{Y - m_{\theta*}^{(c)}(X)\right\}^2}{\text{var}(Y)} = \frac{E\left\{m_{\theta*}^{(c)}(X) - \mu_Y\right\}^2}{\text{var}(Y)}, \quad (6)$$

to be the proportion of the variance of Y, that is, explained by $m_{\theta*}^{(c)}(X)$, and

$$\lambda_{m_{\theta}*}^2 = \frac{\text{MSPE}(m_{\theta*}^{(c)}(X))}{\text{MSPE}(m_{\theta*}(X))} = 1 - \frac{E\left\{m_{\theta*}^{(c)}(X) - m_{\theta*}(X)\right\}^2}{\text{MSPE}(m_{\theta*}(X))}. \quad (7)$$

to be the proportion of the MSPE of $m_{\theta*}(X)$, that is, explained by $m_{\theta*}^{(c)}(X)$.

Remark 2.1 (Interpretation of $\rho^2_{m_{\theta *}}$ and $\lambda^2_{m_{\theta *}}$).—Define the L_2-distance between any two real-valued random variables ξ and η by $d_2(\xi, \eta) = \left\{E(\xi - \eta)^2\right\}^{\frac{1}{2}}$. Figure 1 depicts the geometric relationships between Y, μ_Y, $m_{\theta *}(X)$, $m^{(c)}_{\theta *}(X)$, and $E(Y \mid X)$, where $\mathscr{P}(X)$ denotes the space of all real-valued functions of X, $m^{(c)}_{\theta *}(X)$ is the projection of Y onto the subspace of all linear functions of $m_{\theta *}(X)$, and $E(Y|X)$ is the projection of Y onto $\mathscr{P}(X)$.

It is clear from Figure 1 that the variance decomposition (4) corresponds to the Pythagorean theorem for the triangle $\left(Y, m^{(c)}_{\theta *}(X), \mu_Y\right)$, which leads to the definition of $\rho^2_{m_\theta *}$, and that the prediction error decomposition (5) is the Pythagorean theorem for $(Y, m^{(c)}_{\theta *}(X), m_{\theta *}(X))$, which defines $\lambda^2_{m_{\theta *}}$. Therefore, $\rho^2_{m_\theta *}$ and $\lambda^2_{m_{\theta *}}$ provide distinct, yet complementary information regarding the prediction accuracy of $m_{\theta *}(X)$: $\rho^2_{m_{\theta *}}$ measures its potential predictive power through its corrected version $m^{(c)}_{\theta *}(X)$, whereas $\lambda^2_{m_{\theta *}}$ measures its closeness to $m^{(c)}_{\theta *}(X)$ and quantifies how much prediction error reduction can be achieved with the correction. Together, they provide a complete summary of the predictive accuracy of $m_{\theta *}(X)$. In practice, $\rho^2_{m_\theta *}$ should be used as the primary measure for the potential predictive power of $m_{\theta *}(X)$, whereas $\lambda^2_{m_\theta *}$ should be used as a supplementary measure to indicate (by a value less than 1) if a linear correction is required for $m_{\theta *}(X)$ to achieve its potential predictive power and how much prediction error reduction can be realized with the correction. Finally, L^2 is not to be confused as a lack-of-fit measure for model $\mathscr{F}$. Although $m_{\theta *}(X) = E(Y \mid X)$ implies $\lambda^2_{m_{\theta *}} = 1$, $\lambda^2_{m_{\theta *}}$ may also be 1 even if $m_{\theta *}(X) \neq E(Y \mid X)$, as long as $m^{(c)}_{\theta *}(X) = m_{\theta *}(X)$. Hence, $\lambda^2_{m_{\theta *}} = 1$ simplyindicatesthat no linear correction is required for $m_{\theta *}(X)$ to achieve it potential predictive power and does not necessarily imply that the model is correctly specified. This point is further illustrated by our simulation results in the supplementary materials (Appendix A.2.3, Figures A.4 and A.5: first row).

It is also seen from Figure 1 that the Pythagorean theorem for the triangle $(Y, E(Y \mid X), \mu_Y)$ corresponds to the following well-known variance decomposition

$$\mathrm{var}(Y) = \mathrm{var}(E(Y \mid X)) + E(\mathrm{var}(Y \mid X)) .$$

We refer to

$$\rho^2_{NP} \equiv 1 - \frac{E(Y - E(Y \mid X))^2}{\mathrm{var}(Y)} = \frac{\mathrm{var}(\mu(X))}{\mathrm{var}(Y)}, \tag{8}$$

the proportion of explained variance by $E(Y/X)$, as the *non-parametric coefficient of determination.* Note that $\sqrt{\rho^2_{NP}}$ is the "correlation ratio" studied previously by Renyi (1959).

The next theorem summarizes some fundamental properties of $\rho^2_{m_{\theta^*}}$ and $\lambda^2_{m_{\theta^*}}$

Theorem 2.1.

a. Let $\rho(\xi, \eta)$ denote the correlation coefficient between two random variables ξ and η. Then, $\rho^2_{m_{\theta^*}} = [\rho(Y, m_{\theta^*}(X))]^2$;

b. (Linear prediction). Let BLUE(X) = $a + b^T X$ be the best linear unbiased estimator (BLUE) of Y, where $\left(a, b\right) = \arg\min_{\alpha,\beta} E\left\{Y - \left(\alpha + \beta^T X\right)\right\}^2$. Then, (i) $\mathrm{BLUE}^{(c)}(X) = \mathrm{BLUE}(X)$, (ii) $\lambda^2_{\mathrm{BLUE}} \equiv 1$, and (iii) ρ^2_{BLUE} is equal to the population value of the classical coefficient of determination for linear regression;

c. If $m_{\theta^*}(X) = E(Y \mid X)$, then $\lambda^2_{m_{\theta^*}} \equiv 1$, and $\rho^2_{m_{\theta^*}} = \rho^2_{NP}$, where ρ^2_{NP} is defined in (8);

d. (Maximal ρ^2) Let $P(X)$ be the space of all p-variate functions $Q(X)$ of X. Then $\rho^2_{NP} = \max_{Q \in \mathscr{P}(X)}\left\{\rho^2_Q\right\}$.

2.2. Sample Prediction Accuracy Measures

Let $(Y_1, X_1),\ldots,(Y_n, X_n)$ be a random sample of (Y, X) and $\hat\theta = \hat\theta(Y_1, X_1, \ldots, Y_n, X_n)$ be a sample statistic. We next derive sample versions of $\rho^2_{m_{\theta^*}}$ and $\lambda^2_{m_{\theta^*}}$ for $m_{\hat\theta}(X)$.

By Lemma A.2 (supplementary materials), we have the following sample version of the variance and prediction error decompositions:

$$\sum_{i=1}^{n}\left(Y_i - \bar Y\right)^2 = \sum_{i=1}^{n}\left(m^{(c)}_{\hat\theta}(X_i) - \bar Y\right)^2 + \sum_{i=1}^{n}\left(Y_i - m^{(c)}_{\hat\theta}(X_i)\right)^2, \tag{9}$$

and

$$\begin{aligned}&\sum_{i=1}^{n}(Y_i - m_{\hat\theta}(X_i))^2\\ &\quad= \sum_{i=1}^{n}\left(Y_i - m^{(c)}_{\hat\theta}(X_i)\right)^2 + \sum_{i=1}^{n}\left(m^{(c)}_{\hat\theta}(X_i) - m_{\hat\theta}(X_i)\right)^2,\end{aligned} \tag{10}$$

where $m^{(c)}_{\hat\theta}(x)$ is the ordinary least squares regression function obtained by linearly regressing $Y_1,\ldots, Y_n$ on $m_{\hat\theta}(X_1),\ldots,m_{\hat\theta}(X_n)$.

The sample versions of $\rho^2_{m_{\theta^*}}$ and $\lambda^2_{m_{\theta^*}}$ are therefore, defined by

$$R^2_{m_{\hat\theta}} = \frac{\sum_{i=1}^{n}\left(m^{(c)}_{\hat\theta}(X_i) - \bar Y\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar Y\right)^2}, \tag{11}$$

and

$$L_{m\hat{\theta}}^2 = \frac{\sum_{i=1}^{n}\left(Y_i - m_{\hat{\theta}}^{(c)}(X_i)\right)^2}{\sum_{i=1}^{n}\left(Y_i - m_{\hat{\theta}}(X_i)\right)^2}, \tag{12}$$

where $R_{m\hat{\theta}}^2$ is the proportion of variation of Y explained by $m_{\hat{\theta}}^{(c)}(X)$ and $L_{m\hat{\theta}}^2$, is the proportion of prediction error of $m_{\hat{\theta}}(X)$ explained by $m_{\hat{\theta}}^{(c)}(X)$.

Remark 2.2.—Similar to Theorem 2.1(a), it can be shown that $R_{m\hat{\theta}}^2 = \{r(Y, m_{\hat{\theta}}(X))\}^2$, where $r(Y, m_{\hat{\theta}}(X))$ is the Pearson correlation coefficient between Y and $m_{\hat{\theta}}(X)$. Furthermore, if $m_{\hat{\theta}}(x)$ is the fitted least squares regression line from a linear model, then $L_{m\hat{\theta}}^2 \equiv 1$ and $R_{m\hat{\theta}}^2$ is identical to the classical coefficient of determination for the linear model.

Below, we give the asymptotic properties of $R_{m\hat{\theta}}^2$, and $L_{m\hat{\theta}}^2$.

Theorem 2.2.—Assume conditions (C2)-(C4) of supplementary materials (Appendix A.1) hold. Then, as $n \to \infty$

a. (Consistency) $R_{m\hat{\theta}}^2 \xrightarrow{P} \rho_{m\theta*}^2$, and $L_{m\hat{\theta}}^2 \xrightarrow{P} \lambda_{m\theta*}^2$;

b. (Asymptotic normality) $\sqrt{n}\left(R_{m\hat{\theta}}^2 - \rho_{m\theta*}^2\right) \xrightarrow{d} N\left(0, \sigma_\rho^2\right)$, and $\sqrt{n}\left(L_{m\hat{\theta}}^2 - \lambda_{m\theta*}^2\right) \xrightarrow{d} N\left(0, \sigma_\lambda^2\right)$, where σ_ρ^2 and σ_λ^2 are the asymptotic variances.

The asymptotic results allow one to assess the variability of the sample measures $R_{m\hat{\theta}}^2$ and $L_{m\hat{\theta}}^2$ and obtain confidence interval estimates for the corresponding population parameters. In practice, the bootstrap method (Efron and Tibshirani 1994) or a transformation-based method would be more appealing than the normal approximation method because the sampling distributions of $R_{m\hat{\theta}}^2$ and $L_{m\hat{\theta}}^2$ can be skewed, especially near 0 and 1.

3. Sample Prediction Accuracy Measures for Right-Censored Data

In this section, we extend the sample measures $R_{m\hat{\theta}}^2$ and $L_{m\hat{\theta}}^2$ defined by (11) and (12) to right-censored time-to-event data. Let $T = \min\{Y, C\}$ and $\delta = I(Y \le C)$, where C is an censoring random variable. Assume that one observes a right-censored sample of n independent and identically distributed triplets $(T_1, \delta_1, X_1),\ldots,(T_n, \delta_n, X_n)$ from the distribution of (T, δ, X).

Assume that $\hat{\theta} = \hat{\theta}(T_1, \delta_1, X_1, \ldots, T_n, \delta_n, X_n)$ is a sample statistic. The sample prediction accuracy measures defined in (11) and (12) are no longer directly applicable to right-censored data because Y is not observed on some subjects. In Lemma A.3 (supplementary materials), we show that

$$\sum_{i=1}^{n} w_i\left\{T_i - \overline{T}^{(w)}\right\}^2 = \sum_{i=1}^{n} w_i\left\{m_{\hat{\theta}}^{(wc)}(X_i) - \overline{T}^{(w)}\right\}^2 + \sum_{i=1}^{n} w_i\left\{T_i - m_{\hat{\theta}}^{(wc)}(X_i)\right\}^2, \tag{13}$$

and

$$\sum_{i=1}^{n} w_i\{T_i - m_{\hat{\theta}}(X_i)\}^2 = \sum_{i=1}^{n} w_i\left\{T_i - m_{\hat{\theta}}^{(wc)}(X_i)\right\}^2 + \sum_{i=1}^{n} w_i\left\{m_{\hat{\theta}}^{(wc)}(X_i) - m_{\hat{\theta}}(X_i)\right\}^2, \tag{14}$$

for any set of nonnegative weights $w_1,\ldots,w_n$ satisfying $\sum_{i=1}^{n} w_i = 1$, where $m_{\hat{\theta}}^{(wc)}(x)$ is the fitted regression function from the weighted least squares linear regression of $Y_1,\ldots,Y_n$ on $m_{\hat{\theta}}(X_1),\ldots,m_{\hat{\theta}}(X_n)$ with weight $W = \mathrm{diag}\{w_1,\ldots,w_n\}$. Furthermore, in Lemma A.4 (supplementary materials), we will show that if

$$w_i = \frac{\dfrac{\delta_i}{\widehat{G}(T_i-)}}{\sum_{j=1}^{n}\dfrac{\delta_j}{\widehat{G}(T_j-)}}, \quad i = 1,\ldots,n, \tag{15}$$

where $\widehat{G}$ is the Kaplan-Meier (Kaplan and Meier 1958)estimate of $G(c) = P(C > c)$, then (13) and (14) can be regarded as the right-censored data analogs of the uncensored sample variance decomposition (9) and prediction error decomposition (10), respectively. These results lead to the following right-censored sample prediction accuracy measures.

Definition 3.1.

The right-censored sample versions of $\rho_{m\theta*}^2$ and $\lambda_{m\theta*}^2$ are defined by

$$R_{m\hat{\theta}}^2 = \frac{\sum_{i=1}^{n} w_i\left\{m_{\hat{\theta}}^{(wc)}(X_i) - \overline{T}^{(w)}\right\}^2}{\sum_{i=1}^{n} w_i\left\{T_i - \overline{T}^{(w)}\right\}^2}, \tag{16}$$

and

$$L_{m\hat{\theta}}^2 = \frac{\sum_{i=1}^{n} w_i\left\{T_i - m_{\hat{\theta}}^{(wc)}(X_i)\right\}^2}{\sum_{i=1}^{n} w_i\{T_i - m_{\hat{\theta}}(X_i)\}^2}, \tag{17}$$

where the weight w_i's are defined in (15).

The above defined $R^2_{m\hat{\theta}}$ can be interpreted as an approximate proportion of sample variance of Y explained by $m^{(wc)}_{\hat{\theta}}(X)$ and $L^2_{m\hat{\theta}}$ an approximate proportion of sample mean squared prediction error of $m_{\hat{\theta}}(X)$ explained by $m^{(wc)}_{\hat{\theta}}(X)$. By definition, $0 \le R^2_{m\hat{\theta}} \le 1$ and $0 \le L^2_{m\hat{\theta}} \le 1$.

Theorem 3.1.

a. (Uncensored data). If there is no censoring, then formulas (16) and (17) reduce to the uncensored data definitions (11) and (12), respectively.

b. (Consistency). Assume conditions (C1)-(C5) hold. Then, as $n \to \infty$, $R^2_{m\hat{\theta}} \xrightarrow{P} \rho^2_{m\theta*}$, and $L^2_{m\hat{\theta}} \xrightarrow{P} \lambda^2_{m\theta*}$.

c. (Asymptotic normality). Assume conditions (C1)-(C5) hold.Then, $\sqrt{n}\left(R^2_{m\hat{\theta}} - \rho^2_{m\theta*}\right) \xrightarrow{d} N\left(0, \upsilon^2_{\rho}\right)$ and $\sqrt{n}\left(L^2_{m\hat{\theta}} - \lambda^2_{m\theta*}\right) \xrightarrow{d} N\left(0, \upsilon^2_{\lambda}\right)$, as $n \to \infty$, where υ^2_{ρ} and υ^2_{λ} are the asymptotic variances.

Remark 3.1.

Theorem 3.1 (b) and (c) are derived under condition (C1) of Appendix A.1 that C is independent of X and Y. In the next section, we demonstrate by simulation that the $R^2_{m\hat{\theta}}$ and $L^2_{m\hat{\theta}}$ measures are quite robust even if C depends the covariates X, unless the model is severely misspecified.

4. Simulations

We present several simulation studies to investigate the finite sample properties of the proposed prediction accuracy measures.

Simulation 1:

In this simulation, we use the population ρ^2_{NP} defined by (8) as a benchmark to illustrate the properties and potential weaknesses of two existing R^2-type measures R^2_{SH} and R^2_{SPH} for the Cox model proposed by Schemper and Henderson (2000) and Stare, Perme, and Henderson (2011), respectively. In the simulation, the event time Y is generated from a Cox proportional hazard model: $Y = H_0^{-1}\left[-\log(U) \times \exp\left(-\beta^T X\right)\right]$, where $U \sim$ U(0,1), $H_0^{-1}(t) = 2t^{\frac{1}{\upsilon}}$ is the inverse function of a Weibull cumulative hazard function $H_0(t) = (0.5t)^{\upsilon}$, and $X = 10 \times$ Bernoulli(0.5). We consider nine data settings by varying β and v. We approximate the population ρ^2 value by averaging its sample R^2 values over 10 Monte Carlo samples of size $n = 5000$ with no censoring. The results are summarized by supplementary materials (Figure A.1 in Appendix A.2.1). Table 1 takes a snapshot of Figure A.1 for some selected settings.

Table 1 shows that for Cox's models with the same baseline hazard (or v) but different hazard ratio (or β) (e.g., models 3, 6, 9), R^2_{SPH} and R^2_{SH} have produced the same rankings as

ρ^2_{NP} and thus have correctly reflected the relative predictive power between these models. However, both R^2_{SPH} and R^2_{SH} have failed to distinguish the predictive power between models with the same hazard ratio (or β) but different baseline hazard (or v). For example, they remain a constant 0.49 for models 7, 8, and 9 whose ρ^2_{NP} values are 0.10, 0.33, and 0.80, respectively. This is not surprising for R^2_{SPH} because it is a measure of the explained rank variation, that is, largely determined by β. However, the predictive power of the Cox model is determined by not only β, but also v, where the latter reflects the variability of the outcome variable and is not adequately accounted for by R^2_{SPH}. The R^2_{SH} measure is observed to suffer from the same limitation although it is not as obvious to see from its definition.

Simulation 2:

This simulation investigates finite sample properties of the proposed R^2 and L^2 measures for the Cox model relative to their population values ρ^2 and λ^2 by varying the censoring rate (0%, 10%, 25%, 50%), sample size n (100,250,1000), and data generation setting (Weibull, Log-normal, Inverse Gaussian). For the Weibull setting, data are generated from a Weibull model $\log(Y) = \beta^T X + \sigma W$, where $\beta = 1$, $\sigma = 0.24$, $X \sim U(0,1)$, $W \sim$ standard extreme value distribution. For the lognormal setting, data are generated from $\log(Y) = \beta^T X + \sigma W$, where $\beta = 1$, $\sigma = 0.27$, $X \sim U(0,1)$, $W \sim \mathrm{N}(0,1)$, and $C \sim$ Weibull (shape = 1, shape = b) with b adjusted to produce a given censoring rate. For the inverse Gaussian setting, data are generated from $Y \sim$ Inverse Gaussian $\left(\text{mean} = -\frac{e^{\alpha_0 + \alpha_1 X}}{\beta_0 + \beta_1 X},\ \text{shape} = e^{2*(\alpha_0 + \alpha_1 X)}\right)$, where $\alpha_0 = 3$, $\alpha_1 = -2.5$, $\beta_0 = -1$, $\beta_1 = 0.6$, $X \sim U(0,1)$. For all three data settings, $\rho^2 = 0.50$, and censoring times are generated from $C \sim$ Weibull (shape = 1, scale = b) with b adjusted to produce a given censoring rate.

We first examine the overall fit of the Cox model under each of the three data settings by displaying in Figure 2 the Cox-Snell residual plots for the Cox model based on the first 10 Monte Carlo replications (first row: Weibull; second row: log-normal AFT; third row: inverse Gaussian) with varying sample size (first column: $n = 100$; second column: $n = 250$; third column: $n = 1000$) and censoring rate, CR = 0%. The plots reveal no misspecification of the Cox model under the Weibull setting (first row), mild model misspecification under the log-normal setting (second row), serious model misspecification under the inverse Gaussian setting (third row).

Figure 3 summarizes simulated R^2 (shaded box) and L^2 (unshaded box) values for the Cox model over 1000 Monte Carlo replications using boxplots by censoring rate (0%, 10%, 25%, 50%), sample size (100, 200, 1000), and data generation setting (upper panel: Weibull; middle panel: log-normal AFT; bottom panel: inverse Gaussian). The population values ρ^2 and λ^2 in Figure 3 are approximated by the averaged sample values over 100 Monte Carlo replications of sample size $n = 5000$ with no censoring. We observe from Figure 3 that for all three data settings, the proposed R^2 and L^2 estimate their population values well: their medians agree well with the population values and as expected, their variability increases as the censoring rate increases and decreases as the sample size increases.

In the online supplementary materials (Appendix A.2.2), the we report results from a similar simulation study for the Cox model under different scenarios with $\rho^2 = 0.2$. We also report results from similar simulation studies for the threshold regression model (Lee and Whitmore 2006) in the supplementary materials (Appendix A.2.3). All the simulations give consistent messages that R^2 and L^2 estimate their population counterparts well across all three data settings regardless of whether the model is correctly or misspecified.

We have also run more simulations under more data settings, population ρ^2 values and for other models such as accelerated failure time models. The results are all consistent with what have been discussed above and thus not reported here.

Simulation 3:

In this simulation, we study the sensitivity of the proposed R^2 and L^2 measures defined in Section 3, when the independent censoring assumption (C1) of the Appendix A.1 is perturbed. The simulation setup is similar to the second simulation except that the censoring time C is dependent on the covariate X and that Y and C are conditional independent given the covariate. Specifically, $\log(C) = \gamma_c^T X + \theta_c \times V$, where $X \sim U(0,1)$, $\theta_c = 4$, $V \sim$ extreme value distribution, and γ_c is adjusted to give a given censoring rate. Boxplots of the simulated R^2 and L^2 for the Cox model are depicted in Figure 4. We observe from Figure 4 that the violation of independent censoring assumption has little effect on the performance of the proposed R^2 and L^2 when the Cox model is correctly specified or mildly misspecified (top and middle panels). However, it results in substantial bias for R^2 and L^2 under the last data setting (bottom panel) when the Cox model is severely misspecified.

5. An Example

In this section, we illustrate the use of the proposed prediction accuracy measures on a primary biliary cirrhosis (PBC) data with 312 patients from a randomized Mayo Clinic trial in primary biliary cirrhosis of the liver conducted between 1974 and 1984 (http://astrostatistics.psu.edu/datasets/R/html/survival/html/pbc.html). For illustration purpose, we evaluate and compare the predictive power of the Cox model, the Weibull and log-normal accelerated failure time models, and the threshold regression model for predicting overall survival of individual patients with PBC, using the five covariates (patient's age, log(serum bilirubin concentration), log(serum albumin concentration), log(standardised blood clotting time), and presence of peripheral edema and antidiuretic therapy) employed in the well-known Mayo risk score (MRS) (Dickson et al. 1989).

We first examine the Cox-Snell residual plot based on the PBC data for each of the four predictions models in Figure 5 for overall lack-of-fit. In each plot, we have also overlaid additional Cox-Snell residual plots from 10 bootstrap samples to reflect the variability. Figure 5 does not indicate any serious lack-of-fit for the Cox model, Weibull and log-normal accelerated failure time models. However, it does suggest possible severe lack-of-fit for the threshold regression model, which is thus excluded from further consideration.

Note that Figure 5 only provides a model diagnostic check for lack-of-fit. It does not offer further information regarding the predictive powers of the models under consideration especially because it is difficult to interpret the right tail of a Cox-Snell plot. In fact, plots of the predicted and observed survival times versus the risk scores for the three models in Figure 6 reveal that the Weibull AFT model and the log-normal AFT model could suffer substantial systematic prediction bias for low-risk patients (or long survivors). To assess and compare their predictive powers, we report their R^2 and L^2 values in Table 2.

It is seen from Table 2 that among the three models, the Cox model stands out as the best prediction model with the highest potential predictive power $R^2 = 0.39$, which is consistent with the observation from Figure 6. To account for sampling variabilities, we also computed R^2 values for the three models based on 100 bootstrap samples and summarized the R^2 differences between the Cox model and each of the other two models using boxplots in Figure 7, which confirms the superior potential predictive power of the Cox model to the Weibull and lognormal AFT models. Finally, the Cox model has an associated L^2 value of 0.83, suggesting that a linear correction would be needed to fulfill its potential predictive power and that the linear correction would result in a reduction of $1 - L^2 = 17\%$ of the mean squared prediction error.

6. Discussion

To assess the prediction accuracy of a nonlinear prediction function with right-censored data, we have first proposed a pair of population prediction accuracy parameters ρ^2 and λ^2 and then developed their sample versions R^2 and L^2 for both uncensored and censored data. The R^2 statistic, defined as the proportion of explained variance by its linearly corrected prediction function, quantifies the potential predictive power of the prediction function. The L^2 statistic, defined as the proportion of explained prediction error by its corrected prediction function, measures how close the prediction function is to its corrected prediction. Together, they give a complete summary regarding the prediction accuracy of a nonlinear prediction function. We highlight that the proposed R^2 statistic for right-censored data enjoys an appealing property that it educes to the classical coefficient of determination R^2 for the linear model in the absence of censoring, which is not shared by any other existing pseudo R^2 proposals for right-censored data. Furthermore, L^2 degenerates to 1 for the linear model with uncensored data. In practice, we recommend that R^2 be used as the primary measure to evaluate and compare the potential predictive power between competing prediction models, and L^2 be used as a supplementary measure only for the final prediction model of interest (such as the one with the largest R^2 value) to indicate by a value less 1 if a correction is needed for the prediction function to fulfill its potential predictive power and quantify how much prediction error reduction can be realized by the correction, as illustrated in the real data example in Section 5.

Our simulation results show that the sample R^2 and L^2 measures estimate their corresponding population parameters well with little bias with moderate sample size and censoring rate, regardless of whether the model is correctly or misspecified. The consistency of the proposed measures for right-censored data is derived under a rather strong technical assumption that the censoring time is independent of both the survival time and the covariates. However, our simulation results indicate that even when the independent censoring assumption is violated, the proposed measures still estimate their population counterparts well, except when the model is severely misspecified. Therefore, in the presence of dependent censoring, it is important to routinely perform model diagnostics to identify and eliminate a severely misspecified model before further applying the proposed measures to evaluate its prediction power. Finally, $L^2 = 1$ simply indicates that no correction is needed for the prediction function to achieve its potential prediction power. It should not be used to suggest a good fit of the model to the data as discussed in Remark 2.1.

This article focuses on event time models with a single failure type, time-independent covariates and independently right-censored data. Future efforts to develop prediction accuracy measures for event time models with time-dependent covariates, competing risks, time-varying effects, and other censoring patterns are warranted. It would also be interesting to extend the proposed measures to weighted R^2 and L^2 that naturally follow from some weighted versions of the variance and prediction error decompositions similar to Lemma A.3, which could be useful to evaluate the local predictive performance for some subpopulation of interest. Our team is also investigating the application of the proposed R^2 measure for node-splitting in survival tree regression and survival random forest.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the associate editor and a referee for their insightful and constructive comments that have led to significant improvements of the article.

Funding

The research of Gang Li was partly supported by National Institute of Health Grants P30 CA16042, UL1TR000124-02, and P50 CA211015.

References

Ash A, and Shwartz M (1999), "R2: A Useful Measure of Model Performance When Predicting a Dichotomous Outcome," Statistics in Medicine, 18, 375–384. [1] [PubMed: 10070680]

Cox DR (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society,* Series B, 34, 187–220. [2]

Cox DR, and Snell EJ (1989), Analysis of Binary Data (Vol. 32), New York: CRC Press [1]

Cox DR, and Wermuth N (1992), "A Comment on the Coefficient of Determination for Binary Responses," The American Statistician, 46, 1–4. [1]

Dickson ER, Grambsch PM, Fleming TR, Fisher LD, and Langworthy A (1989), "Prognosis in Primary Biliary Cirrhosis: Model for Decision Making," Hepatology, 10, 1–7. [8] [PubMed: 2737595]

Efron B (1978), "Regression and ANOVA with Zero-One Data: Measures of Residual Variation," Journal of the American Statistical Association, 73, 113–121. [1]

Efron B, and Tibshirani RJ (1994), An Introduction to the Bootstrap, New York: CRC Press [4]

Goodman LA (1971), "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications," Technometrics, 13, 33–61. [1]

Graf E, Schmoor C, Sauerbrei W, and Schumacher M (1999), "Assessment and Comparison of Prognostic Classification Schemes for Survival Data," Statistics in Medicine, 18, 2529–2545. [2] [PubMed: 10474158]

Haberman SJ (1982), "Analysis of Dispersion of Multinomial Responses," Journal of the American Statistical Association, 77, 568–580. [1]

Harrell FE, Califf RM, Pryor DB, Lee KL, and Rosati RA (1982), "Evaluating the Yield of Medical Tests," JAMA, 247, 2543–2546. [1] [PubMed: 7069920]

Hilden J (1991), "The Area Under the ROC Curve and its Competitors," Medical Decision Making, 11, 95–101. [1] [PubMed: 1865785]

Kaplan E, and Meier P (1958), "Nonparametric Estimation From Incomplete Observations," Journal of the American Statistical Association, 53, 457–481. [5]

Kent JT (1983), "Information Gain and a General Measure of Correlation," Biometrika, 70, 163–173. [1]

Kent JT, and O'Quigley J (1988), "Measures of Dependence for Censored Survival Data," Biometrika, 75, 525–534. [2]

Korn EL, and Simon R (1990), "Measures of Explained Variation for Survival Data," Statistics in Medicine, 9, 487–503. [2] [PubMed: 2349402]

Lee M-LT, and Whitmore GA (2006), "Threshold Regression for Survival Analysis: Modeling Event Times by a Stochastic Process Reaching a Boundary," Statistical Science, 21, 501–513. [6]

Maddala GS (1986), *Limited-Dependent and Qualitative Variables in Econometrics* (No. 3). New York: Cambridge University Press [1]

Magee L (1990), "R2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests," The American Statistician, 44, 250–253. [1]

McFadden D (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in Frontiers in Econometrics, ed. Zarembka P, New York: Wiley [1]

Mittlbock M, and Schemper M, (1996), "Explained Variation for Logistic Regression," Statistics in Medicine, 15, 1987–1997. [1] [PubMed: 8896134]

Nagelkerke NJ (1991), "A Note on a General Definition of the Coefficient of Determination," Biometrika, 78, 691–692. [1]

O'Quigley J, Xu R, and Stare J (2005), "Explained Randomness in Proportional Hazards Models," Statistics in Medicine, 24, 479–489. [2] [PubMed: 15532086]

Rényi A (1959), "On Measures of Dependence," Acta Mathematica Hungarica, 10,441–451. [4]

Royston P, and Sauerbrei W (2004), "A New Measure of Prognostic Separation in Survival Data," Statistics in Medicine, 23, 723–748. [2] [PubMed: 14981672]

Schemper M and Henderson R (2000), "Predictive Accuracy and Explained Variation in Cox Regression," Biometrics, 56, 249–255. [2,5] [PubMed: 10783803]

Stare J, Perme MP, and Henderson R (2011), " Measure of Explained Variation for Event History Data," Biometrics, 67, 750–759. [2,5] [PubMed: 21155749]

Theil H (1970), "On the Estimation of Relationships Involving Qualitative Variables," American Journal of Sociology, 76, 103–154. [1]

Zheng B, and Agresti A (2000), "Summarizing the Predictive Power of a Generalized Linear Model," Statistics in Medicine, 19, 1771–1781. [1] [PubMed: 10861777]

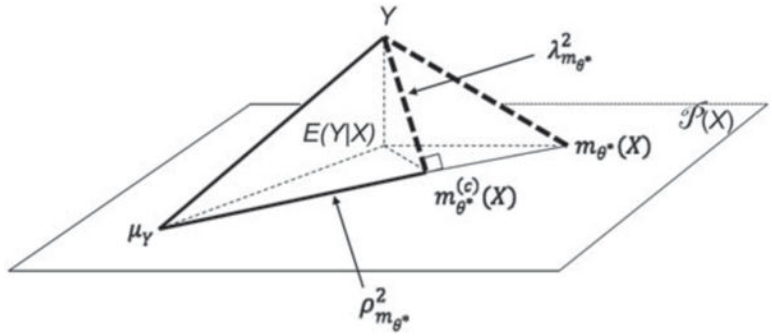

Figure 1.

Geometric interpretation of $\rho^2_{m\theta*}$ and $\lambda^2_{m\theta*}$.

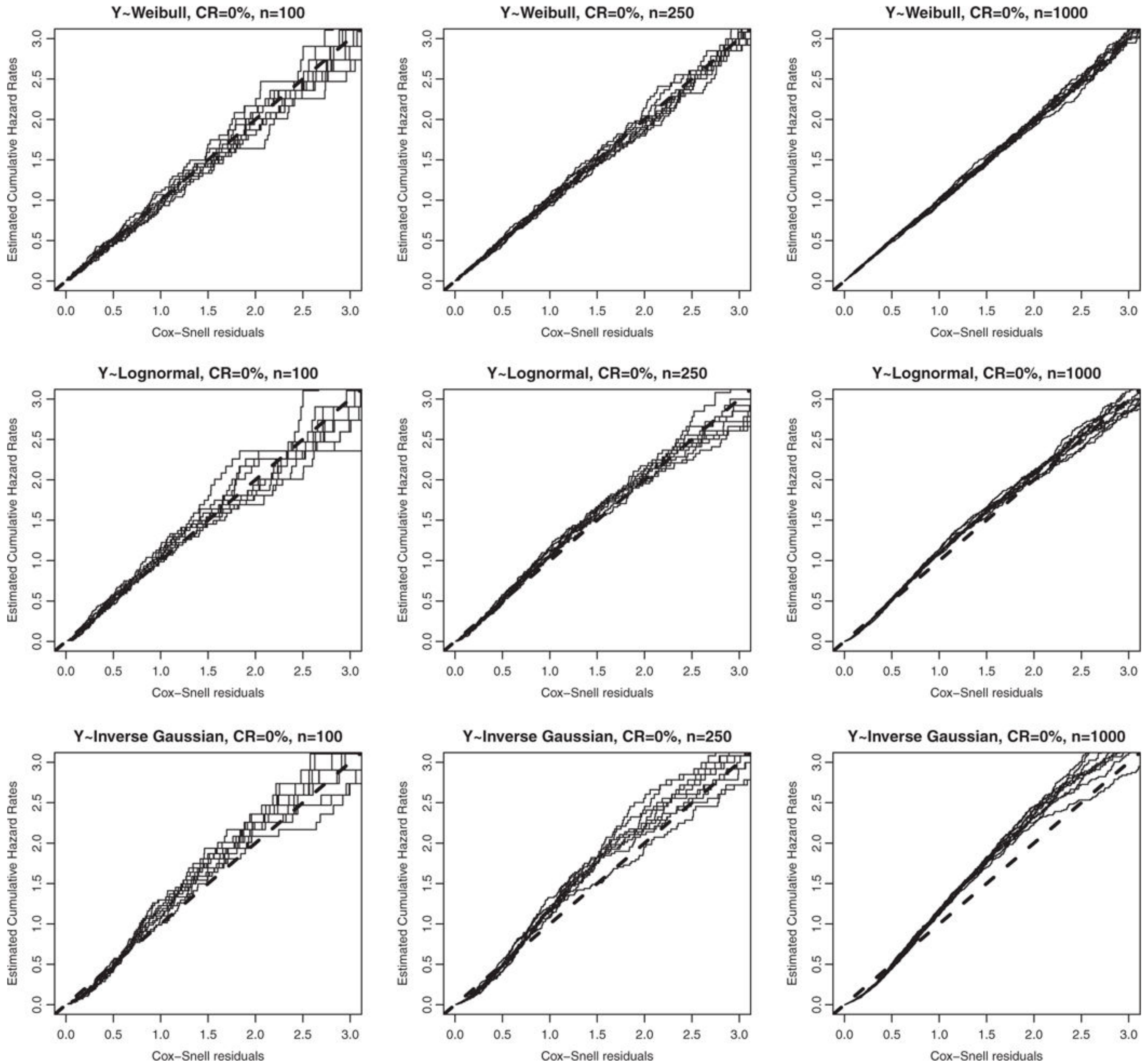

Figure 2.
(Cox’s model with independent censoring; censoring rate, CR = 0%) Cox-Snell residual plot for the Cox model based on the first 10 Monte Carlo samples with censoring rate CR = 0%, varying sample size (first column: $n = 100$; second column: $n = 250$; third column: $n = 1000$), and varying data generation setting (first row: Weibull; second row: log-normal; third row: inverse Gaussian).

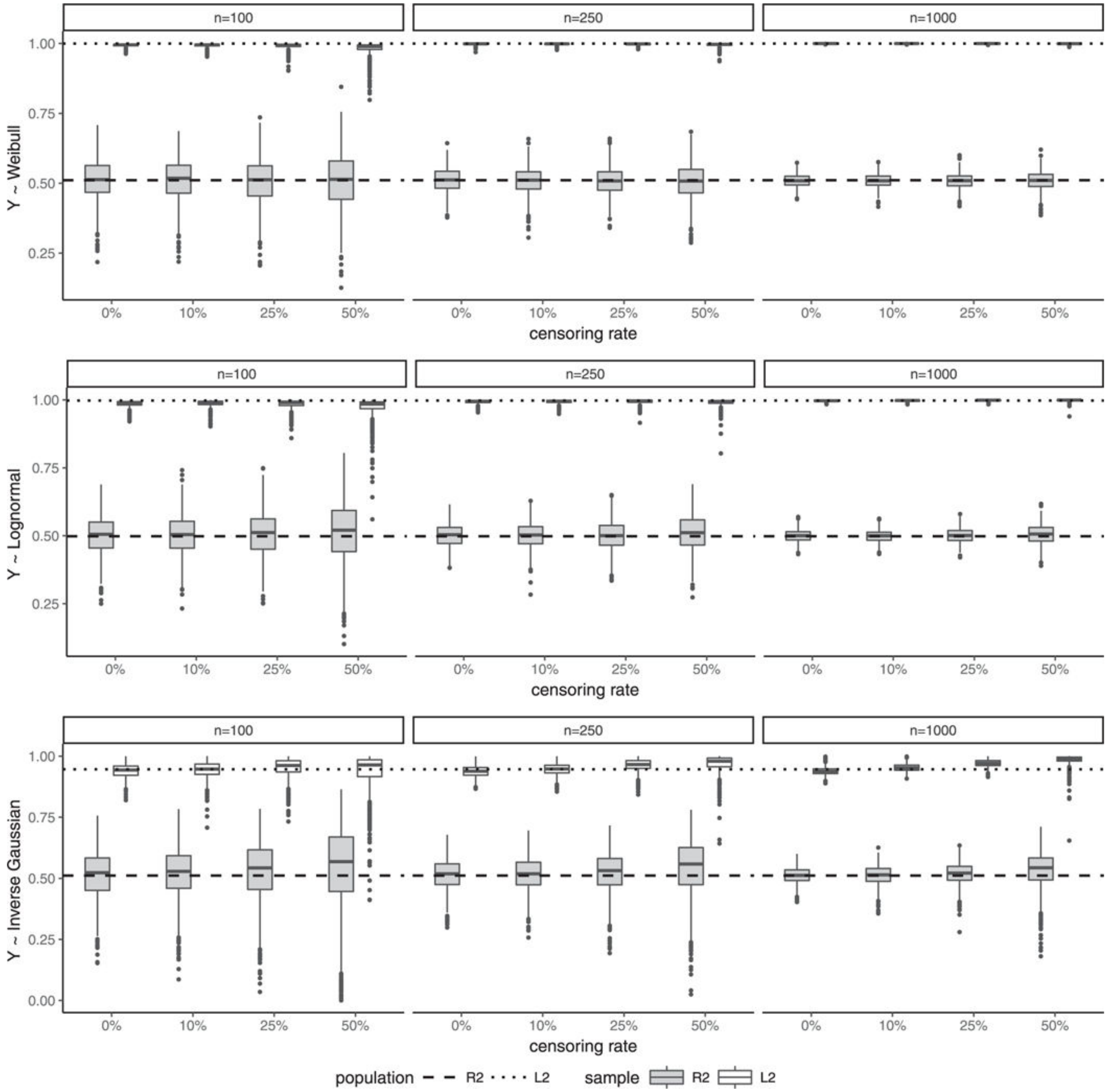

Figure 3.
(Independent censoring) Boxplots of simulated R^2 (shaded box) and L^2 (unshaded box) for the Cox model by censoring rate (0%, 10%, 25%, 50%), sample size (100,250,1000), and data generation setting (upper panel: Weibull; middle panel: log-normal; bottom panel:inverse Gaussian).

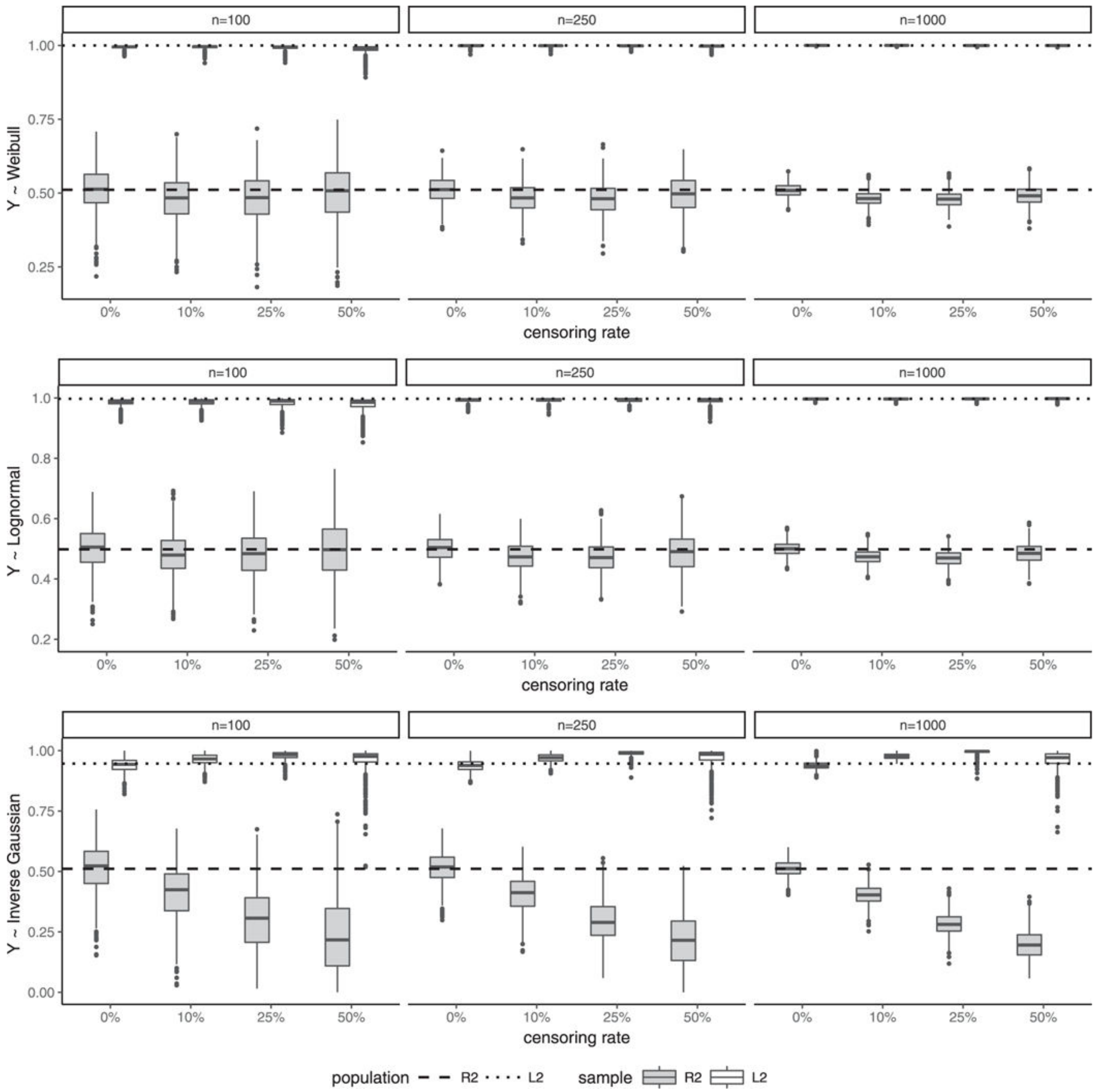

Figure 4.

(Dependent censoring) Boxplots of simulated R^2 (shaded box) and L^2 (unshaded box) for the Cox model by censoring rate (0%, 10%, 25%, 50%), sample size (100,200,1000), and data generation setting (upper panel: Weibull; middle panel: log-normal AFT; bottom panel: inverse Gaussian).

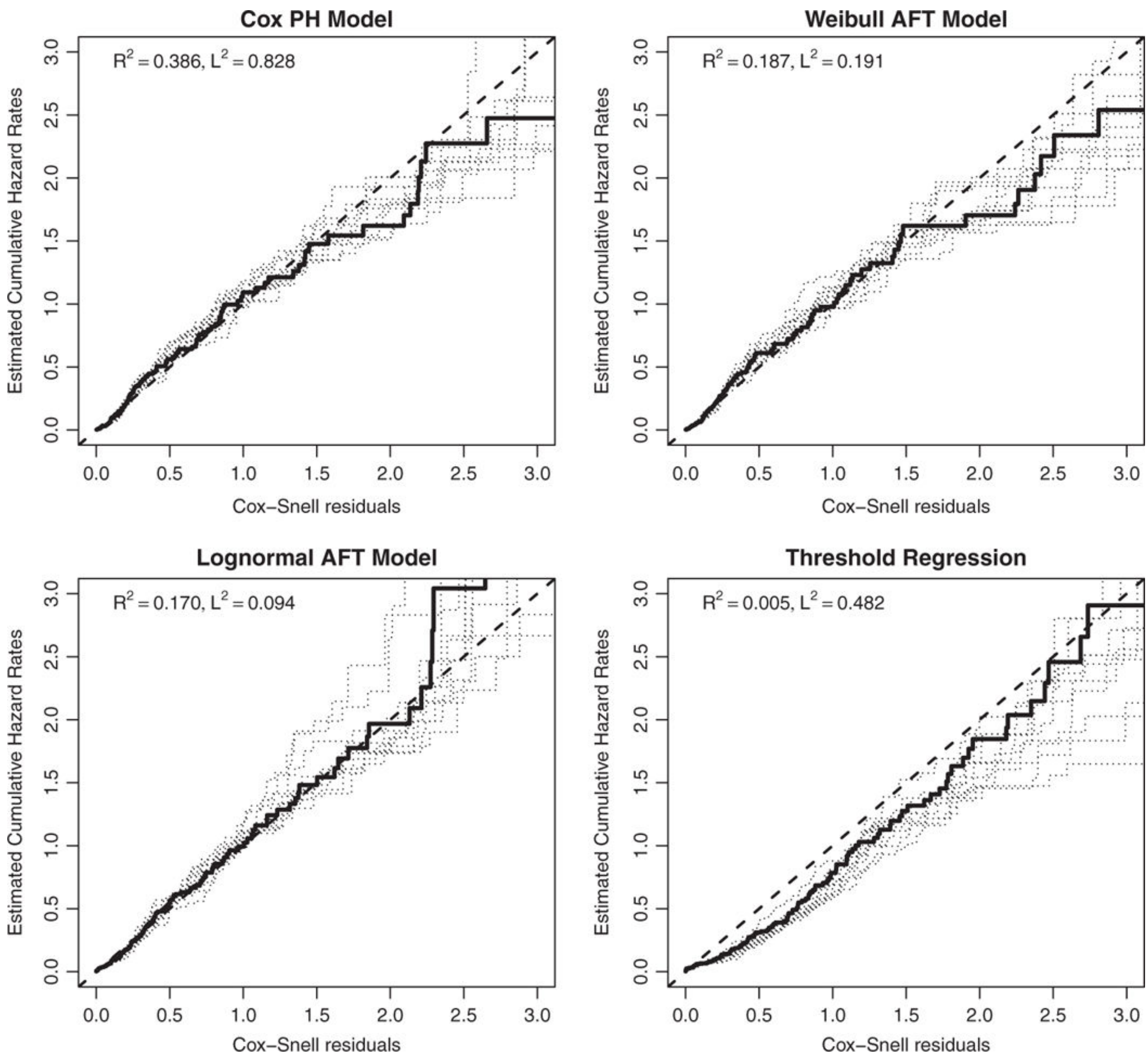

Figure 5.

(PBCdata) Cox-Snell residual plots for the Cox model, Weibull AFT model, log-normal AFT model, and threshold regression model. For each model, the solid line is based on the observed PBC data and the dotted lines are based on 10 bootstrap samples. Deviations from the 45° line indicate possible lack-of-fit to the data.

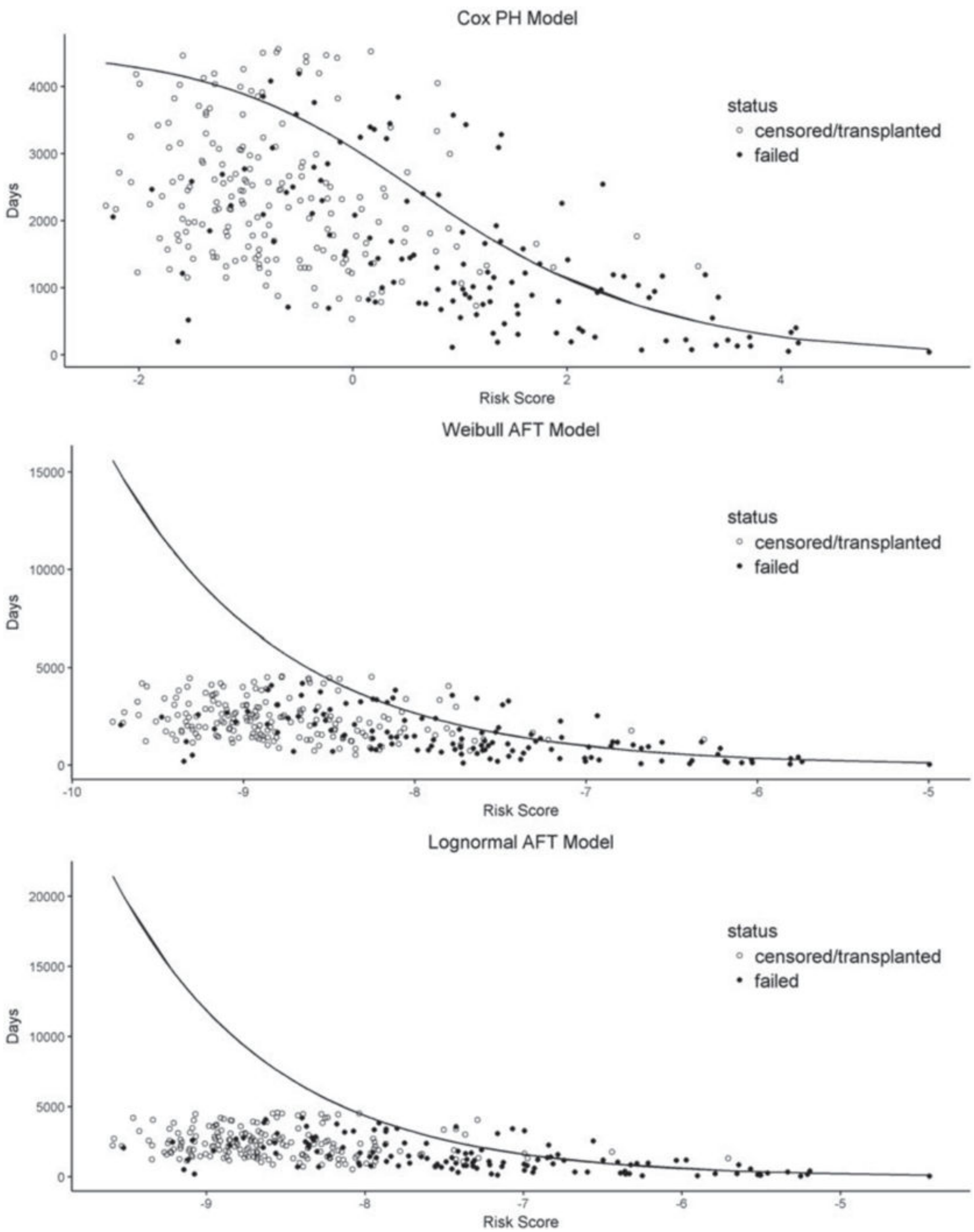

Figure 6.
(PBC data) Predicted (solid line) and observed (solid dot: uncensored; censored: circle) survival times (in days) versus risk score for the Cox model (top panel), Weibull AFT model (middle panel), and log-normal AFT model (bottom panel).

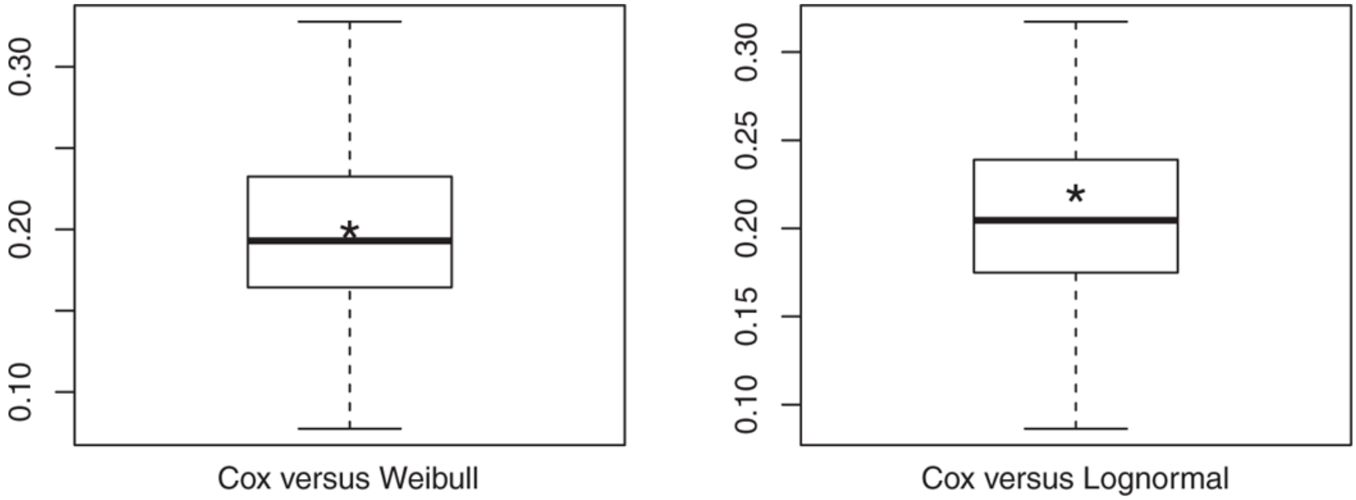

Figure 7.
(PBC data) Boxplots of the R^2 differences between different models based on 100 bootstrap samples from the PBC data (Left: Cox's model versus Log-normal AFT model; Right: Cox's model versus Weibull AFT model). The asterisk in each boxplot represents the R^2 difference between the two models based on the observed PBC data.

Table 1.

Simulated population proportion $\left(\rho^2_{\mathrm{NP}}\right)$ of explained variance by the Cox (1972) model and population (R^2_{SPH} and R^2_{SH} of Schemper and Henderson (2000) and Stare, Perme, and Henderson (2011).

Model	β	v	ρ^2_{NP}	R^2_{SPH}	R^2_{SH}
1	0.1	0.5	0.07	0.23	0.10
2	0.1	1	0.14	0.23	0.10
3	0.1	10	0.15	0.23	0.10
4	0.2	0.5	0.09	0.38	0.28
5	0.2	1	0.27	0.38	0.28
6	0.2	10	0.40	0.38	0.28
7	0.5	0.5	0.09	0.49	0.49
8	0.5	1	0.33	0.49	0.49
9	0.5	10	0.80	0.49	0.49

Table 2.

(PBC data) R^2 values of different survival regression models.

Model	Cox PH model	Weibull AFT model	Log-normal AFT model
R^2	0.39	0.19	0.17
L^2	0.83	0.19	0.09