8-2022

Survival Analysis And The OLS AFT

Sanjuka Lemonge
sanjuka.johanalemonge@siu.edu

SURVIVAL ANALYSIS AND THE OLS AFT

by

Sanjuka Johana Lemonge

B.S., University of Kelaniya, 2019

A Research Paper
Submitted in Partial Fulfillment of the Requirements for the
Master of Science Degree

School of Mathematical and Statistical Sciences
in the Graduate School
Southern Illinois University Carbondale
August, 2022

RESEARCH PAPER APPROVAL

SURVIVAL ANALYSIS AND THE OLS AFT

by

Sanjuka Johana Lemonge

A Research Paper Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Mathematics

Approved by:

David J. Olive

Michael Sullivan

Yaser Samadi

Graduate School
Southern Illinois University Carbondale
July 1, 2022

AN ABSTRACT OF THE RESEARCH PAPER OF

SANJUKA JOAHANA LEMONGE, for the Master of Science degree in MATHEMATICS, presented on JULY 1, 2022, at Southern Illinois University Carbondale.

TITLE: SURVIVAL ANALYSIS AND THE OLS AFT

MAJOR PROFESSOR: Dr. David J. Olive

This research paper examines testing the accelerated failure time (AFT) survival regression model with ordinary least squares (OLS), and presents one case studies using survival analysis.

KEY WORDS: Survival Analysis, AFT, OLS

ACKNOWLEDGMENTS

I would like to take this opportunity to thank my research advisor, Dr. David Olive for overseeing my Master's project and Dr. Michael Sullivan and Dr. Yaser Samadi for sitting on my committee. I would also like to thank all the professors of School of Mathematical and Statistical Sciences, SIU for their instruction and care over the past two years. Finally I want to say to my family, thank you for all of your support and encouragement. You have pushed me to succeed and I could not have done it without you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

SURVIVAL ANALYSIS

The first four chapters follows Olive (2022a) closely.

Definition 1. Let $Y \geq 0$ be the time until an event occurs. Then Y is called the survival time or time until event. The survival time is censored if the event of interest has not been observed. Let Y_i be the ith survival time. Let Z_i be the time the ith observation (possibly an individual or machine) leaves the study for any reason other than the event of interest. Then Z_i is the time until the ith observation is censored. Then the right censored survival time T_i of the ith observation is $T_i = \min(Y_i, Z_i)$. Let $\delta_i = 0$ if T_i is (right) censored ($T_i = Z_i$) and let $\delta_i = 1$ if T_i is not censored ($T_i = Y_i$). Then the univariate survival analysis data is $(T_1, \delta_1), (T_2, \delta_2), ..., (T_n, \delta_n)$. Alternatively, the data is $T_1, T_2^*, T_3, ..., T_{n-1}^*, T_n$ where the * means that the case was (right) censored. Sometimes the asterisk * is replaced by a plus +, and Y_i, y_i or t_i can replace T_i.

Definition 2. i) The cumulative distribution function (cdf) of Y is $F(t) = P(Y \leq t)$. Since $Y \geq 0$, $F(0) = 0, F(\infty) = 1$, and $F(t)$ is nondecreasing.

ii) The probability density function (pdf) of Y is $f(t) = F'(t)$.

iii) The survival function of Y is $S(t) = P(Y > t)$. $S(0) = 1, S(\infty) = 0$ and $S(t)$ is nonincreasing.

iv) The hazard function of Y is $h(t) = \dfrac{f(t)}{1 - F(t)}$ for $t > 0$ and $F(t) < 1$. Note that $h(t) \geq 0$ if $F(t) < 1$.

v) The cumulative hazard function of Y is $H(t) = \int_0^t h(u)du$ for $t > 0$. It is true that $H(0) = 0, H(\infty) = \infty$, and $H(t)$ is nondecreasing.

Assume $Y \geq 0$. Then $F(0) = 0$, $S(0) = 1$, and $H(0) = 1$. Note that $S(\infty) = 0$ implies that $H(\infty) = \infty$ where $\lim_{t \to \infty} H(t) = H(\infty)$. Note that $0 \leq F(t) \leq 1$, $0 \leq S(t) \leq 1$, $f(t) \geq 0$, $h(t) \geq 0$, and $H(t) \geq 0$.

Given one of $F(t), f(t), S(t), h(t)$ or $H(t)$, the following theorem shows how to find

the other 4 quantities for $t > 0$. Each of these five quanities completely determines the distribution of the random variable. In reliability analysis, the *reliability function* $R(t) = S(t)$, and in economics, Mill's ratio $= 1/h(t)$. In actuarial sciences, $h(t)$ is the *force of mortality*.

Theorem 1.

A) $F(t) = \int_0^t f(u)du = 1 - S(t) = 1 - \exp[-H(t)] = 1 - \exp[-\int_0^t h(u)du]$.

B) $f(t) = F'(t) = -S'(t) = h(t)[1 - F(t)] = h(t)S(t) = h(t)\exp[-H(t)] = H'(t)\exp[-H(t)]$.

C) $S(t) = 1 - F(t) = 1 - \int_0^t f(u)du = \int_t^\infty f(u)du = \exp[-H(t)] = \exp[-\int_0^t h(u)du]$.

D)

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} = \frac{-S'(t)}{S(t)} = - \frac{d}{dt}\log[S(t)] = H'(t).$$

E) $H(t) = \int_0^t h(u)du = -\log[S(t)] = -\log[1 - F(t)]$.

Example 1. Suppose $Y \sim EXP(\lambda)$ where $\lambda > 0$, then $h(t) = \lambda$ for $t > 0$, $f(t) = \lambda e^{-\lambda t}$ for $t > 0$, $F(t) = 1 - e^{-\lambda t}$ for $t > 0$, $S(t) = e^{-\lambda t}$ for $t > 0$, $H(t) = \lambda t$ for $t > 0$ and $E(Y) = 1/\lambda$. The exponential distribution is the only distribution of a continuous random variable Y with a constant hazard function since $h(t)$ completely determines the distribution of the random variable Y. Derive $H(t), S(t), F(t)$, and $f(t)$ from the constant hazard function $h(t) = \lambda$ for $t > 0$ and some $\lambda > 0$.

Solution: $H(t) = \int_0^t h(u)du = \int_0^t \lambda du = \lambda t$ for $t > 0$.

$S(t) = e^{-H(t)} = e^{-\lambda t}$, for $t > 0$.

$F(t) = 1 - S(t) = 1 - e^{-\lambda t}$ for $t > 0$.

Finally, $f(t) = h(t)S(t) = \lambda e^{-\lambda t} = F'(t)$ for $t > 0$.

Example 2. If $Y \sim$ Weibull(γ, λ) where $\gamma > 0$ and $\lambda > 0$, then $h(t) = \lambda\gamma t^{\gamma-1}$ for $t > 0$, $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^\gamma)$ for $t > 0$, $F(t) = 1 - \exp(-\lambda t^\gamma)$ for $t > 0$, $S(t) = \exp(-\lambda t^\gamma)$ for $t > 0$, $H(t) = \lambda t^\gamma$ for $t > 0$. The Weibull$(\lambda, \gamma = 1)$ distribution is the EXP(λ) distribution. The hazard function can be increasing, decreasing or constant. Hence the Weibull distribution

often fits reliability data well, and the Weibull distribution is an important distribution in reliability analysis. Derive $H(t), S(t), F(t)$, and $f(t)$ if $Y \sim \text{Weibull}(\lambda, \gamma)$.

Solution:

$$H(t) = \int_0^t h(u)du = \int_0^t \lambda\gamma u^{\gamma-1}du = \lambda\gamma\frac{u^\gamma}{\gamma}\Big|_0^t = \lambda t^\gamma \quad \text{for} \quad \text{t} > 0.$$

$S(t) = \exp[-H(t)] = \exp[-\lambda t^\gamma]$, for $t > 0$.

$F(t) = 1 - S(t) = 1 - \exp[-\lambda t^\gamma]$ for $t > 0$.

Finally, $f(t) = h(t)S(t) = \lambda\gamma t^{\gamma-1}\exp[-\lambda t^\gamma]$ for $t > 0$.

CHAPTER 2

WEIBULL AND EXPONENTIAL REGRESSION

In a *1D regression model*, the response variable Y is conditionally independent of the $p \times 1$ vector of predictors $\boldsymbol{x}$ given the sufficient predictor $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | SP \text{ or } \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x} | \mathrm{h}(\boldsymbol{x}), \tag{2.1}$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The estimated sufficient predictor $ESP = \hat{h}(\boldsymbol{x})$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ where $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$.

Definition 3. For parametric proportional hazards regression models, the baseline function is parametric and the parameters are estimated via maximum likelihood. Then as a 1D regression model, $SP = \boldsymbol{\beta}_P^T\boldsymbol{x}$, and the hazard function

$$h_{Y|SP}(t) \equiv h_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}_P^T\boldsymbol{x})h_{0,P}(t) = \exp(SP)h_{0,P}(t)$$

where the parametric baseline function $h_{0,P}$ depends on k unknown parameters but does not depend on the predictors $\boldsymbol{x}$. The survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = [S_{0,P}(t)]^{\exp(\boldsymbol{\beta}_P^T\boldsymbol{x})} = [S_{0,P}(t)]^{\exp(SP)}, \tag{2.2}$$

and

$$\hat{S}_{\boldsymbol{x}}(t) = [\hat{S}_{0,P}(t)]^{\exp(\hat{\boldsymbol{\beta}}_P^T\boldsymbol{x})} = [\hat{S}_{0,P}(t)]^{\exp(ESP)}. \tag{2.3}$$

The following univariate results will be useful for Exponential and Weibull regression. If Y has a Weibull distribution, $Y \sim W(\gamma, \lambda)$, then $S_Y(t) = \exp(-\lambda t^\gamma)$ where t, λ and γ are positive. If $\gamma = 1$, then Y has an Exponential distribution, $Y \sim EXP(\lambda)$ where $E(Y) = 1/\lambda$. Now V has a smallest extreme value distribution, $V \sim SEV(\theta, \sigma)$, if

$$S_V(t) = P(V > t) = \exp\left(-\exp\left(\frac{t-\theta}{\sigma}\right)\right)$$

where $\sigma > 0$ while t and θ are real. If $Z \sim SEV(0,1)$, then $V = \theta + \sigma Z \sim SEV(\theta, \sigma)$ since the SEV distribution is a location scale family. Also, $V = \log(Y) \sim SEV(\theta = -\sigma \log(\lambda), \sigma = 1/\gamma)$, and $Y = e^V \sim W(\gamma = 1/\sigma, \lambda = e^{-\theta/\sigma})$.

If Y_i follows a Weibull regression model, then $\log(Y_i)$ follows an accelerated failure time (AFT) model: $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i$ where the e_i are iid $SEV(0,1)$, and $\log(Y)|\boldsymbol{x} \sim SEV(\alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}, \sigma)$.

Definition 4. The Weibull proportional hazards regression (WPH) model or Weibull regression model is a parametric proportional hazards model with $Y|\boldsymbol{x} \sim W(\gamma = 1/\sigma, \lambda_{\boldsymbol{x}})$ where

$$\lambda_{\boldsymbol{x}} = \exp\left[-\left(\frac{\alpha}{\sigma} + \frac{\boldsymbol{\beta}_A^T \boldsymbol{x}}{\sigma}\right)\right] = \lambda_0 \exp(\boldsymbol{\beta}_P^T \boldsymbol{x})$$

with $\lambda_0 = \exp(-\alpha/\sigma)$ and $\boldsymbol{\beta}_P = -\boldsymbol{\beta}_A/\sigma$. Thus for $t > 0$, $P(Y > t|\boldsymbol{x}) =$

$$S_{\boldsymbol{x}}(t) = \exp(-\lambda_{\boldsymbol{x}} t^\gamma) = \exp(-\lambda_0 \exp(\boldsymbol{\beta}_P^T \boldsymbol{x}) t^\gamma) = [\exp(-\lambda_0 t^\gamma)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})} =$$

$$[S_{0,P}(t)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})}.$$

As a 1D regression model, $Y|SP \sim W(\gamma, \lambda_0 \exp(SP))$. Also,

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}_P^T \boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}_P^T \boldsymbol{x}_i) h_0(t)$$

where $h_0(t) = h_0(t|\boldsymbol{\theta}) = \lambda_0 \gamma t^{\gamma - 1}$ is the Weibull baseline function. Exponential regression is the special case of Weibull regression where $\sigma = 1$. Hence $Y|\boldsymbol{x} \sim W(1, \lambda_{\boldsymbol{x}}) \sim EXP(\lambda_{\boldsymbol{x}})$.

CHAPTER 3

ACCELERATED FAILURE TIME MODELS

Definition 5. For a parametric *accelerated failure time* model,

$$\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i \tag{3.1}$$

where the e_i are iid from a location scale family. Let $SP = \boldsymbol{\beta}_A^T \boldsymbol{x}$. Then as a 1D regression model, $\log(Y)|SP = \alpha + SP + e$. The parameters are again estimated by maximum likelihood and the survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|\boldsymbol{x}}(t) = S_0\left(\frac{t}{\exp(\boldsymbol{\beta}_A^T \boldsymbol{x})}\right),$$

and

$$\hat{S}_{\boldsymbol{x}}(t) = \hat{S}_0\left(\frac{t}{\exp(\hat{\boldsymbol{\beta}}_A^T \boldsymbol{x})}\right)$$

where $\hat{S}_0(t)$ depends on $\hat{\alpha}$ and $\hat{\sigma}$.

For the AFT model, $h_i(t) = h_{\boldsymbol{x}}(t) = e^{-SP} h_0(t/e^{SP})$ and $S_i(t) = S_{\boldsymbol{x}}(t) = S_0(t/\exp(SP))$ where $SP = \boldsymbol{\beta}_A^T \boldsymbol{x}$. If $S_{\boldsymbol{x}}(t_{\boldsymbol{x}}(\rho)) = 1 - \rho$ for $0 < \rho < 1$, then $t_{\boldsymbol{x}}(\rho)$ is the ρth percentile. For the accelerated failure time model,

$$t_{\boldsymbol{x}}(\rho) = t_0(\rho) \exp(\boldsymbol{\beta}_A^T \boldsymbol{x})$$

where $t_0(\rho) = \exp(\sigma e_i(\rho) + \alpha)$ and $S_{e_i}(e_i(\rho)) = P(e_i > e_i(\rho)) = 1 - \rho$. Note that the estimated percentile ratio is free of ρ, $\hat{\sigma}$ and $\hat{\alpha}$

$$\frac{\hat{t}_{\boldsymbol{x}_1}(\rho)}{\hat{t}_{\boldsymbol{x}_2}(\rho)} = \exp(\hat{\boldsymbol{\beta}}_A^T(\boldsymbol{x}_1 - \boldsymbol{x}_2)).$$

The *acceleration factor* $= e^{-SP}$ and $t_{0,\rho} = e^{-SP} t_{\boldsymbol{x},\rho}$. The median survival times are related by $t_{0,0.5} = e^{-SP} t_{\boldsymbol{x},0.5}$. If $e^{-SP} < 1$, then the median survival time of $\boldsymbol{x} >$ the median survival time of $\mathbf{0}$, a result that is good if the event is death, but bad if the event is time until recovery. Note that $H_{\boldsymbol{x}}(t) = -\log S_{\boldsymbol{x}}(t) = -\log(S_0(t/e^{SP})) = H_0(t/e^{SP})$.

Remark 1. Assume $x_i > 0$. Then $\beta_i > 0$ increases $\log(Y_i)$ and Y_i, while $\beta_i < 0$ decreases $\log(Y_i)$ and Y_i. For the Cox PH regression model, $h_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}^T\boldsymbol{x})h_0(t)$. Hence $\beta_i > 0$ increases hazard and decreases Y_i, while $\beta_i < 0$ decreases hazard and increases Y_i. In the WPH model, $\boldsymbol{\beta}_P = -\boldsymbol{\beta}_A/\sigma$.

Definition 6. The *Weibull AFT* satisfies $\log(Y)|(\alpha + \boldsymbol{\beta}_A^T\boldsymbol{x}) \sim SEV(\alpha + \boldsymbol{\beta}_A^T\boldsymbol{x}, \sigma)$. The *Exponential AFT* is the special case with $\sigma = 1$.

Theorem 2. Weibull regression models, including Exponential regression models, are the only models where Y follows a proportional hazards regression model and $\log(Y)$ follows an accelerated failure time model.

If the Weibull PH regression model holds for Y_i, then $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T\boldsymbol{x}_i + \sigma e_i$ where $e_i \sim SEV(0, 1)$. Thus $\log(Y)|\boldsymbol{x} \sim SEV(\alpha + \boldsymbol{\beta}_A^T\boldsymbol{x}, \sigma)$, and the $\log(Y_i)$ follows a parametric accelerated failure time model. Two other important AFTs are i) the lognormal AFT where $\log(Y)|\boldsymbol{x} \sim N(\alpha + \boldsymbol{\beta}_A^T\boldsymbol{x}, \sigma^2)$ where the Y_i are lognormal and the $e_i \sim N(0, 1)$ are normal, and ii) the loglogistic AFT where $\log(Y)|\boldsymbol{x} \sim L(\alpha + \boldsymbol{\beta}_A^T\boldsymbol{x}, \sigma)$ 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.

A case consists of the measurements on a person or thing. Let $(\boldsymbol{x}_i^T, Y_i)^T$ be the ith case. For example, people sick from a deadly disease who go to 3 hospitals, where Y_i is the survival time. As noted by Olive (2022b), if the cases are iid and the censoring is independent of the cases, then the uncensored cases $(\boldsymbol{x}_i^T, Y_i)^T$ (where the Y_i are uncensored) may not follow the multiple linear regression model since the censoring causes the Y_i to follow a truncated distribution. However, OLS may be useful for testing $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$.

CHAPTER 4

EXAMPLES AND SIMULATIONS

Example.

A small simulation was done using 5000 runs. So an observed coverage in [0.94, 0.96] gives no reason to doubt that the confidence interval (CI) has the nominal coverage of 0.95. For Weibull regression, the cases were $(Z_i, \delta_i, \boldsymbol{x}_i)$ where $Z_i = Y_i$ is uncensored if $\delta_i = 1$, and Z_i is right censored if $\delta_i = 0$. Hence $Z_i = T_i$. R code similar to that of Zhou (2001) was used to generate data from the Weibull proportional hazards regression model. Then $SP = \boldsymbol{x}_i^T \boldsymbol{\beta} = 1x_{i,1} + \cdots + 1x_{i,k} \sim N(0, a^2)$ for $i = 1, ..., n$. The simulations used $a = 1$ where $\boldsymbol{\beta} = \boldsymbol{\beta}_P = (1, ..., 1, 0, ..., 0)^T$ with k ones and $p - k$ zeros. Then $\boldsymbol{\beta}_A = (-1/\gamma, ..., -1/\gamma, 0, ..., 0)^T$ where $\gamma = 1/\sigma$ for the Weibull regression model.

Let $\boldsymbol{x} = (1 \ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $p \times 1$ vector of nontrivial predictors. 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 $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\text{Cov}(\boldsymbol{z}_i) = \boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1 + (p-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi + (p-2)\psi^2]$. Hence the correlations are $cor(z_i, z_j) = \rho = (2\psi + (p-2)\psi^2)/(1 + (p-1)\psi^2)$ for $i \neq j$. Then $\sum_{j=1}^{k} z_j \sim N(0, k\sigma_{ii} + k(k-1)\sigma_{ij}) = N(0, v^2)$. Let $\boldsymbol{u} = a\boldsymbol{z}/v$. Then $cor(x_i, x_j) = \rho$ for $i \neq j$ where x_i and x_j are nontrivial predictors. 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 $\boldsymbol{u}_i$ cluster about the line in the direction of $(1, ..., 1)^T$. Then $SP = \boldsymbol{x}_i^T \boldsymbol{\beta} = 1x_{i,1} + \cdots + 1x_{i,k} \sim N(0, a^2)$ for $i = 1, ..., n$.

If the cases are iid from some population, we conjecture that the OLS fit to the uncensored cases gives a consistent estimator for $\boldsymbol{\gamma}$ where $\boldsymbol{\gamma}$ is a biased estimator of $\boldsymbol{\beta}$. Then we can test $H_0 : (\beta_{i1}, ..., \beta_{ik})^T = \boldsymbol{0}$ with $H_0 : (\gamma_{i1}, ..., \gamma_{ik})^T = \boldsymbol{0}$ using the OLS fit to the uncensored cases. Testing $H_0 : \beta_1 - \beta_2 = 0$ with $H_0 : \gamma_1 - \gamma_2 = 0$ may not be possible due to the bias of the estimator.

Simulations were done in *R*. See R Core Team (2018). The *R* code below gives some output. The simulation used 5000 runs, n=100,1000, p= 4,8, k=1,2, $\psi = 0, 1/\sqrt{p}, 0.9$, and $\gamma = 0.1, 0.2, 1, 5, 10$. In the program arguments, gam = γ and psi = ψ. For the output shown below, $\boldsymbol{\beta}_A = (-0.2, 0.0, 0.0, 0.0)^T$. OLS confidence intervals were made for β_i where $\beta_0 = \alpha$. The coverage was the proportion of runs the CI contained β_i, but the CI for α gave the proportion of runs where the CI contained 0, corresponding the test $H_0 : \alpha = 0$. Since H_0 is not true for α, we would like the coverage to be close to 0, where the power = 1− coverage.. The program also did the OLS partial F test for H_0: the reduced model is good, where the reduced model contained α and the first k predictors. Hence H_0 was true. We expect that the coverage for the first k β_i may not be close to the nominal 0.95, while the coverage for the last $p - k$ β_i may be close to the nominal 0.95. The coverage for the partial F test, redtest, was the proportion of runs where H_0 was not rejected. Again we expect that the coverage may be close to the nominal 0.95. The average confidence interval (CI) lengths were scaled by multiplying the CI length by $\sqrt{n}$. Two line for each run were given. The first line corresponds to the coverages, while the second line corresponds to the scaled CI lengths. CIs for α, β_1, β_2, β_{p-1}, and β_p were tabled.

```
source("http://parker.ad.siu.edu/Olive/slpack.txt")
args(aftolssim)
function (n = 100, p = 4, k = 1, nruns = 100, psi = 0, a = 1,
    gam = 5,  clam = 0.1, alpha = 0.05)


aftolssim(n=100,p=4,nruns=5000,psi=0,gam=5)
$olslen
[1] 1.097681 1.108201 1.105579 1.106598 1.105377
$olscov
[1] 0.0018 0.9498 0.9480 0.9444 0.9528
$redcov
```

```
[1] 0.9482

$betaaft

[1]  0.0 -0.2  0.0  0.0  0.0

$k

[1] 1

$coef

   Intercept            X1            X2            X3            X4

-0.148247421 -0.170822942  0.027490842 -0.002682143  0.009487977

$nunc

[1] 91.037
```

Table 4.1. AFT with OLS

n	p	k	ψ	α	β_1	β_2	β_{p-1}	β_p	Ftest
100	4	1	0	0.0018	0.9498	0.9480	0.9444	0.9528	0.9482
len				1.0977	1.1082	1.1056	1.1066	1.1055	

CHAPTER 5

SIMULATION RESULTS

Table 5.1. p=4,gamma=5

n	p	k	ψ	α	β_1	β_2	β_{p-1}	β_p	Ftest
100	4	1	0	0.0032	0.9456	0.9496	0.9512	0.9498	0.9476
len				1.0948	1.1071	1.1023	1.1029	1.1037	
1000	4	1	0	0	0.9448	0.9524	0.947	0.9536	0.9484
len				1.0659	1.0684	1.0668	1.0663	1.0667	
100	4	2	0	0.0036	0.9486	0.9474	0.9486	0.9486	0.95
len				1.0977	1.5671	1.5679	1.5655	1.5654	
1000	4	2	0	0	0.9522	0.9502	0.9506	0.9474	0.9518
len				1.0657	1.5096	1.5095	1.5088	1.5089	
100	4	1	0.9	0.0032	0.9526	0.948	0.9524	0.9518	0.9496
len				1.0973	17.7497	17.7533	17.7811	17.7482	
1000	4	1	0.9	0	0.9526	0.9518	0.949	0.95	0.9564
len				1.0655	17.1021	17.0985	17.0956	17.0991	
100	4	2	0.9	0.0036	0.9502	0.954	0.9432	0.9512	0.9466
len				1.0994	35.5643	35.5047	35.5607	35.5601	
1000	4	2	0.9	0	0.9482	0.9466	0.9496	0.9512	0.9486
len				1.0656	34.1852	34.1754	34.1812	34.1888	
100	4	1	$1/\sqrt{p}$	0.001	0.9542	0.953	0.9512	0.9502	0.9556
len				1.0961	2.5530	2.5524	2.5562	2.5480	
1000	4	1	$1/\sqrt{p}$	0	0.9516	0.9474	0.9546	0.951	0.9496
len				1.0657	2.4607	2.4584	2.4611	2.4602	
100	4	2	$1/\sqrt{p}$	0.0028	0.9486	0.9484	0.9484	0.945	0.9488
len				1.0950	4.9017	4.9140	4.8997	4.9138	
1000	4	2	$1/\sqrt{p}$	0	0.9534	0.952	0.9506	0.9474	0.9516
len				1.0653	4.7374	4.7405	4.7390	4.7391	

Table 5.2. p=8,gamma=5

n	p	k	ψ	α	β_1	β_2	β_{p-1}	β_p	Ftest
100	8	1	0	0.0048	0.9516	0.9464	0.9478	0.9542	0.9484
len				1.1265	1.1387	1.1361	1.1361	1.1366	
1000	8	1	0	0	0.9512	0.9502	0.9508	0.9536	0.9464
len				1.0676	1.0711	1.0687	1.0682	1.0690	
100	8	2	0	0.0032	0.954	0.9512	0.9496	0.9458	0.9494
len				1.1269	1.6085	1.6079	1.6046	1.6058	
1000	8	2	0	0	0.944	0.9546	0.9538	0.9446	0.9486
len				1.0682	1.5134	1.5133	1.5111	1.5114	
100	8	1	0.9	0.0058	0.9496	0.948	0.9588	0.9462	0.9456
len				1.1252	27.4294	27.3914	27.3715	27.3836	
1000	8	1	0.9	0	0.957	0.9464	0.9534	0.9486	0.95
len				1.0671	25.8083	25.7993	25.8110	25.7976	
100	8	2	0.9	0.0042	0.947	0.952	0.9446	0.9526	0.9532
len				1.1241	54.7259	54.7746	54.8008	54.7028	
1000	8	2	0.9	0	0.9544	0.9532	0.9502	0.9572	0.9554
len				1.0689	51.6534	51.6679	51.6588	51.6562	
100	8	1	$1/\sqrt{p}$	0.0062	0.9472	0.9556	0.9492	0.944	0.9458
len				1.1227	3.5160	3.5120	3.5150	3.5184	
1000	8	1	$1/\sqrt{p}$	0	0.9502	0.9508	0.9498	0.9474	0.9492
len				1.0672	3.3171	3.3134	3.3139	3.3150	
100	8	2	$1/\sqrt{p}$	0.0052	0.9504	0.9516	0.9542	0.9552	0.9522
len				1.1244	6.8857	6.8749	6.8745	6.8714	
1000	8	2	$1/\sqrt{p}$	0	0.9452	0.9518	0.9506	0.9528	0.9506
len				1.0683	6.4884	6.4862	6.4878	6.4837	

Table 5.3. p=4,gamma=10

n	p	k	ψ	α	β_1	β_2	β_{p-1}	β_p	Ftest
100	4	1	0	0.0048	0.9456	0.9546	0.944	0.9524	0.9538
len				0.5462	0.5512	0.5497	0.5507	0.5519	
1000	4	1	0	0	0.9504	0.95	0.9504	0.953	0.9532
len				0.5320	0.5327	0.5326	0.5324	0.5325	
100	4	2	0	0.0036	0.949	0.9506	0.9554	0.9502	0.9532
len				0.5467	0.7802	0.7792	0.7799	0.7780	
1000	4	2	0	0	0.949	0.9528	0.9568	0.9496	0.9546
len				0.5316	0.7526	0.7527	0.7524	0.7524	
100	4	1	0.9	0.0046	0.9524	0.951	0.951	0.9456	0.9496
len				0.5468	8.8284	8.8537	8.8728	8.8472	
1000	4	1	0.9	0	0.9488	0.9522	0.9428	0.946	0.946
len				0.5313	8.5238	8.5305	8.5315	8.5332	
100	4	2	0.9	0.0054	0.9482	0.955	0.9566	0.9492	0.9506
len				0.5471	17.6864	17.6935	17.7191	17.7183	
1000	4	2	0.9	0	0.9452	0.9542	0.9516	0.947	0.953
len				0.5315	17.0573	17.0561	17.0600	17.0525	
100	4	1	$1/\sqrt{p}$	0.0046	0.9496	0.9482	0.9476	0.9506	0.945
len				0.5460	1.2682	1.2717	1.2696	1.2689	
1000	4	1	$1/\sqrt{p}$	0	0.9482	0.9442	0.9496	0.9514	0.9446
len				0.5318	1.2281	1.2277	1.2282	1.2282	
100	4	2	$1/\sqrt{p}$	0.0056	0.9504	0.9526	0.9522	0.9536	0.9496
len				0.5474	2.4567	2.4504	2.4517	2.4550	
1000	4	2	$1/\sqrt{p}$	0	0.9498	0.9506	0.9536	0.9498	0.9548
len				0.5322	2.3676	2.3680	2.3682	2.3680	

Table 5.4. p=8,gamma=10

n	p	k	ψ	α	β_1	β_2	β_{p-1}	β_p	Ftest
100	8	1	0	0.0062	0.953	0.9448	0.9488	0.947	0.9434
len				0.5619	0.5666	0.5666	0.5656	0.5666	
1000	8	1	0	0	0.9474	0.9508	0.9508	0.953	0.9548
len				0.5327	0.5336	0.5332	0.5332	0.5329	
100	8	2	0	0.0072	0.9508	0.95	0.9516	0.9472	0.9458
len				0.5598	0.7978	0.7991	0.7963	0.7988	
1000	8	2	0	0	0.9484	0.95	0.9552	0.9472	0.9514
len				0.5332	0.7550	0.7550	0.7546	0.7546	
100	8	1	0.9	0.0074	0.95	0.9542	0.9494	0.9498	0.9542
len				0.5598	13.6062	13.6213	13.6545	13.6824	
1000	8	1	0.9	0	0.9492	0.9444	0.9494	0.9504	0.9492
len				0.5331	12.8792	12.8896	12.8873	12.8843	
100	8	2	0.9	0.0086	0.9456	0.955	0.9474	0.9468	0.9454
len				0.5597	27.2493	27.2192	27.3153	27.3061	
1000	8	2	0.9	0	0.9508	0.9432	0.9466	0.9504	0.9514
len				0.5331	25.7720	25.7557	25.7709	25.7832	
100	8	1	$1/\sqrt{p}$	0.004	0.949	0.951	0.9434	0.9454	0.9504
len				0.5595	1.7543	1.7483	1.7535	1.7515	
1000	8	1	$1/\sqrt{p}$	0	0.947	0.9432	0.9482	0.954	0.951
len				0.5328	1.6553	1.6562	1.6559	1.6561	
100	8	2	$1/\sqrt{p}$	0.0056	0.9554	0.9446	0.9514	0.947	0.9518
len				0.5610	3.4389	3.4320	3.4336	3.4416	
1000	8	2	$1/\sqrt{p}$	0	0.952	0.9424	0.9502	0.95	0.9496
len				0.5328	3.2334	3.2345	3.2357	3.2341	

Table 5.5. p=4,gamma=0.1

n	p	k	ψ	α	β_1	β_2	β_{p-1}	β_p	Ftest
100	4	1	0	0	0.1994	0.9462	0.9482	0.9472	0.9434
len				64.8313	68.3369	60.1241	60.1381	60.1601	
1000	4	1	0	0	0	0.9546	0.9496	0.9516	0.9484
len				62.562	65.2185	57.5180	57.5337	57.5694	
100	4	2	0	0	0.4296	0.4308	0.9478	0.9476	0.9514
len				65.2778	91.6704	91.5314	85.5791	85.6785	
1000	4	2	0	0	0	0	0.9468	0.9468	0.9466
len				62.5842	87.0644	87.0277	81.3782	81.3757	
100	4	1	0.9	0	0.9486	0.9524	0.951	0.9518	0.9548
len				65.0397	970.5270	970.0337	969.3072	969.7589	
1000	4	1	0.9	0	0.8908	0.9482	0.9464	0.9494	0.9462
len				62.5195	922.6177	922.4574	923.0181	922.7911	
100	4	2	0.9	0	0.9492	0.9534	0.9494	0.95	0.946
len				65.3116	1941.2176	1946.4222	1952.8835	1943.4188	
1000	4	2	0.9	0	0.9416	0.9384	0.9516	0.9472	0.948
len				62.5349	1845.4996	1844.4384	1843.8134	1843.7230	
100	4	1	$1/\sqrt{p}$	0	0.7058	0.9518	0.9574	0.9478	0.9556
len				65.1461	143.4593	139.498	139.5655	139.6569	
1000	4	1	$1/\sqrt{p}$	0	0.0084	0.9492	0.9502	0.9504	0.9524
len				62.5034	136.1708	132.6081	132.5496	132.5400	
100	4	2	$1/\sqrt{p}$	0	0.8878	0.8856	0.9514	0.9484	0.9486
len				65.0247	270.5358	270.1237	268.1231	268.245	
1000	4	2	$1/\sqrt{p}$	0	0.3484	0.3506	0.9452	0.9446	0.9458
len				62.5512	257.4557	257.6739	255.6722	255.7433	

Table 5.6. p=8,gamma=0.1

n	p	k	ψ	α	β_1	β_2	β_{p-1}	β_p	Ftest
100	8	1	0	0	0.2394	0.9434	0.9498	0.9502	0.947
len				67.6905	71.3756	62.8130	62.8707	62.7660	
1000	8	1	0	0	0	0.9564	0.9552	0.9452	0.9508
len				62.6620	65.3352	57.6591	57.6390	57.6388	
100	8	2	0	0	0.4474	0.439	0.9454	0.953	0.9486
len				67.3270	94.5821	94.5079	88.2303	88.5117	
1000	8	2	0	0	0	0	0.9506	0.947	0.955
len				62.7739	87.1929	87.2982	81.6529	81.6756	
100	8	1	0.9	0	0.9426	0.9504	0.9452	0.9554	0.946
len				67.4179	1509.143	1513.018	1512.034	1510.2	
1000	8	1	0.9	0	0.9298	0.953	0.9492	0.952	0.9534
len				62.7318	1394.644	1394.838	1393.779	1393.703	
100	8	2	0.9	0	0.9498	0.9522	0.949	0.951	0.952
len				67.6638	3028.196	3029.977	3036.264	3031.546	
1000	8	2	0.9	0	0.946	0.9416	0.948	0.9508	0.9472
len				62.6508	2783.915	2785.493	2784.229	2783.029	
100	8	1	$1/\sqrt{p}$	0	0.8314	0.951	0.956	0.9506	0.9488
len				67.6548	197.5671	195.1421	194.6639	195.0724	
1000	8	1	$1/\sqrt{p}$	0	0.0834	0.9488	0.9514	0.9578	0.9468
len				62.7311	181.8244	179.0762	178.9908	179.0377	
100	8	2	$1/\sqrt{p}$	0	0.9092	0.9156	0.943	0.947	0.9398
len				67.289	379.5497	380.088	379.076	378.743	
1000	8	2	$1/\sqrt{p}$	0	0.5844	0.5918	0.9518	0.9474	0.947
len				62.7317	351.5133	351.7104	350.3533	350.379	

Table 5.7. p=4,gamma=0.2

n	p	k	ψ	α	β_1	β_2	β_{p-1}	β_p	Ftest
100	4	1	0	0	0.2874	0.9476	0.9502	0.9504	0.947
len				31.3747	33.4087	29.6287	29.6546	29.6216	
1000	4	1	0	0	0	0.951	0.9468	0.9482	0.9484
len				30.2424	31.8861	28.2881	28.2924	28.3152	
100	4	2	0	0	0.509	0.5038	0.951	0.9522	0.9514
len				31.4606	44.6611	44.8055	42.0355	41.9561	
1000	4	2	0	0	0	0	0.9502	0.9526	0.9558
len				30.2224	42.5778	42.5745	39.9780	39.9926	
100	4	1	0.9	0	0.9468	0.9518	0.952	0.952	0.9526
len				31.5959	478.8191	478.2536	476.5612	478.4111	
1000	4	1	0.9	0	0.9132	0.9498	0.9458	0.9514	0.9536
len				30.2521	454.3176	454.0204	453.9011	454.1279	
100	4	2	0.9	0	0.952	0.9394	0.9512	0.9598	0.956
len				31.4480	950.1755	950.2043	952.0222	950.7384	
1000	4	2	0.9	0	0.945	0.9356	0.9452	0.95	0.9488
len				30.2392	907.3382	907.0891	907.3986	906.9019	
100	4	1	$1/\sqrt{p}$	0	0.7752	0.9496	0.9458	0.9522	0.9468
len				31.5295	70.4546	68.6984	68.6576	68.7149	
1000	4	1	$1/\sqrt{p}$	0	0.025	0.9458	0.956	0.946	0.9488
len				30.2576	66.9508	65.2555	65.3091	65.2618	
100	4	2	$1/\sqrt{p}$	0	0.8958	0.8954	0.9502	0.9438	0.9422
len				31.4559	132.9280	132.9377	131.9290	131.7833	
1000	4	2	$1/\sqrt{p}$	0	0.4302	0.4168	0.9468	0.951	0.9458
len				30.2206	126.4858	126.5364	125.6751	125.6365	

Table 5.8. p=8,gamma=0.2

n	p	k	ψ	α	β_1	β_2	β_{p-1}	β_p	Ftest
100	8	1	0	0	0.295	0.9526	0.9496	0.9534	0.953
len				32.5228	34.5777	30.7187	30.8117	30.6650	
1000	8	1	0	0	0	0.948	0.9494	0.9542	0.9512
len				30.3193	31.9731	28.3617	28.3891	28.3765	
100	8	2	0	0	0.5134	0.5384	0.9444	0.9516	0.9574
len				32.4324	46.1332	46.1596	43.3077	43.2915	
1000	8	2	0	0	0	0	0.95	0.9474	0.9442
len				30.3122	42.7498	42.7055	40.1117	40.0951	
100	8	1	0.9	0	0.9488	0.9508	0.9412	0.945	0.9492
len				32.5756	742.1757	741.4085	743.0718	744.468	
1000	8	1	0.9	0	0.9274	0.9486	0.9528	0.9464	0.9486
len				30.314	684.7783	684.8700	684.8004	685.1093	
100	8	2	0.9	0	0.9472	0.9522	0.9476	0.95	0.9532
len				32.4979	1483.61	1481.41	1480.275	1484.24	
1000	8	2	0.9	0	0.9384	0.944	0.9468	0.9476	0.9496
len				30.3301	1370.817	1370.123	1370.954	1369.419	
100	8	1	$1/\sqrt{p}$	0	0.8528	0.9448	0.9498	0.9504	0.9492
len				32.5601	96.7917	95.4206	95.3259	95.3800	
1000	8	1	$1/\sqrt{p}$	0	0.1462	0.9536	0.9494	0.9494	0.952
len				30.3362	89.3526	88.0950	88.0602	88.1289	
100	8	2	$1/\sqrt{p}$	0	0.9312	0.9282	0.9514	0.9504	0.948
len				32.4215	186.0931	186.4105	185.5374	185.3135	
1000	8	2	$1/\sqrt{p}$	0	0.6424	0.6538	0.9484	0.9502	0.9462
len				30.3014	172.6545	172.632	172.0048	171.9607	

Table 5.9. p=4,gamma=1

n	p	k	ψ	α	β_1	β_2	β_{p-1}	β_p	Ftest
100	4	1	0	0.0002	0.8896	0.9506	0.9538	0.9482	0.9502
len				5.5959	5.8426	5.6052	5.6142	5.6134	
1000	4	1	0	0	0.3322	0.9486	0.9564	0.9498	0.9506
len				5.4293	5.6327	5.3976	5.4001	5.3977	
100	4	2	0	0	0.9206	0.9112	0.9522	0.9436	0.9466
len				5.5739	8.0746	8.0663	7.8930	7.8899	
1000	4	2	0	0	0.5932	0.6034	0.95	0.9498	0.9492
len				5.4248	7.7978	7.7920	7.6322	7.6319	
100	4	1	0.9	0	0.9476	0.9522	0.9446	0.9514	0.9496
len				5.5780	89.7420	89.8072	89.7428	89.6237	
1000	4	1	0.9	0	0.9472	0.9508	0.95	0.9468	0.9516
len				5.4308	86.6672	86.6599	86.6317	86.6336	
100	4	2	0.9	0.0004	0.952	0.9552	0.9502	0.9462	0.9512
len				5.6161	180.5797	180.6311	180.2662	180.7399	
1000	4	2	0.9	0	0.954	0.9514	0.9478	0.9544	0.9538
len				5.4254	172.8756	173.0179	172.8876	172.9414	
100	4	1	$1/\sqrt{p}$	0	0.9368	0.9536	0.946	0.9458	0.9446
len				5.5869	13.0553	12.9426	12.9112	12.9560	
1000	4	1	$1/\sqrt{p}$	0	0.8028	0.954	0.955	0.9508	0.9472
len				5.4316	12.5569	12.4568	12.4529	12.4624	
100	4	2	$1/\sqrt{p}$	0.0004	0.951	0.9512	0.9494	0.9526	0.9472
len				5.5957	24.9579	25.0127	24.9203	24.9523	
1000	4	2	$1/\sqrt{p}$	0	0.9074	0.9132	0.9436	0.946	0.939
len				5.4247	24.0332	24.0352	23.9714	23.9902	

Table 5.10. p=8,gamma=1

n	p	k	ψ	α	β_1	β_2	β_{p-1}	β_p	Ftest
100	8	1	0	0.0004	0.885	0.9484	0.9456	0.953	0.949
len				5.7289	5.9786	5.7300	5.7275	5.7328	
1000	8	1	0	0	0.3308	0.9498	0.9492	0.9476	0.9528
len				5.4433	5.6435	5.4131	5.4132	5.4103	
100	8	2	0	0.0002	0.927	0.9218	0.9482	0.9462	0.949
len				5.7373	8.3083	8.3026	8.1173	8.1315	
1000	8	2	0	0	0.5886	0.5902	0.9532	0.9488	0.9496
len				5.4382	7.8157	7.8120	7.6439	7.6471	
100	8	1	0.9	0.0006	0.9504	0.9498	0.9504	0.9464	0.9512
len				5.7230	138.5736	138.6580	138.7049	138.4863	
1000	8	1	0.9	0	0.9496	0.9492	0.9494	0.947	0.9518
len				5.4371	130.6414	130.5683	130.5454	130.6157	
100	8	2	0.9	0.0006	0.9482	0.9482	0.9486	0.9462	0.9474
len				5.7213	276.8517	276.5351	277.0885	276.9547	
1000	8	2	0.9	0	0.9524	0.9446	0.9508	0.9506	0.953
len				5.4424	261.3473	261.3177	261.3832	261.4592	
100	8	1	$1/\sqrt{p}$	0.0002	0.943	0.946	0.9504	0.954	0.9514
len				5.7534	17.9925	17.9235	17.8525	17.9150	
1000	8	1	$1/\sqrt{p}$	0	0.8718	0.9546	0.9502	0.9444	0.9492
len				5.4369	16.8578	16.7918	16.7847	16.7889	
100	8	2	$1/\sqrt{p}$	0.0002	0.9484	0.9568	0.9494	0.9492	0.9536
len				5.7323	34.8391	34.9193	34.8409	34.8250	
1000	8	2	$1/\sqrt{p}$	0	0.9252	0.9312	0.9448	0.9518	0.9496
len				5.4380	32.8580	32.8374	32.8165	32.8213	

CHAPTER 6

SURVIVAL MODELS

The Kaplan-Meier method is a nonparametric statistical method for estimating survival function from survival data. By the Kaplan-Meier method, the survival function can be represented as:

$$\hat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i}) \tag{6.1}$$

where t_i is the time when the event occurs, d_i is the number of events occurring at time t_i, and n_i is the number of survived individuals at time t_i. We use the Kaplan-Meier graph to display the survival probability with time variation, namely the approximate survival function curve.

The Cox proportional hazard model, also known as the Cox model, is a semi-parametric statistical method, use to estimate the impact of covariates on time when there are covariates besides event and time in the survival data. In addition, the Cox model can also be applied to predict the survival probability at a certain time point. The model is based on two hypotheses:

- Proportional Hazard hypothesis: the hazard rate is in proportion to the covariates, which means the influence of covariates on hazard does not change with time
- Logarithmic Linear hypothesis: the covariates have a linear relation with the logarithmic hazard ratio.

The Cox proportional hazards regression (PH) model is

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{_{Y_i|\boldsymbol{\beta}'\boldsymbol{x}_i}}(t) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)h_0(t)$$

where $h_0(t)$ is the unknown baseline function and $\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)$ is the hazard ratio.

CHAPTER 7

REAL DATA ANALYSIS

The 'Survival Dataset'which is used for this analysis, consists of some information on the survival in a particular organization for 701 employees. The dataset has 701 observations and 9 variables. Description of the variables are as below. (https://github.com/davidcaughlin/R-Tutorial-Data-Files/blob/master/Survival.csv)

- ID- Employee Identification Number
- Start date – Start date in the organization
- Gender - Gender of the employee (Man, Woman)
- Race – Race of the employee (Black, Hispanic Latino, White)
- Pay hourly – Last hourly payment they received in Dollars
- Pay sat – Pay satisfaction level in the survey (1 to 5 scale)
- Turnover – 0= Still in organization, 1=Voluntary turnover, 2= Involuntary turnover
- Turnover date – Turnover date in the organization
- LOS – How long someone work in the organization in days

Inspect Length of Service (LOS) Variable distribution

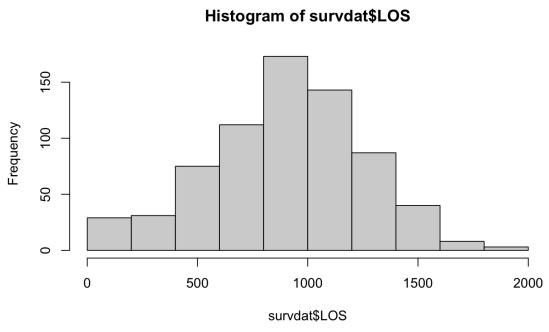

Figure 7.1. Histogram of Length of Service

'Length of Service' seems to be normally distributed.

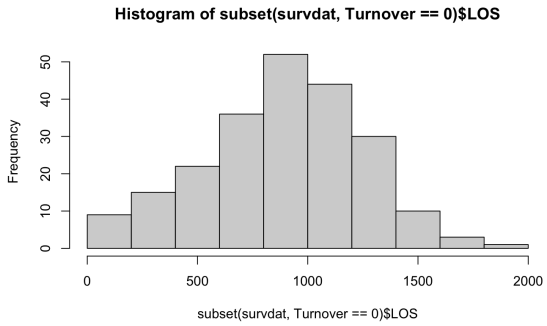

Figure 7.2. Histogram of Length of Service

The distribution for the length of service for those people who still stay in the organization seems to be normally distributed.

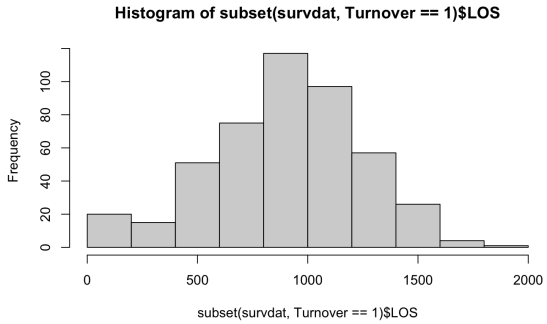

Figure 7.3. Histogram of Length of Service

The distribution for the length of service for those people those people who voluntarily turned over seems to be normally distributed.

Kaplan-Meier Analysis

Censoring Variable.

Voluntary turnover is considered as the focal event and anything other than voluntary turnover (if someone left the organization or they're still in the organization) to be considered as right censored. Threfore, new variable created as 'censored'.

KM-Model 1

```
Call: survfit(formula = Surv(LOS, censored) ~ 1,
data = survdat, type = "kaplan-meier")

      n events median 0.95LCL 0.95UCL
[1,] 701    463   1059    1022    1095
```

Among 701 total observations, 463 experience the focal event which is voluntary turnover. The median length of service before someone experienced the event of voluntary turnover is 1059 days. The 95% confidence interval of the lower limit is 1022 days, and the upper 95% confidence interval limit is 1095 days. Median is a point estimate between these two limits here for the 95% confidence interval.

```
Call: survfit(formula = Surv(LOS, censored) ~ 1,
data = survdat, type = "kaplan-meier")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   73    701      12   0.9829 0.00490       0.9733       0.9925
  110    682       1   0.9814 0.00510       0.9715       0.9915
  146    681       3   0.9771 0.00566       0.9661       0.9883
  183    677       4   0.9713 0.00632       0.9590       0.9838
```

According to the life table, 73 days after they started the organization, 12 people experience the event of voluntary turnover. Which means when their length of service was 73 days, 12 people voluntarily turned over. After 110 days one person experienced the event again and so on. The 'n.risk' column indicates the average number of individuals at risk for the relevant time interval.

The survival column contains the proportion of individuals who survived past or through the time interval in question and it's referred to as the cumulative survival rate. For example, the survival rate is 0.9814 for 73 days to 110 days. It implies that 98.14% of the people have yet to experience the event in question.

Plot cumulative survival rates

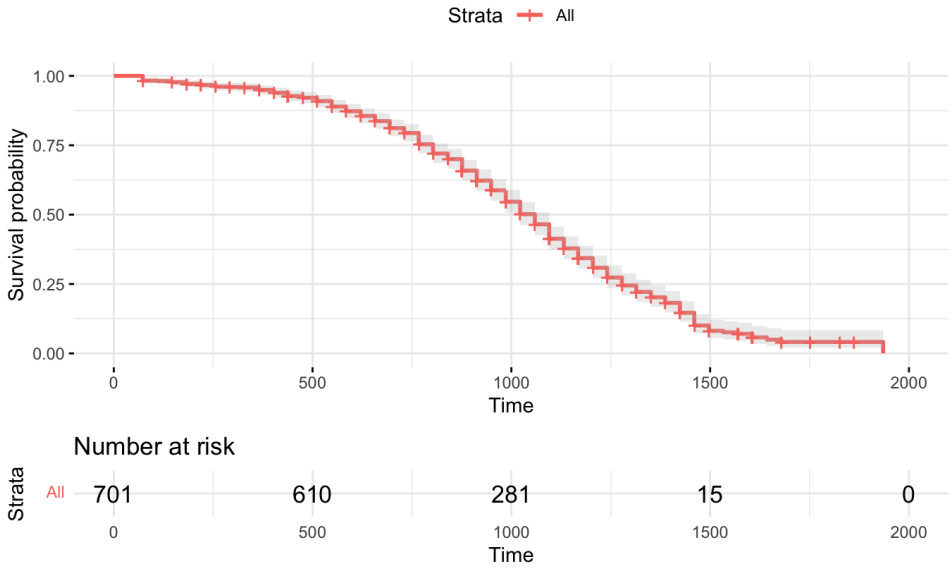

Figure 7.4. Cumulative survival rates

The plot shows cumulative survival rates over time. The x-axis represents the number of days for the time to event. The y-axis represents the cumulative survival probability. Confidence intervals are shaded in the plot. In addition to the graph, a table shows some

more information. 701 people are at risk at the point of zero days. When it comes to 500 days, 610 people are at risk and so on. When it gets to 2000 days everybody has experienced the event voluntary turnover.

Kaplan-Meier Analysis-Models with categorical covariates

KM-Model 2

```
Call: survfit(formula = Surv(LOS, censored) ~ Race, data = survdat,
    type = "kaplan-meier")
```

	n	events	median	0.95LCL	0.95UCL
Race=Black	283	190	1022	986	1095
Race=HispanicLatino	79	57	1022	876	1132
Race=White	339	216	1059	1022	1095

Now the Kaplan Meier model is fitted using the 'Race' variable as a covariate. Out of 283 black individuals, 190 experienced the event of voluntary turnover by the end. 57 Hispanic Latino individuals out of 79 experienced the event of voluntary turnover. 216 people who have identified as White in terms of their race experienced voluntary turnover out of 339. The median length of service for the three groups are 1022, 1022 and 1059 respectively.

Life tables for the three independent race groups are as shown below.

```
Call: survfit(formula = Surv(LOS, censored) ~ Race, data = survdat,
    type = "kaplan-meier")

               Race=Black
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   73    283       6   0.9788 0.00856       0.9622       0.9957
  146    272       3   0.9680 0.01049       0.9477       0.9888
  183    268       3   0.9572 0.01210       0.9337       0.9812
  256    264       1   0.9535 0.01258       0.9292       0.9785

            Race=HispanicLatino
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   73     79       2   0.9747  0.0177      0.94065        1.000
  256     77       1   0.9620  0.0215      0.92079        1.000
  438     74       3   0.9230  0.0302      0.86569        0.984
  475     71       1   0.9100  0.0325      0.84859        0.976

               Race=White
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   73    339       4   0.9882 0.00586       0.9768        1.000
  110    333       1   0.9852 0.00656       0.9725        0.998
  183    332       1   0.9823 0.00718       0.9683        0.996
  219    330       3   0.9733 0.00877       0.9563        0.991
```

Plot cumulative survival rate curves for categories of Race variable and Log rank test

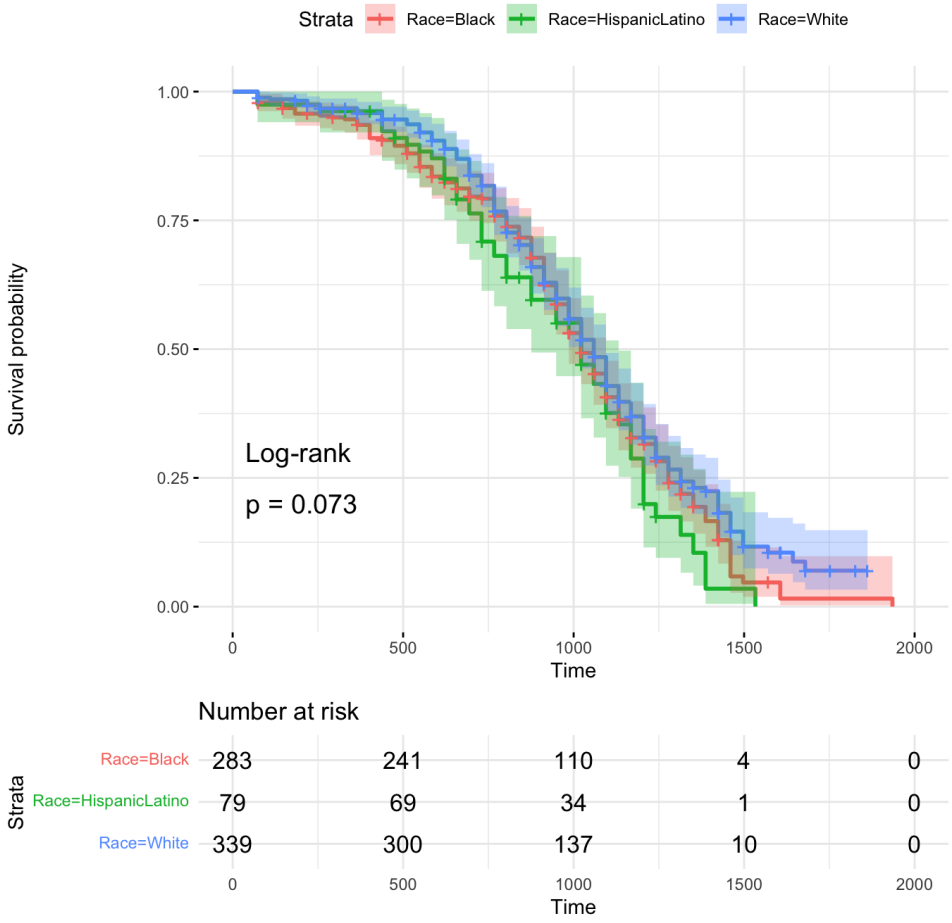

Figure 7.5. Cumulative survival rates for categories of Race

The plot indicates three different stratums for three different race groups. In other words, black individuals are represented by red color while the Hispanic Latino individuals and the white individuals show in green and blue respectively. It looks like there are some differences between three cumulative survival rate curves represented here.

Using the log rank test, we can see whether those cumulative survival rate curves are significantly different between the different race groups. The p-value is .073 which is greater

than our conventional .05 alpha level. It implies that these cumulative survival rates don't seem to differ between the different race groups.

According to the risk table, which is now separated by the three different race groups, 283, 79 and 339 individuals from Black, Hispanic Latino and White respectively are at risk at the point of zero days. When it comes to 500 days, 241, 69, 300 people are at risk and so on. When it gets to 2000 days, eventually everybody has experienced the event voluntary turnover.

KM-Model 3

```
Call: survfit(formula = Surv(LOS, censored) ~ Gender, data = survdat,
    type = "kaplan-meier")

                n events median 0.95LCL 0.95UCL
Gender=Man    340    213   1022     986    1095
Gender=Woman 361    250   1059     986    1095
```

Another Kaplan Meier model was fitted using the 'Gender' variable as a covariate. Out of 340 male individuals, 213 experienced the event of voluntary turnover by the end. 250 female individuals out of 361 experienced the event of voluntary turnover. The median length of service for the males and females are 1022 and 1059 respectively.

Life tables for the two independent gender groups are as shown below.

```
Call: survfit(formula = Surv(LOS, censored) ~ Gender, data = survdat,
    type = "kaplan-meier")

                Gender=Man
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   73    340       3   0.9912 0.00507      0.98129        1.000
  146    333       1   0.9882 0.00587      0.97677        1.000
  183    332       2   0.9822 0.00718      0.96827        0.996
  219    330       2   0.9763 0.00828      0.96020        0.993


                Gender=Woman
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   73    361       9   0.9751 0.00821       0.9591        0.991
  110    349       1   0.9723 0.00865       0.9555        0.989
  146    348       2   0.9667 0.00946       0.9483        0.985
  183    345       2   0.9611 0.01020       0.9413        0.981
```

Cumulative survival rate curves for categories of Gender variable and Log rank test

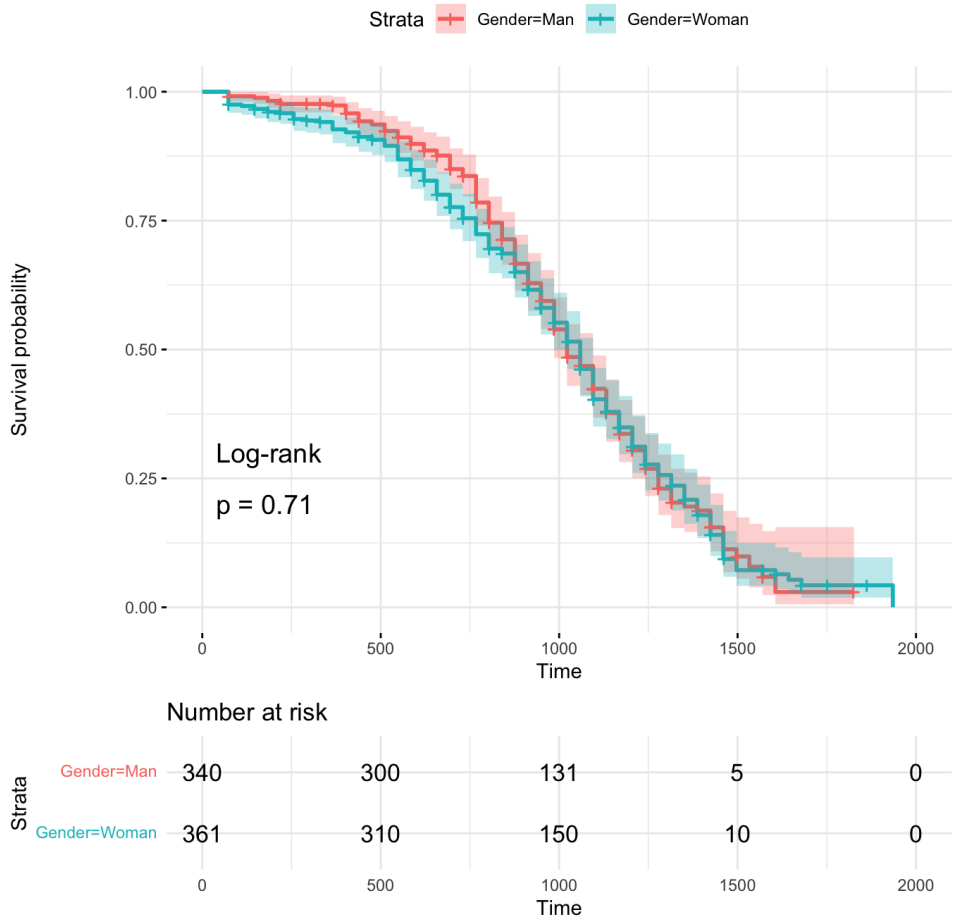

Figure 7.6. Cumulative survival rates for categories of Gender

The plot indicates two different stratums for two different gender groups. In other words, male individuals are represented by red color while female individuals are shown in green. It looks like there are some differences between the two cumulative survival rate curves.

Using the log rank test, we can see whether those cumulative survival rate curves are significantly different between the different gender groups. The p-value is .071 which is greater than our conventional .05 alpha level, implying that these cumulative survival rates

don't seem to differ between the different gender groups.

According to the risk table, which is now separated by the two different gender groups, 340 and 361 individuals from males and females respectively are at risk at that point of zero days. When it comes to 500 days, 300 and 310 people are at risk and so on. When it gets in to 2000 days eventually everybody has experienced the event voluntary turnover.

Estimate Cox Proportional Hazards (PH) Model- Cox Regression Model

Cox Regression Model 1

```
Call:
coxph(formula = Surv(LOS, censored) ~ Race, data = survdat)

  n= 701, number of events= 463

                      coef exp(coef) se(coef)      z Pr(>|z|)
RaceHispanicLatino  0.17936   1.19645  0.15147  1.184    0.236
RaceWhite          -0.14365   0.86619  0.09998 -1.437    0.151

                   exp(coef) exp(-coef) lower .95 upper .95
RaceHispanicLatino    1.1964     0.8358    0.8891     1.610
RaceWhite             0.8662     1.1545    0.7120     1.054

Concordance= 0.522  (se = 0.014 )
Likelihood ratio test= 5.13  on 2 df,   p=0.08
Wald test            = 5.28  on 2 df,   p=0.07
Score (logrank) test = 5.3  on 2 df,   p=0.07
```

This model compares the cumulative survival rates across the three different race categories using the 'Black' race category as the reference group. As given in the Kaplan-Meier analysis, 463 people experienced voluntary turnover which is the event in question out of 701 sample size.

The concordance is an aggregate estimate of how well the model predicts individual people's experience of the event. It represents the proportion of pairs of individuals who experience of the event in question. A concordance value of 0.522 indicates that the model really does no better than chance. In other words, there's a 50-50 chance of correctly predicting which individuals incur the event before the other. The race covariate and the model slightly improve the accuracy of our predictions but not by too much.

We have the likelihood ratio test, Wald test and the log rank test information. These tests are asymptotically equivalent to one another when assuming we have a large enough sample size. When we have smaller sample sizes it is better to use the log rank test. The log rank test gives the same p-value we obtained from Kaplan-Meier analysis implying that cumulative survival rates don't seem to differ between the different race groups.

Cox Regression Model 2

```
Call:
coxph(formula = Surv(LOS, censored) ~ Race + Pay_hourly + Pay_sat,
    data = survdat)

  n= 663, number of events= 445
   (38 observations deleted due to missingness)

                      coef exp(coef) se(coef)      z Pr(>|z|)
RaceHispanicLatino  0.12087   1.12848  0.15796  0.765 0.444154
RaceWhite          -0.04446   0.95651  0.10367 -0.429 0.668008
Pay_hourly         -0.10971   0.89609  0.03276 -3.349 0.000812 ***
Pay_sat             0.14135   1.15183  0.08135  1.738 0.082289 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                   exp(coef) exp(-coef) lower .95 upper .95
RaceHispanicLatino    1.1285     0.8862    0.8280    1.5380
RaceWhite             0.9565     1.0455    0.7806    1.1720
Pay_hourly            0.8961     1.1160    0.8404    0.9555
Pay_sat               1.1518     0.8682    0.9821    1.3509

Concordance= 0.55  (se = 0.016 )
Likelihood ratio test= 15.97  on 4 df,   p=0.003
Wald test            = 16  on 4 df,   p=0.003
Score (logrank) test = 16.02  on 4 df,   p=0.003
```

This model includes two additional continuous covariates called ‘hourly pay rate for people’ as well as their ‘pay satisfaction ratings’ together with the categorical covariate ’race’. Now the sample size is 663 instead of 701 due to the missingness of 38 observations in ‘pay satisfaction variable’. Out of 663, 445 people experienced voluntary turnover.

The ‘Hourly pay rate’ variable is statistically significant because the p-value is less than the conventional alpha value of 0.05 meaning that variable is statistically associated with the risk for experiencing voluntary turnover and it is a negative association. This implies that the higher hourly payments have the less risk for experiencing voluntary turnover within the time period of the study in question. All the other three variables are not statistically significant.

Also, 0.8961 is the hazard ratio value for Hourly pay rate variable, and 1-0.8961 = 0.1039, which means the risk of an individual experiencing the focal event is going to decrease by about 10.4% for every additional dollar an individual earned.

The concordance value is now 0.55 which has increased by 0.028. Further, the log rank test is statistically significant since the p-value is 0.003 (less than 0.05) which indicates that this model fit the data significantly better than a null model with no covariates in the model.

Estimating Overall risk for a specific individual using the Cox Regression Model 2

Let's consider a hypothetical person who is a Hispanic Latino with hourly pay rate 16 and pay satisfaction level is 4 on the 1 to 5 pay satisfaction scale.

```
> RaceHL = 1
> RaceW = 0
> PH = 16
> PS= 4.00
>
> Log_overallrisk = .121*RaceHL -0.044*RaceW -0.110*PH + 0.141*PS
> print(Log_overallrisk)
[1] -1.075
> exp(Log_overallrisk)
[1] 0.3412978
```

The log overall risk is -1.075. We can get the overall risk by the exponentiated value which is 0.341. This means that the overall risk of voluntarily turning over is 0.341 for this individual who is a Hispanic Latino with hourly pay rate 16 and pay satisfaction level is 4 on the 1 to 5 pay satisfaction scale. Moreover, 1-0.3413 = 0.6587, implying that this individual is 65.9% less likely to quit when compared to an individual with scores of zero on each of the covariates in the model.

An individual with scores of zero on each of the covariates in the model is someone who is black, earning zero dollars an hour which is probably not legal and have a score of zero on the pay satisfaction scale, but the pay satisfaction scale can range from one to five. It's not a meaningful value. To overcome this problem, we can use some techniques and fit the Cox regression model 3.

Cox Regression Model 3

```
coxph(formula = Surv(LOS, censored) ~ HL_Race + c_pay_hourly +
    c_pay_sat, data = survdat)

  n= 663, number of events= 445
   (38 observations deleted due to missingness)

                 coef exp(coef) se(coef)      z Pr(>|z|)
HL_RaceBlack -0.12087   0.88615  0.15796 -0.765 0.444154
HL_RaceWhite -0.16533   0.84761  0.15773 -1.048 0.294567
c_pay_hourly -0.10971   0.89609  0.03276 -3.349 0.000812 ***
c_pay_sat     0.14135   1.15183  0.08135  1.738 0.082289 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
HL_RaceBlack    0.8862     1.1285    0.6502    1.2077
HL_RaceWhite    0.8476     1.1798    0.6222    1.1547
c_pay_hourly    0.8961     1.1160    0.8404    0.9555
c_pay_sat       1.1518     0.8682    0.9821    1.3509

Concordance= 0.55  (se = 0.016 )
Likelihood ratio test= 15.97  on 4 df,   p=0.003
Wald test            = 16  on 4 df,   p=0.003
Score (logrank) test = 16.02  on 4 df,   p=0.003
```

Using the grand mean center of the continuous variables (for example: a person's hourly pay of 13.23 and we're subtracting the overall sample mean, the resulting value is going to be the grand mean centered value), instead of comparing to someone with making zero pay which would probably be illegal, we can compare them to someone who's making average pay. Similarly with the pay satisfaction variable instead of comparing them to someone who has a completely theoretically meaningless value (zero) on the scale of one to five, we can use average score based on this sample.

Moreover, in the Cox regression model 3, the reference group of race variable has changed to Hispanic Latino which was Black earlier.

The 'Hourly pay rate' variable is still statistically significant because the p-value is less than the conventional alpha value of 0.05. All the other three variables are again not statistically significant.

0.8961 is again the same hazard ratio value for Hourly pay rate variable. 1-0.8961 = 0.1039, which means the risk of an individual experiencing the focal event is going to decrease by about 10.4% for every additional dollar an individual earned.

Estimating Overall risk for a specific individual using the Cox Regression Model 3

Let's consider the same hypothetical person who is a Hispanic Latino with hourly pay rate 16 and pay satisfaction level is 4 on the 1 to 5 pay satisfaction scale. (Hourly pay rate variable and pay satisfaction variable are grand mean centered here.)

```
> RaceB = 0
> RaceW = 0
> PH = 16.00 - mean(survdat$Pay_hourly,na.rm = TRUE)
> PS= 4.00 - mean(survdat$Pay_sat,na.rm = TRUE)
>
> Log_overallrisk = -.121*RaceB -0.165*RaceW -0.110*PH + 0.141*PS
> print(Log_overallrisk)
[1] -0.1623714
> exp(Log_overallrisk)
[1] 0.8501254
```

The log overall risk is -0.1624. We can get the overall risk by exponentiating that value which is 0.850. This means that the overall risk of voluntarily turning over is 0.850 for this individual who is a Hispanic Latino with hourly pay rate 16 and pay satisfaction level is 4 on the 1 to 5 pay satisfaction scale. Moreover, 1-0.85 = 0.15, implying that this individual is 15% less likely to quit when compared to an individual with scores of zero on each of the covariates in the model.

In other word, an individual with scores of zero on each of the covariates in the model is someone who is Hispanic Latino, earning average hourly pay and have average pay satisfaction. Now, this value is more meaningful.

Nested model comparison of Cox regression model 1 and Cox regression model 2

The Cox regression model 1 had only one covariate, 'race' with 701 people in the sample. But in the Cox regression model 2, we found out that were 38 people had missing data for the 'pay satisfaction ratings' variable. Therefore, the resulting sample was 663 people for that model. In order to compare the two models, we need to make sure, we're comparing models with the same sample sizes. So, we must eliminate those 38 observations from Cox regression model 1.

```
 Call:
coxph(formula = Surv(LOS, censored) ~ Race, data = drop_na(survdat,
    LOS, censored, Race, Pay_hourly, Pay_sat))

  n= 663, number of events= 445

                     coef exp(coef) se(coef)      z Pr(>|z|)
RaceHispanicLatino  0.1487    1.1604   0.1576  0.944    0.345
RaceWhite          -0.1029    0.9022   0.1016 -1.013    0.311

                   exp(coef) exp(-coef) lower .95 upper .95
RaceHispanicLatino    1.1604     0.8618    0.8520     1.580
RaceWhite             0.9022     1.1084    0.7393     1.101

Concordance= 0.514  (se = 0.014 )
Likelihood ratio test= 2.83  on 2 df,   p=0.2
Wald test            = 2.9  on 2 df,   p=0.2
Score (logrank) test = 2.91  on 2 df,   p=0.2
```

```
Call:
coxph(formula = Surv(LOS, censored) ~ Race + Pay_hourly + Pay_sat,
    data = drop_na(survdat, LOS, censored, Race, Pay_hourly,
        Pay_sat))

  n= 663, number of events= 445

                      coef exp(coef) se(coef)      z Pr(>|z|)
RaceHispanicLatino  0.12087   1.12848  0.15796  0.765 0.444154
RaceWhite          -0.04446   0.95651  0.10367 -0.429 0.668008
Pay_hourly         -0.10971   0.89609  0.03276 -3.349 0.000812 ***
Pay_sat             0.14135   1.15183  0.08135  1.738 0.082289 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                   exp(coef) exp(-coef) lower .95 upper .95
RaceHispanicLatino    1.1285     0.8862    0.8280    1.5380
RaceWhite             0.9565     1.0455    0.7806    1.1720
Pay_hourly            0.8961     1.1160    0.8404    0.9555
Pay_sat               1.1518     0.8682    0.9821    1.3509

Concordance= 0.55  (se = 0.016 )
Likelihood ratio test= 15.97  on 4 df,   p=0.003
Wald test            = 16  on 4 df,   p=0.003
Score (logrank) test = 16.02  on 4 df,   p=0.003
```

```
> anova(cox_reg1,cox_reg2)
Analysis of Deviance Table
 Cox model: response is  Surv(LOS, censored)
 Model 1: ~ Race
 Model 2: ~ Race + Pay_hourly + Pay_sat
   loglik  Chisq Df P(>|Chi|)
1 -2459.6
2 -2453.1 13.141  2  0.001401 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to the Anova, the chi-square test associated p-value is less than 0.05. That indicates that the full model with all the covariates included (race, hourly pay rate, pay satisfaction ratings) fits the data significantly better than the smaller nested model that only had the race variable. In other words, the Cox regression model 2 is significantly better than the Cox regression model 1.

Parametric Models

Exponential model 1

```
Call:
survreg(formula = Surv(LOS, censored) ~ Race + Pay_hourly + Pay_sat,
    data = survdat, dist = "exponential")
                     Value Std. Error     z      p
(Intercept)         6.2143     0.5261 11.81 <2e-16
RaceHispanicLatino -0.0326     0.1574 -0.21 0.8358
RaceWhite           0.0183     0.1028  0.18 0.8589
Pay_hourly          0.0923     0.0323  2.86 0.0042
Pay_sat            -0.0847     0.0813 -1.04 0.2978

Scale fixed at 1

Exponential distribution
Loglik(model)= -3645.1   Loglik(intercept only)= -3650
Chisq= 9.83 on 4 degrees of freedom, p= 0.043
Number of Newton-Raphson Iterations: 4
n=663 (38 observations deleted due to missingness)
```

Exponential model 2

```
Call:
survreg(formula = Surv(LOS, censored) ~ Gender + Pay_hourly +
    Pay_sat, data = survdat, dist = "exponential")
              Value Std. Error     z      p
(Intercept)  6.2788     0.5353 11.73 <2e-16
GenderWoman -0.0705     0.0959 -0.74 0.4623
Pay_hourly   0.0910     0.0319  2.85 0.0043
Pay_sat     -0.0858     0.0813 -1.06 0.2910

Scale fixed at 1

Exponential distribution
Loglik(model)= -3644.9   Loglik(intercept only)= -3650
Chisq= 10.26 on 3 degrees of freedom, p= 0.016
Number of Newton-Raphson Iterations: 4
n=663 (38 observations deleted due to missingness)
```

Weibull model 1

```
Call:
survreg(formula = Surv(LOS, censored) ~ Race + Pay_hourly + Pay_sat,
    data = survdat, dist = "weibull")
                     Value Std. Error      z      p
(Intercept)         6.7159     0.1826  36.78 <2e-16
RaceHispanicLatino -0.0363     0.0554  -0.66 0.5118
RaceWhite           0.0153     0.0362   0.42 0.6731
Pay_hourly          0.0366     0.0116   3.16 0.0016
Pay_sat            -0.0479     0.0285  -1.68 0.0934
Log(scale)         -1.0453     0.0387 -27.02 <2e-16

Scale= 0.352

Weibull distribution
Loglik(model)= -3412   Loglik(intercept only)= -3419.2
Chisq= 14.4 on 4 degrees of freedom, p= 0.0061
Number of Newton-Raphson Iterations: 7
n=663 (38 observations deleted due to missingness)
```

Weibull model 2

```
Call:
survreg(formula = Surv(LOS, censored) ~ Gender + Pay_hourly +
    Pay_sat, data = survdat, dist = "weibull")
               Value Std. Error      z       p
(Intercept)  6.6882     0.1864  35.88 < 2e-16
GenderWoman  0.0076     0.0340   0.22 0.82300
Pay_hourly   0.0389     0.0115   3.38 0.00071
Pay_sat     -0.0492     0.0285  -1.73 0.08419
Log(scale)  -1.0443     0.0387 -26.95 < 2e-16

Scale= 0.352

Weibull distribution
Loglik(model)= -3412.4   Loglik(intercept only)= -3419.2
Chisq= 13.58 on 3 degrees of freedom, p= 0.0035
Number of Newton-Raphson Iterations: 7
n=663 (38 observations deleted due to missingness)
```

Log-Logistic model 1

```
Call:
survreg(formula = Surv(LOS, censored) ~ Race + Pay_hourly + Pay_sat,
    data = survdat, dist = "loglogistic")
                    Value Std. Error      z      p
(Intercept)         6.5435     0.2236  29.27 <2e-16
RaceHispanicLatino -0.0233     0.0668  -0.35 0.7269
RaceWhite           0.0109     0.0433   0.25 0.8015
Pay_hourly          0.0360     0.0135   2.67 0.0077
Pay_sat            -0.0340     0.0340  -1.00 0.3167
Log(scale)         -1.2610     0.0404 -31.19 <2e-16

Scale= 0.283

Log logistic distribution
Loglik(model)= -3453.5   Loglik(intercept only)= -3458.1
Chisq= 9.12 on 4 degrees of freedom, p= 0.058
Number of Newton-Raphson Iterations: 4
n=663 (38 observations deleted due to missingness)
```

Log-Logistic model 2

```
Call:
survreg(formula = Surv(LOS, censored) ~ Gender + Pay_hourly +
    Pay_sat, data = survdat, dist = "loglogistic")
               Value Std. Error      z      p
(Intercept)  6.5613     0.2270  28.91 <2e-16
GenderWoman -0.0260     0.0402  -0.65  0.518
Pay_hourly   0.0361     0.0134   2.70  0.007
Pay_sat     -0.0347     0.0340  -1.02  0.307
Log(scale)  -1.2604     0.0404 -31.17 <2e-16

Scale= 0.284

Log logistic distribution
Loglik(model)= -3453.5   Loglik(intercept only)= -3458.1
Chisq= 9.26 on 3 degrees of freedom, p= 0.026
Number of Newton-Raphson Iterations: 4
n=663 (38 observations deleted due to missingness)
```

For all the six models, pay hourly rate is an important variable. The Weibull regression model appears to fit better since the chisq pvalue is much smaller than that for the Exponential and loglogisitic regression models.

CHAPTER 8

CONCLUSIONS

The simulations were done in *R*. See R Core Team (2018). Programs are in the collection of functions *slpack.txt*. See (http://parker.ad.siu.edu/Olive/slpack.txt). Table 1 was made with `aftolssim`.

If the cases are iid from some population, we conjecture that the OLS fit to the uncensored cases gives a consistent estimator for $\boldsymbol{\gamma}$ where $\boldsymbol{\gamma}$ is a biased estimator of $\boldsymbol{\beta}$. Then we can test $H_0 : (\beta_{i1}, ..., \beta_{ik})^T = \mathbf{0}$ with $H_0 : (\gamma_{i1}, ..., \gamma_{ik})^T = \mathbf{0}$ using the OLS fit to the uncensored cases. Testing $H_0 : \beta_1 - \beta_2 = 0$ with $H_0 : \gamma_1 - \gamma_2 = 0$ may not be possible due to the bias of the estimator.

When the $\beta_i = 0$, the confidence intervals usually contained 0. The CIs for α usually did not contain 0, which was what was tested, and the tests were good since $\alpha \neq 0$. The tests for the reduced model were also good.

REFERENCES

[1] Olive, D.J. (2022a), *Survival Analysis*, online course notes, see (http://parker.ad.siu.edu/Olive/survbk.htm).

[2] Olive, D.J. (2022b), "OPLS Regression with IID Cases," preprint at (http://parker.ad.siu.edu/Olive/ppregiid.pdf).

[3] R Core Team (2018), "R: a language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

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

[5] https://github.com/davidcaughlin/R-Tutorial-Data-Files/blob/master/Survival.csv

VITA

Graduate School
Southern Illinois University

Sanjuka Johana Lemonge

sanjuka.johanalemonge@siu.edu

University of Kelaniya
Bachelor of Science, (Honours) Statistics, March 2019

Research Paper Title:
Survival Analysis and the OLS AFT

Major Professor: Dr. David J. Olive