Estimating Covariances and Goodness of Fit Plots for Accelerated Failure Time Models

David J. Olive and Sanjuka Johana Lemonge*
Southern Illinois University

October 26, 2025

Abstract

Let the response variable Y be the time until an event such as death. Assume that there are p predictors $x_1, ..., x_p$ and that the response variable is right censored. Several survival regression models, including accelerated failure time models, have the form $Z = \log(Y) = \alpha_Z + \boldsymbol{x}_i^T \boldsymbol{\beta}_Z + e$. This paper gives a simple method for estimating the covariances $Cov(x_i, Z)$ for some of these models. Plots are given for checking the goodness of fit of accelerated failure time models. Plots for checking the proportional hazards regression model are often harder to use.

KEY WORDS: Buckley James estimator, Multiple linear regression, Proportional hazards regression, Survival regression, Weibull regression.

1 INTRODUCTION

This section reviews some survival regression models, including plots for checking the models. The response variable $Y > 0$ is the time until an event such as death. Let the $p \times 1$ vector of predictor variables $\boldsymbol{x} = (x_1, ..., x_p)^T$. Let the sufficient predictor $SP = \alpha + \boldsymbol{x}^T \boldsymbol{\beta}$, and let the estimated sufficient predictor $ESP = \hat{\alpha} + \boldsymbol{x}^T \hat{\boldsymbol{\beta}}$ where it is possible that $\alpha = \hat{\alpha} = 0$. The ESP is sometimes called the estimated risk score.

Assume that the response variable is right censored so Y is not observed. Instead, the right censored survival time $T_i = \min(Y_i, W_i)$ where Y_i is independent of the censoring time W_i. Also $\delta_i = 0$ if $T_i = W_i$ is censored and $\delta_i = 1$ if $T_i = Y_i$ is uncensored. Hence the data is $(T_i, \delta_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$.

For an accelerated failure time model, the log transformation of the response variable results in a multiple linear regression model. Hence multiple linear regression models will be useful. Now let the response variable Y be for multiple linear regression, so Y

*David J. Olive is Professor, School of Mathematical & Statistical Sciences, Southern Illinois University, Carbondale, IL 62901, USA.

need not be a nonnegative time until event. A useful multiple linear regression model is $Y|\boldsymbol{x}^T\boldsymbol{\beta} = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$ or $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ or

$$Y_i = \alpha + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1}$$

for $i = 1, ..., n$. Assume that the e_i are independent and identically distributed (iid) with expected value $E(e_i) = 0$ and variance $V(e_i) = \sigma^2$. In matrix form, this model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}, \tag{2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times (p+1)$ matrix with ith row $(1, \boldsymbol{x}_i^T)$, $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta}^T)^T$ is a $(p+1) \times 1$ vector , and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Also $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

For a multiple linear regression model with heterogeneity, assume model (1) holds with $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma_e} = diag(\sigma_i^2) = diag(\sigma_1^2, ..., \sigma_n^2)$ is an $n \times n$ positive definite matrix. When the σ_i^2 are known, weighted least squares is often used. Under regularity conditions, the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\phi}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ can be shown to be a consistent estimator of $\boldsymbol{\phi}$.

For estimation with ordinary least squares, let the $p \times p$ covariance matrix of $\boldsymbol{x}$ be $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma_x} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T]$ and let the $p \times 1$ vector $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma_{xY}} = E[(\boldsymbol{x} - E(\boldsymbol{x})(Y - E(Y))] = (\text{Cov}(x_1, Y), ..., \text{Cov}(x_p, Y))^T$. Let

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \text{ and } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}).$$

Then the OLS estimators for model (1) are $\hat{\boldsymbol{\phi}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, $\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T\overline{\boldsymbol{x}}$, and

$$\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}.$$

For a multiple linear regression model with iid cases, $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ under mild regularity conditions, while $\hat{\alpha}_{OLS}$ is a consistent estimator of $E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x})$.

For a parametric accelerated failure time (AFT) model,

$$Z_i = \log(Y_i) = \alpha + \boldsymbol{\beta}_A^T\boldsymbol{x}_i + \sigma e_i \tag{3}$$

where the e_i are iid from a location scale family. The parameters are estimated by maximum likelihood.

The Weibull proportional hazards regression model or Weibull regression model is

$$Y|SP \sim W(\gamma = 1/\sigma, \lambda_0 \exp(SP))$$

where Y has a Weibull $W(\gamma, \lambda)$ distribution if the probability density function of Y is

$$f(y) = \lambda\gamma y^{\gamma-1}\exp[-\lambda y^{\gamma}] \text{ for } y > 0.$$

This regression model can also be fit using the Cox (1972) proportional hazards regression model. Let the sufficient predictor $SP = \boldsymbol{x}^T\boldsymbol{\beta}_P$. If $Y|\boldsymbol{x}^T\boldsymbol{\beta}_P$ satisfies a Weibull regression

model, then $Z = \log(Y) = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}_A + e_i$ satisfies a Weibull AFT with $\lambda_0 = \exp(-\alpha/\sigma)$ and $\boldsymbol{\beta}_P = -\boldsymbol{\beta}_A/\sigma$. Exponential regression is the special case where $\sigma = 1$.

Two other important AFTs are i) the lognormal AFT where $\log(Y)|\boldsymbol{x}^T\boldsymbol{\beta}_A \sim N(\alpha + \boldsymbol{x}^T\boldsymbol{\beta}_A, \sigma^2)$ where the Y_i are lognormal and the $e_i \sim N(0,1)$ are normal, and ii) the loglogistic AFT where $\log(Y)|\boldsymbol{x}^T\boldsymbol{\beta}_A \sim L(\alpha + \boldsymbol{x}^T\boldsymbol{\beta}_A, \sigma)$ where the Y_i are loglogistic and the $e_i \sim L(0,1)$ are logistic. For the loglogistic AFT, Y follows a proportional odds model. Y does not follow a proportional hazards regression model for the loglogistic and lognormal AFTs.

The Buckley and James (1979) estimator $(\hat{\alpha}_{BJ}, \hat{\boldsymbol{\beta}}_{BJ})$ is a nonparametric survival regression method for models of the form (3), and is a competitor for the parametric AFTs. When there is no censoring, this estimator is equivalent to the ordinary least squares estimator for multiple linear regression.

Often the log transformation results in a linear model with heterogeneity:

$$Z_i = \log(Y_i) = \alpha_Z + \boldsymbol{x}_i^T\boldsymbol{\beta}_Z + e_i \tag{4}$$

where the e_i are independent with expected value $E(e_i) = 0$ and variance $V(e_i) = \sigma_i^2$. For the AFT and the Buckley James estimator, the variance is constant: $V(e_i) = \sigma^2$ does not depend on i.

The Cox (1972) proportional hazards regression model is a semiparametric model with $SP = \boldsymbol{\beta}_C^T\boldsymbol{x}$ and hazard function

$$h_{Y|SP}(t) = \exp(\boldsymbol{\beta}_C^T\boldsymbol{x})h_0(t) = \exp(SP)h_0(t)$$

where the baseline hazard function $h_0(t)$ is left unspecified. The survival function is

$$S_{Y|SP}(t) = P(Y > t|SP = \boldsymbol{\beta}_C^T\boldsymbol{x}) = [S_0(t)]^{\exp(\boldsymbol{\beta}_C^T\boldsymbol{x})} = [S_0(t)]^{\exp(SP)} \tag{5}$$

where $S_0(t)$ is the baseline survival function. If $\boldsymbol{x} = \boldsymbol{0}$ is within the range of the predictors, then the baseline survival and hazard functions correspond to the survival and hazard functions of $\boldsymbol{x} = \boldsymbol{0}$. The cumulative hazard function $H_{Y|SP}(t) = -\log(S_{Y|SP}(t))$ for $t > 0$, and the hazard function $h_{Y|SP}(t) = \frac{d}{dt}H_{Y|SP}(t)$ for $t > 0$. High hazard implies low survival times while low hazard implies long survival times.

The literature for checking the goodness of fit of the proportional hazards model is fairly large. See, for example, Arjas (1988), Feng et al. (2017), Gill and Schumaker (1987), Kay (1984), Lin and Wei (1991), Marzec and Marzec (1997), Ng'andu (1997), Quantin et al. (1996) and Yu et al. (2008).

Grambsch and Therneau (1994) give a useful graphical check. Suppose the ith case had an uncensored survival time t_i. Let the scaled Schoenfeld residual for the ith observation and jth variable x_j be $r^*_{P,j}(t_i)$. For each variable, plot the t_i versus the $r^*_{P,j}(t_i) + \hat{\beta}_j$ and add the loess curve. If the loess curve is approximately horizontal for each of the p plots, then the proportional hazards assumption is reasonable. Alternatively, fit a line to each plot and test that each of the p slopes is equal to 0. The R function `cox.zph` makes both the plots and tests.

2 Estimating $\Sigma_{\boldsymbol{x}Z}$ for Some Censored Survival Regression Models

This section derives an estimator for $\boldsymbol{\Sigma}_{\boldsymbol{x}Z} = \text{Cov}(\boldsymbol{x}, Z)$ where the right censored Z_i are not observed. Let the ordinary least squares (OLS) estimator be $\hat{\boldsymbol{\beta}}_{OLS}$. Assume that the cases $(\boldsymbol{x}_i, Y_i)$ are iid. Since model (4) is a multiple linear regression model, under mild regularity conditions, the population parameter vectors are equal: $\boldsymbol{\beta}_Z = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Z}$. Thus $\boldsymbol{\Sigma}_{\boldsymbol{x}Z} = \text{Cov}(\boldsymbol{x})\boldsymbol{\beta}_Z = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_Z$. When the response Y_i is censored, several survival regression models give consistent estimators $\hat{\boldsymbol{\beta}}_Z$ of $\boldsymbol{\beta}_Z$. Hence

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_Z. \tag{6}$$

The Buckley James estimator is equivalent to the OLS estimator when there is no censoring, and censoring does not change the population parameter. Although the population parameters are the same, the estimators are very different. The OLS estimator is not a consistent estimator of $\boldsymbol{\beta}_Z = \boldsymbol{\beta}_{OLS}$ when censoring occurs. Hence $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}$ is not a consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{x}Z}$ if the response variable is censored.

If an accelerated failure time model is used, then two estimators are $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z}(A) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z}(B) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{BJ}$. These two estimators require consistent estimators of $\boldsymbol{\beta}_Z = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Z}$, which occurs, for example, if the cases $(\boldsymbol{x}_i, Y_i)$ are iid from some population with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and covariance vector $\boldsymbol{\Sigma}_{\boldsymbol{x}Z}$. The survival times Y_i can be right censored, but the predictor variables $x_1, ..., x_p$ are not censored. Note that the predictor variables that have the highest absolute correlation with Z have the highest values of $|\widehat{\text{Cov}}(x_i, Z)|/\sqrt{\hat{V}(x_i)}$.

Although this technique is a simple plug in estimator, to our knowledge, the technique has not been suggested for survival regression models before. In the literature, there are several estimators for the correlation $Cor(X, Y)$ where X and Y are survival times. These estimators usually use maximum likelihood estimation or multiple imputation assuming that (X, Y) are iid from a bivariate normal distribution. See, for example, Barchard and Russell (2024), Li et al. (2018), and Lyles, Fan, and Chuachoowong (2001). For covariance estimation, see references in Pesonen, Pesonen, and Nevalainen (2015).

3 The EE Plot

It is important to check that a parametric AFT model is reasonable with the Buckley James estimator before using Equation (6). Make an EE plot of $ESPBJ = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_{BJ}$ versus $ESPA = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_A$. For the Weibull AFT, also plot $ESPPH = -\hat{\sigma}\boldsymbol{x}^T\hat{\boldsymbol{\beta}}_P$ versus the above two ESPs, where PH stands for the Cox proportional hazards estimator. The plotted points in the EE plot should scatter tightly about the identity line with zero intercept and unit slope. Lack of fit is suggested if the plotted points do not cluster tightly about the identity line. The identity line is included in the EE plots as a visual aid. Suppose the two estimators of $\boldsymbol{\beta} = \boldsymbol{\beta}_Z$ are consistent. Then as $n \to \infty$, the correlation of the plotted points goes to 1 in probability for any finite interval, e.g., from the 1st percentile to the 99th percentile of $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}_{BJ}$.

For the Exponential regression model, $\sigma = 1$ and $\boldsymbol{\beta}_P = -\boldsymbol{\beta}_A$. Since the Exponential regression model with $\hat{\boldsymbol{\beta}}_E$ is a special case of the Weibull regression model with $\hat{\boldsymbol{\beta}}_W$, plots can use $ESPE = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_E$, $ESPBJ$, $ESPPH$ or $-\boldsymbol{x}^T\hat{\boldsymbol{\beta}}_P$, and $ESPA = ESPW$ where $A = W$ is the Weibull regression that estimates $\hat{\sigma}$.

The EE plots are much easier to use than other plots in the literature, since tight clustering about the identity line suggests goodness of fit.

4 Some Other Plots

Let $\log(T_i) = \hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T\boldsymbol{x}_i + r_i$ where the r_i are residuals. Collett (2003, p. 231) defines a standardized residual $r_{Si} = r_i/\hat{\sigma}$. For accelerated failure time models, a *log censored response (LCR) plot* is a plot of $\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T\boldsymbol{x}_i$ versus $\log(T_i)$ on the vertical axis with plotting symbol 0 for censored cases and + for uncensored cases. The identity line with unit slope and zero intercept is added to the plot, and the vertical deviations from the identity line $= r_i$. This plot is useful to check for cases with unusual survival times. The plot is also useful for checking for linearity (which is often rather weak due to the censored response and skewed data).

Ritov (1990) and Tsiatis (1990) consider the censored survival regression model $Z = t(Y) = \alpha_Z + \boldsymbol{x}_i^T\boldsymbol{\beta}_Z + e$ where t is a known monotone function. Then a plot of $\hat{\alpha}_Z + \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_Z$ versus $t(T_i)$ is useful. A power transformation has the form $W = t_\lambda(T) = T^\lambda$ for $\lambda \neq 0$ and $W = t_0(T) = \log(T)$ for $\lambda = 0$ where

$$\lambda \in \Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}.$$

The modified power transformation family

$$t_\lambda(T_i) \equiv T_i^{(\lambda)} = \frac{T_i^\lambda - 1}{\lambda} \tag{7}$$

for $\lambda \neq 0$ and $T_i^{(0)} = \log(T_i)$.

A graphical method for response transformations makes the transformation plot of $\hat{\alpha} + \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ versus $t_\lambda(T_i)$ using the Buckley James estimator with $W = t_\lambda(T)$ as the "response" for $\lambda \in \Lambda_L$. Then a candidate response transformation $Z = t_{\lambda^*}(T)$ has the transformation plot that suggests a linear model is appropriate.

Smith (2002, p. 191) notes that the Buckley James estimator replaces the censored T_i by their estimated conditional expected values, resulting in a renovated response variable T_i^*. Using the T_i^* instead of the T_i may increase the linearity of the plot.

The *slice survival plot* divides the ESP into J groups of roughly the same size. For each group j with n_j cases, the model estimated survival function $\hat{S}_j(t)$ is computed using the $\boldsymbol{x}$ corresponding to the "median ESP" of the group (the kth order statistic of the ESP in group j, where $k = 1 + floor[(n_j - 1)/2]$). Let $\hat{S}_{KMj}(t)$ be the Kaplan Meier estimator computed from the survival times (T_i, δ_i) in the jth group. For each group, $\hat{S}_j(t)$ is plotted and $\hat{S}_{KMj}(t_i)$ is plotted as circles at the uncensored event times t_i. The survival regression model is reasonable if the circles "tracks $\hat{S}_j$ well" in each of the J plots.

If the slice widths go to zero, but the number of cases per slice increases to ∞ as $n \to \infty$, then the Kaplan Meier estimator and the model estimator converge to $S_{Y|SP}(t)$ if the model holds. Simulations suggest that the two survival functions are "close" for moderate n and nine slices. For small n and skewed predictors, some slices may be too wide in that the model is correct but $\hat{S}_{KMj}(t)$ is not a good approximation of $S_{Y|SP}(t)$ where SP corresponds to the $\boldsymbol{x}$ used to compute $\hat{S}_j(t)$.

For the Cox model, if pointwise confidence interval (CI) bands are added to the plot, then $\hat{S}_{KMj}$ "tracks $\hat{S}_j$ well" if most of the plotted circles do not fall very far outside the pointwise CI bands since these pointwise bands are not as wide as simultaneous bands. Collett (2003, pp. 241-243) places several observed Kaplan Meier curves with fitted curves on the same plot.

Survival regression is the study of the conditional survival $S_{Y|SP}(t)$, and the slice survival plot is a useful tool for visualizing $S_{Y|SP}(t)$ in the background of the data. Suppose the jth slice is narrow so that $ESP \approx w_j$. If the model is reasonable, $ESP \approx SP$, and the number of uncensored cases in the jth slice is not too small, then $S_{Y|SP=w_j}(t) \approx \hat{S}_j(t) \approx \hat{S}_{KMj}(t)$. (These quantities approximate $[\hat{S}_0(t)]^{\exp(w_j)}$ for the Cox model.) Thus the nonparametric Kaplan Meier estimator is used to check the model estimator $\hat{S}_j(t)$ in each slice.

The slice survival plot tailored to the Cox model is closely related to May and Hosmer (1998) test. Also, van Houwelingen et al. (2006) use similar ideas, but place the J Kaplan Meier curves on one plot and the J Cox survival curves on another plot. The ESP is a scalar while $\boldsymbol{x}$ is a $p \times 1$ vector. Using the ESP instead of $\boldsymbol{x}$ in plots is an important dimension reduction technique (and is similar to using a scalar valued minimal sufficient statistic instead of the p-dimensional sufficient statistic $\boldsymbol{x}$.) Plots have been suggested by several authors with $\boldsymbol{x}$ divided into J groups instead of the ESP. For example, see Miller (1981, p. 168). Hosmer and Lemeshow (1999, pp. 141–145) suggests making plots based on the quartiles of the ith predictor x_i, and note that a problem with Cox survival curves, $\hat{S}_{\boldsymbol{x}}(t) = [\hat{S}_0(t)]^{\exp(\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x})} = [\hat{S}_0(t)]^{\exp(ESP)}$, is that they may use inappropriate extrapolation. Using the ESP results in narrow slices with many cases, and adding Kaplan Meier curves shows if there is extrapolation.

5 Examples and Simulations

Example 1. The ovarian cancer data is from Collett (2003, p. 187-190) and Edmunson et al. (1979). The response variable is the survival time of $n = 26$ ovarian cancer patients in days with predictors *age* in years and *treat* (1 for cyclophosphamide alone and 2 for cyclophosphamide combined with adriamycin). See Figure 1 for the three EE plots for the ovarian cancer data, where ESPW=ESPA. The Weibull AFT appears to be appropriate for this data set. Then $\widehat{Cov}(age, Z) = -0.1286$, $\widehat{Cov}(treat, Z) = -7.90408$, $\widehat{Cov}(age, Z)/\sqrt{\hat{V}(age)} = -0.2522$, and $\widehat{Cov}(treat, Z)/\sqrt{\hat{V}(treat)} = -0.7840$. Hence $|\widehat{Cor}(treat, Z)| \approx 3|\widehat{Cor}(age, Z)|$. Figure 2 shows the EE plots when an Exponential AFT is used instead of the Weibull AFT. Now the plotted points do not cluster about the identity line. Hence the Exponential AFT should not be used. Figure 3 shows the

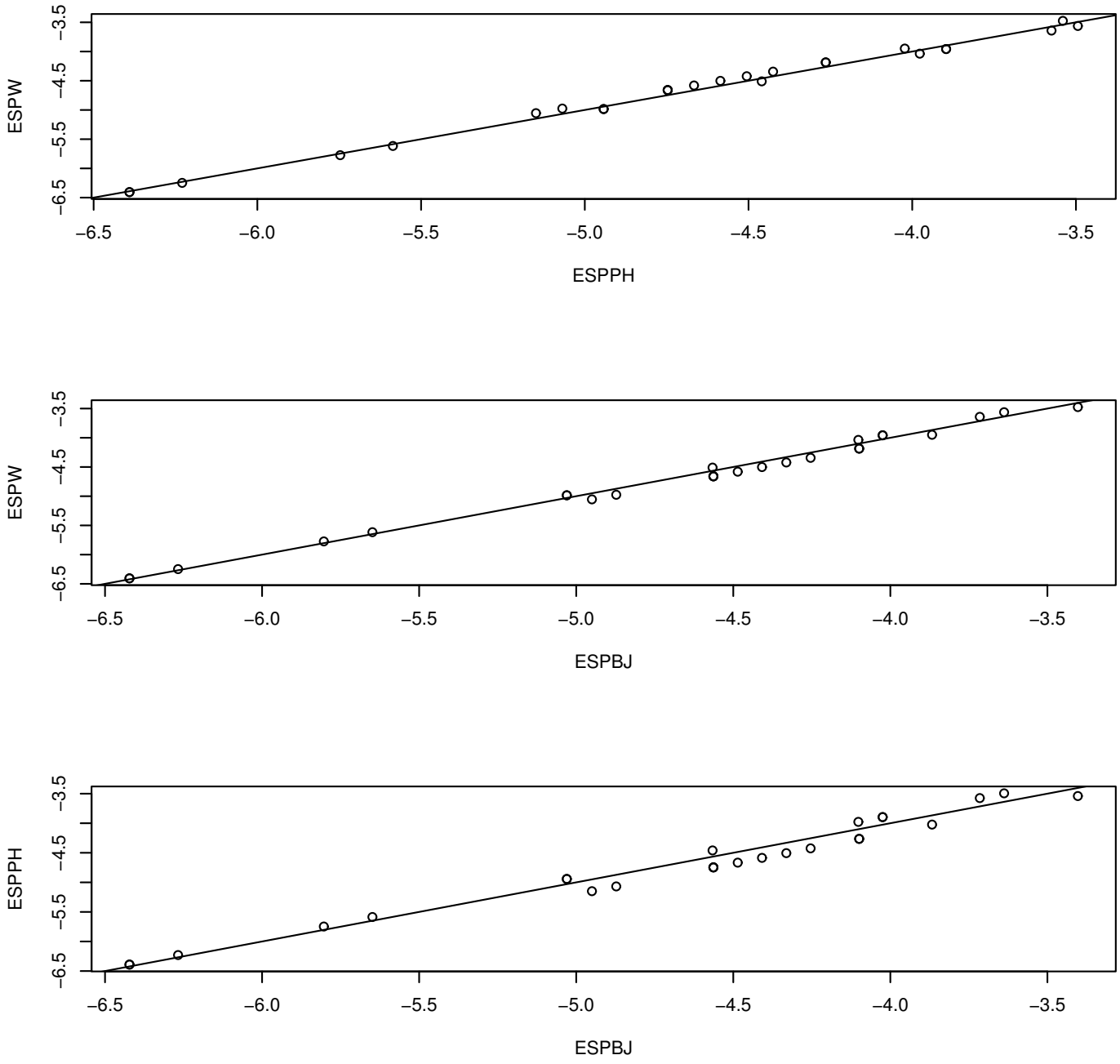

Figure 1: Three EE Plots for the Ovarian Cancer Data

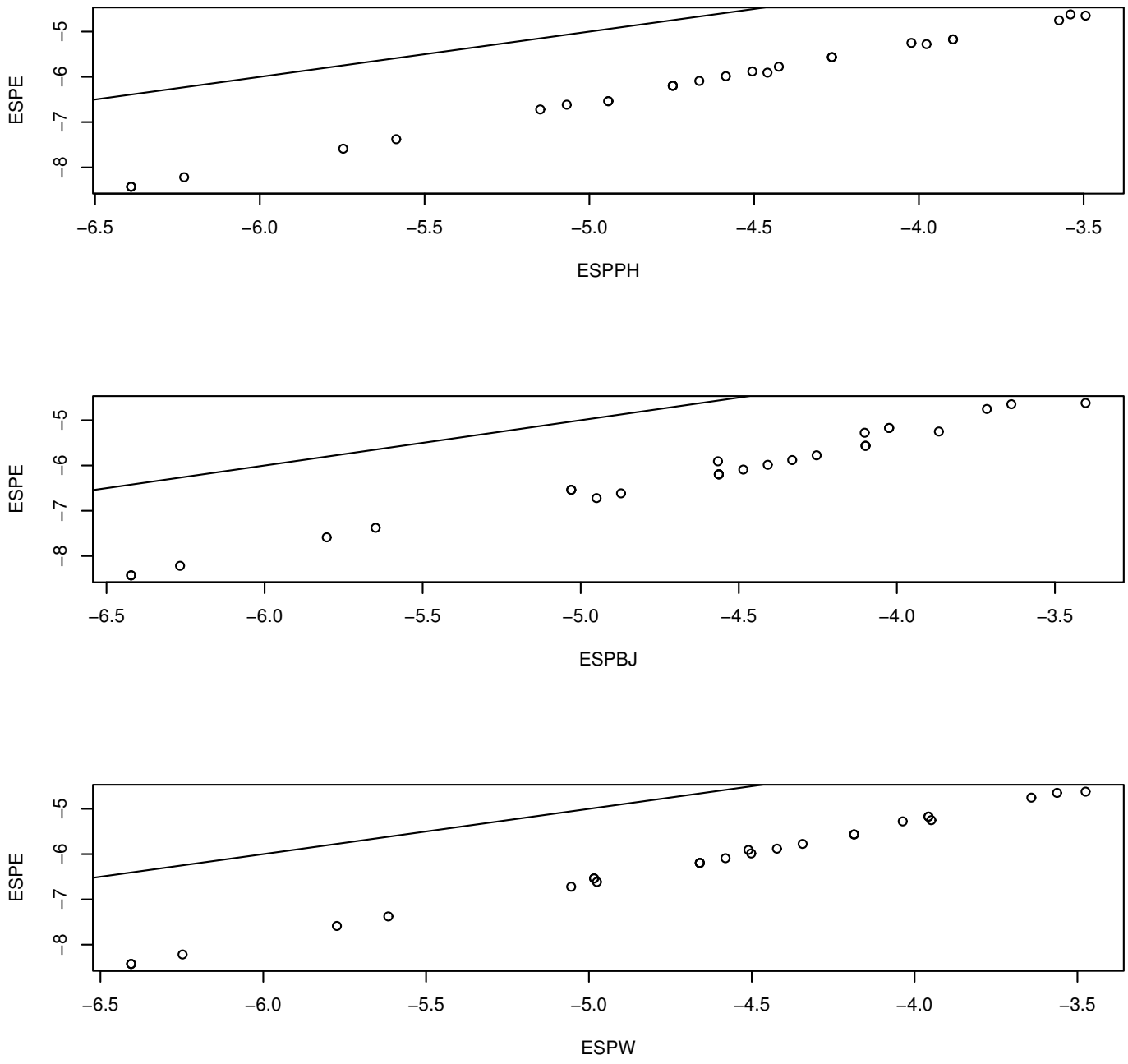

Figure 2: Three EE Plots with the Exponential AFT

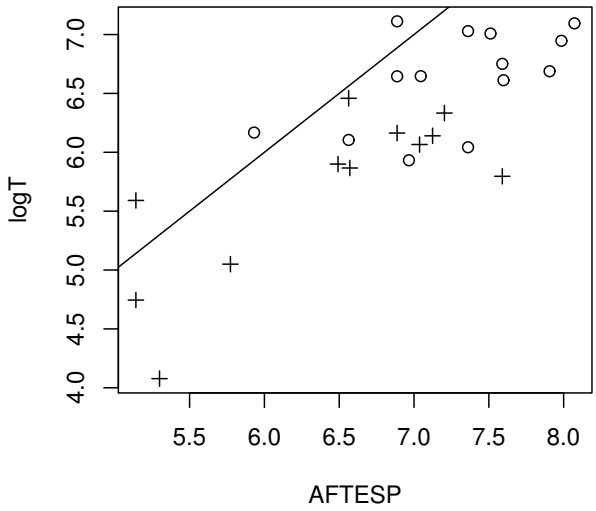

Figure 3: LCR Plot for Ovarian Cancer Data

LCR plot using the Weibull AFT.

Example 2. R contains a data set *nwtco* where the response variable Y is the time until relapse with $n = 4028$. The model used predictors *histol* = tumor histology from central lab, *instit* = tumor histology from local institution, *age* in months, and *stage* of disease from 1 to 4 (treated as a continuous variable). In Figure 4, the Grambsch and Therneau (1994) plots suggest that the Cox model is not valid since not all of the loess curves are flat, and the global test has p-value ≈ 0. The slice survival plot in Figure 5 shows that the Cox survival estimators and Kaplan Meier estimators are nearly identical in the six slices, suggesting that the Cox model is a reasonable approximation to the data. The greatest contributors to lack of fit seem to be the predictors age and stage corresponding to the bottom two plots of Figure 4, and survival for small ESP corresponding to the upper left plot in Figure 5.

5.1 $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z}$ Simulation

R code similar to that of Zhou (2001) was used to generate a Weibull regression data set with parameter vector $\boldsymbol{\beta}_P$. Then the Weibull AFT parameter vector $\boldsymbol{\beta} = \boldsymbol{\beta}_Z = \boldsymbol{\beta}_A = -\sigma\boldsymbol{\beta}_P = -(1/\gamma)\boldsymbol{\beta}_p$. Hence $\boldsymbol{\Sigma}_{\boldsymbol{x}Z} = -\gamma \mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta}_P$. The simulation used $\boldsymbol{\beta}_A = -(1/\gamma, ..., 1/\gamma, 0, ..., 0)^T$ with $p-k$ zeroes and $\boldsymbol{\beta}_P = (1, ..., 1, 0, ..., 0)^T$ with k ones and $p-k$ zeroes. The population $\boldsymbol{\Sigma}_{\boldsymbol{x}Z} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_A$ was computed.

In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_p(\boldsymbol{0}, \boldsymbol{I})$ where the p elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $p \times p$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\mathrm{Cov}(\boldsymbol{x}_i) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1+p\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi+(p-1)\psi^2]$.

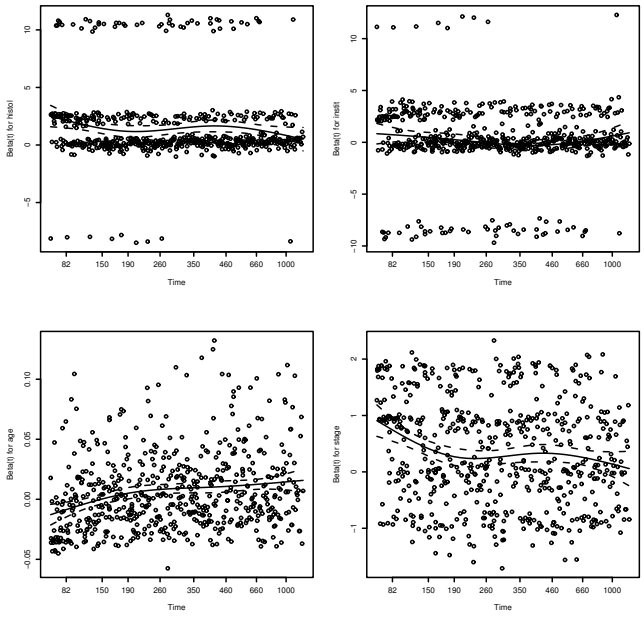

Figure 4: Grambsch and Therneau Plots for NWTCO Data

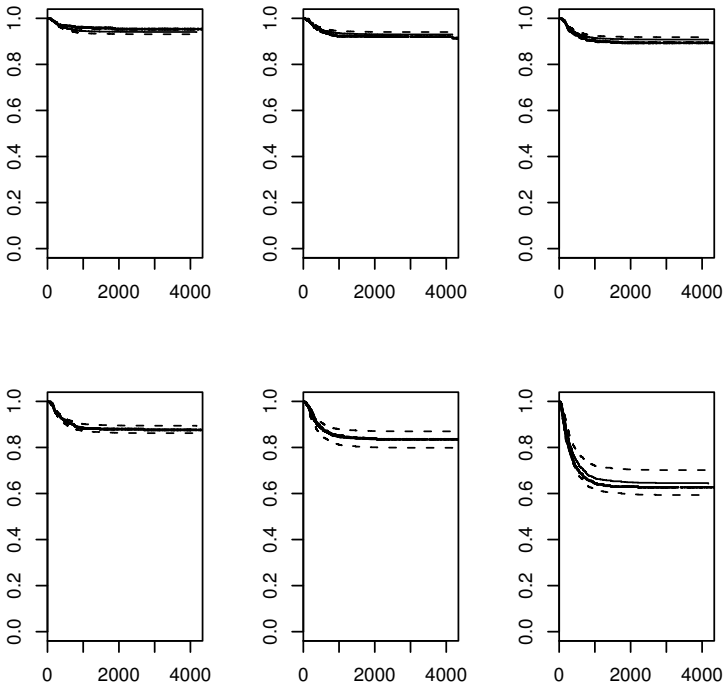

Figure 5: Slice Survival Plot for NWTCO Data: Dashes Are Pointwise CI Bands

Table 1: $\boldsymbol{\Sigma}_{\boldsymbol{x}Z} = (-1, 0, 0, 0)^T$

(n, p, ψ, k)	est	σ_{1Z}	σ_{2Z}	σ_{3Z}	σ_{4Z}
(100,4,0,1)	samp	-0.9993	-0.0022	-0.0024	-0.0010
	SD	0.1935	0.1620	0.1623	0.1630
(100,4,0,1)	AFT	-1.0030	-0.0014	-0.0014	0+
	SD	0.1855	0.1491	0.1496	0.1505
(100,4,0,1)	BJ	-1.0019	-0.0023	-0.0016	-0.0009
	SD	0.2024	0.1688	0.1689	0.1694

Hence the correlations are $cor(x_i, x_j) = \rho = (2\psi + (p-1)\psi^2)/(1 + p\psi^2)$ for $i \neq j$. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors $\boldsymbol{x}_i$ cluster about the line in the direction of $(1, ..., 1)^T$.

Then 5000 runs are used to get the estimators. The means and standard deviations of the estimators are given. In the simulation, the uncensored values of Z are known. Hence the first estimator is the usual sample covariance vector $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z}$. For real data, this estimator can not be computed since only censored values of Z are known. The second estimator is $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z}(A) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_A$ from the Weibull AFT. The third estimator is $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z}(BJ) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{BJ}$ using the Buckley James estimator. Let $\boldsymbol{\Sigma}_{\boldsymbol{x}Z} = (\sigma_{1Z}, ..., \sigma_{pZ})^T$. Table 1 gives 2 lines per simulation scenario. The first line gives the means while the second line gives the standard deviations. A value of 0+ means the absolute value was less than 0.00005.

All three estimators worked well. It is not surprising that a correctly specified AFT would slightly outperform the Buckley James estimator (have the smallest standard deviations).

6 Conclusions

The Harrell (2015) `rms` library is useful for the Buckley James estimator. For more on estimators for model (4), see, for example, Heller and Simonoff (1990), James and Smith (1984), Lai and Ying (1991), Lin and Wei (1992), and Zeng and Lin (2007). The Kaplan Meier estimator is due to Kaplan and Meier (1958).

Under iid cases, $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ even if heterogeneity is present. Hence $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_Z$ where, for example, $\hat{\boldsymbol{\beta}}_Z$ is one of the estimators studied by Yu, Liu, and Chen (2024).

Similar ideas can be used for other censored regression models, such as a Tobit regression, provided that the cases are iid, and the population parameter vector $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$.

Software

The *R* software was used in the simulations. See R Core Team (2024). Programs are in the collection of *R* functions *survpack.txt*, available from (http://parker.ad.siu.edu/Olive/survpack.txt). The function `BJcovxz` generates a Weibull regression data set with right censored survival times. The function `BJcovxzsim` was used for Table 1. The

function `vnwtco` was used to produce plots for Example 2. The function `vovar` is useful for Example 1. The function `bphgfit` makes a slice survival plot if there is a single covariate $x = 1$ for group (treatment) 1 and $x = 0$, otherwise. The function `bphsim3` simulates the above function. The function `phgfit2` makes a slice survival plot when $\boldsymbol{x}$ is a $p \times 1$ vector. Both of these functions are for the Cox model. The function `phsim5` simulates the above function. The function `wphsim` simulates the slice survival plot for Weibull regression. The function `wregsim2` simulates EE plots for Weibull regression.

Some R code for producing the simulation and Figure 1 appears in Johana Lemonge (2025). The two data sets are available from (http://parker.ad.siu.edu/Olive/survdata.txt).

References

Arjas, E. (1988), "A Graphical Method for Assessing Goodness of Fit in Cox's Proportional Hazards Model," *Journal of the American Statistical Association*, 83, 204-212.

Barchard, K.A., and Russell, J.A. (2024), "Distorted Correlations among Censored Data: Causes, Effects, and Correction," *Behavior Research Methods*, 56, 1207-1228.

Buckley, J. and James, I. (1979), "Linear Regression with Censored data," *Biometrika*, 66, 429-436.

Collett, D. (2003), *Modelling Survival Data in Medical Research*, 2nd ed., Chapman & Hall/CRC, Boca Raton, FL.

Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

Edmunson, J.H., Fleming, T.R., Decker, D.G., Malkasian, G.D., Jorgenson, E.O., Jeffries, J.A., Webb, M.J., and Kvols, L.K. (1979), "Different Chemotherapeutic Sensitivities and Host Factors Affecting Prognosis in Advanced Ovarian Carcinoma Versus Minimal Residual Disease," *Cancer Treatment Reports*, 63, 241-247.

Feng, C., Wang, H., Zhang, Y., Han, Y., Liang, Y., and Tu, X.M. (2017), "On Testing Proportionality in the Cox Regression Model by Andersen's Plot," *Communications in Statistics: Theory and Methods*, 46, 3489-3500.

Gill, R.D., and Schumaker, M. (1987), "A Simple Test of the Proportional Hazards Assumption," *Biometrika*, 74, 289-300.

Grambsch, P.M., and Therneau, T.M. (1994), "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals," *Biometrika*, 81, 515-526.

Harrell, F.E. (2015), *Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis*, 2nd ed., Springer, New York, NY.

Heller, G., and Simonoff, J. S. (1990), "A Comparison of Estimators for Regression with a Censored Response Variable," *Biometrika*, 77, 515-520.

Hosmer, D.W., and Lemeshow, S. (1999), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Wiley, NY.

James, I.R., and Smith, P.J. (1984), "Consistency Results for Linear Regression with Censored Data," *The Annals of Statistics*, 12, 590-600.

Johana Lemonge, S. (2025), *OLS Testing with Predictors Scaled to Have Unit Sample Variance*, PhD thesis, Southern Illinois University, (http://parker.ad.siu.edu/Olive/sSanjuka.pdf).

Kaplan, E.L., and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53, 457-481.

Kay, R. (1984), "Goodness-of-Fit Methods for the Proportional Hazards Model: a Review," *Reviews of Epidemiology Santé Publications*, 32, 185-198.

Lai, T.L., and Ying, Z. (1991), "Large-Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data," *Annals of Statistics*, 19, 1370-1402.

Li, Y., Gillespie, B.W., Shedden, K. and Gillespie, J.A. (2018), "Profile Likelihood Estimation of the Correlation Coefficient in the Presence of Left, Right or Interval Censoring and Missing Data," *The R Journal,* 10, 159-179.

Lin, D.Y., and Wei, L.J. (1991), "Goodness-of-Fit Tests for the General Cox Regression Model," *Statistica Sinica*, 1, 1-17.

Lin, J.S., and Wei, L.J. (1992), "Linear Regression Analysis Based on Buckley-James Estimating Equation," *Biometrics*, 48, 679-681.

Lyles, R.H., Fan, D., and Chuachoowong, R. (2001), "Correlation Coefficient Estimation Involving a Left Censored Laboratory Assay Variable," *Statistics in Medicine*, 20, 2921-2933.

Marzec, L., and Marzec, P. (1997), "Generalized Martingale-Residual Processes for Goodness-of-Fit Inference in Cox's Type Regression Models," *The Annals of Statistics*, 25, 683-714.

May, S., and Hosmer, D.W. (1998), "A Simple Method for Calculating a Goodness-of-Fit Test for the Proportional Hazards Model," *Lifetime Data Analysis*, 4, 109-120.

Miller, R. (1981), *Survival Analysis*, Wiley, NY.

Ng'andu, N.H. (1997), "An Empirical Comparison of Statistical Tests for Assessing the Proportional Hazards Assumption of Cox's Model," *Statistics in Medicine*, 16, 611-626.

Pesonen, M., Pesonen, H., and Nevalainen, J. (2015), "Covariance Estimation for Left-Censored Data," *Computational Statistics & Data Analysis*, 92, 13-25.

Quantin, C., Moreau, T., Asselain, B., Maccario, J., and Lellouch, J. (1996), "A Regression Survival Model for Testing the Proportional Hazards Hypothesis," *Biometrics*, 52, 874-885.

R Core Team (2024), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Ritov, Y. (1990), "Estimation in a Linear Regression with Censored Data," *Annals of Statistics*, 18, 303-328.

Smith, P.J. (2002), *Analysis of Failure and Survival Data*, Chapman and Hall/CRC, Boca Raton, FL.

Tsiatis, A.A. (1990), "Estimating Regression Parameters Using Linear Rank Tests for Censored Data," *Annals of Statistics*, 18, 354-372.

van Houwelingen, H.C., Bruinsma, T., Hart, A.A.M. Veer, L.J., and Wessels, L.F.A. (2006), "Cross-validated Cox Regression on Microarray Gene Expression Data," *Statistics in Medicine*, 25, 3201-3216.

Yu, L., Liu, L., and Chen, D.G. (2024), "Extending Buckley-James Method for Heteroscedastic Survival Data," *Journal of Statistical Computation and Simulation*, 94, 1776-1792.

Yu, Q., Chappell, R., Wong, G.Y.C., Hsu, Y., and Mazur, M., (2008), "Relationship Between Cox, Lehmann, Weibull, and Accelerated Lifetime Models," *Communications*

in Statistics Theory and Methods, 37, 1458-1470.

Zeng, D., and Lin, D.Y. (2007), "Efficient Estimation for the Accelerated Failure Time Model," *Journal of the American Statistical Association*, 102, 1387-1396.

Zhou, M. (2001), "Understanding the Cox Regression Models with Time–Change Covariates," *The American Statistician*, 55, 153-155.