Estimating Covariances for Some Survival Regression Models

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

July 30, 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.

KEY WORDS: Accelerated failure time models, Buckley James estimator, Multiple linear regression, Weibull regression.

1 INTRODUCTION

This section reviews some survival regression 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 $SP = \boldsymbol{x}^T \boldsymbol{\beta}$, and let the estimated sufficient predictor $ESP = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}$.

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 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

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

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

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 $\text{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma}_{\boldsymbol{e}} = diag(\sigma_i^2) = diag(\sigma_1^2, ..., \sigma_n^2)$ 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 $\boldsymbol{\phi}$. See, for example, White (1980).

For estimation with ordinary least squares, let the covariance matrix of $\boldsymbol{x}$ be $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T]$ and $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = E[(\boldsymbol{x} - E(\boldsymbol{x})(Y - E(Y))]$. 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 nonparametric 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.

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

This section derives an estimator for $\boldsymbol{\Sigma}_{\boldsymbol{x}Z} = \mathrm{Cov}(\boldsymbol{x}, Z)$ 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, $\boldsymbol{\beta}_Z = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Z}$. Thus $\boldsymbol{\Sigma}_{\boldsymbol{x}Z} = \mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta}_Z = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_Z$. When the response Y_i is censored, several models give consistent estimators $\hat{\boldsymbol{\beta}}_Z$ of $\boldsymbol{\beta}_Z$. Hence

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_Z. \tag{5}$$

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 $|\widehat{\mathrm{Cov}}(x_i, Z)|/\sqrt{\hat{V}(x_i)}$.

3 Example and Simulations

It is important to check that a parametric AFT model is reasonable with the Buckley James before using Equation (5). 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. The identity line is included in the EE plots as a visual aid.

Example. 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 $|\widehat{Cor}(treat, Z)| \approx 3|\widehat{Cor}(age, Z)|$.

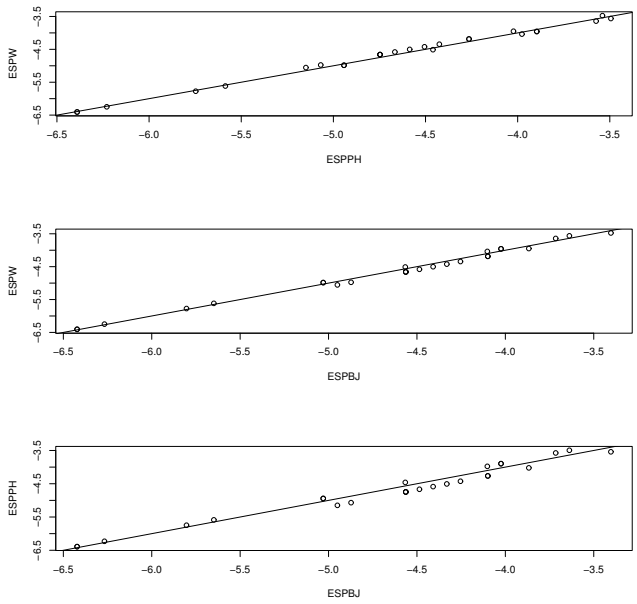

Figure 1: Three EE Plots for the Ovarian Cancer Data

3.1 $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z}$ 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 $\boldsymbol{\Sigma}_{\boldsymbol{x}Z} = -\gamma \text{Cov}(\boldsymbol{x})\boldsymbol{\beta}_P$. 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 $\text{Cov}(\boldsymbol{x}_i) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1+p\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi+(p-1)\psi^2]$. 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 that 0.00005.

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

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).

4 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), Lai and Ying (1991), Lin and Wei (1992), and Yu, Liu, and Chen (2024).

Under iid cases, $\hat{\boldsymbol{\beta}}_{OLS}$ still estimates $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ when 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).

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, Gillespie, Shedden, and Gillespie (2018), and Lyles, Fan, and Chuachoowong (2001).

Software

The *R* software was used in the simulations. See R Core Team (2024). Programs are in the Olive (2025) 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 using a method similar to that of Zhou (2001).

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

References

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.

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.

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).

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, 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.

Olive, D.J. (2025), *Survival Analysis*, online course notes, see (http://parker.ad.siu.edu/Olive/survbk.htm).

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

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

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

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