BIS 630 HW 6

Joanna Chen

4/16/2020

We will use the data from the AIDS Clinical Trials Group (HW06_ACTG.csv), a double-blind, placebo-controlled trial that compared a drug regimen including indinavir (IDV) to a drug regimen without IDV (tx) in HIV-infected patients. The primary outcome measure was time to AIDS defining event or death. Variables of interest in this homework are:
tx: 1=Treatment includes IDV; 0=Control group (treatment without IDV)
cd4: Baseline CD4 count (cells/milliliter)
age: Age at enrollment (years)
time: Time to AIDS defining event or death
censor: 1=AIDS defining diagnosis or death; 0=Otherwise
Hint: Throughout this question, you are asked to consider the impact of a 50-unit increase in CD4 count. A 1- unit change in cd4scaled = cd4/50 is equivalent to a 50-unit change in cd4. Thus, you should create and use cd4scaled in your models.

```
setwd("~/Downloads/survival HW6")
library(readr)
library(data.table)
library(flexsurv)
```

```
## Loading required package: survival
```

```
library(survival)
dt <- read_csv("HW06_ACTG.csv")
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   time = col_double(),
##   censor = col_double(),
##   tx = col_double(),
##   txgrp = col_double(),
##   sex = col_double(),
##   raceth = col_double(),
##   ivdrug = col_double(),
##   hemophil = col_double(),
##   karnof = col_double(),
##   cd4 = col_double(),
##   priorzdv = col_double(),
##   age = col_double()
## )
```

```
setDT(dt)
dt$cd4scaled = dt$cd4/50
dt$agez = (dt$age - mean(dt$age))/sd(dt$age)
```

1. [25 points] Fit an exponential model containing treatment, CD4 count, and age.

```
fitex = survreg(Surv(time, censor) ~ tx + cd4scaled + agez, data=dt, dist="exponential")
summary(fitex)
```

```
##
## Call:
## survreg(formula = Surv(time, censor) ~ tx + cd4scaled + agez,
##     data = dt, dist = "exponential")
##               Value Std. Error     z       p
## (Intercept)  6.7148     0.1575 42.64 < 2e-16
## tx           0.6803     0.2150  3.16  0.0016
## cd4scaled    0.8269     0.1269  6.52 7.3e-11
## agez        -0.2405     0.0986 -2.44  0.0147
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -817.1   Loglik(intercept only)= -856.6
##  Chisq= 79.08 on 3 degrees of freedom, p= 4.8e-17
## Number of Newton-Raphson Iterations: 7
## n= 1151
```

```
# -2*logLik(fitex)
# AIC(fitex)
```

a. Using the fitted model, estimate the time ratio with a 95% confidence interval comparing treatment to control. Estimate the time ratio with a 95% confidence interval for a 50-unit increase in CD4 count. Estimate the time ratio with a 95% confidence interval for a 1-unit increase in age.

We know that

$$\hat{TR}_k = e^{\hat{\beta}_k} \quad 95\%\text{CITR}_k : \left(e^{\text{lcl}(\hat{\beta}_k)}, e^{\text{ucl}(\hat{\beta}_k)}\right).$$

We know that $ln(\hat{TR}_k) = 0.6803$ for tx and 95% CI $ln(TR) = (0.6803 - 1.96 * 0.2150, 0.6803 + 1.96 * 0.2150) = (0.2589, 1.1017)$. Therefore, 95% CI TR) for tx = (exp(0.2589), exp(1.1017)) = (1.30, 3.01).

Similary, $ln(\hat{TR}_k) = 0.8269$ for cd4scaled and 95% CI $ln(TR) = (0.8269 - 1.96 * 0.1269, 0.8269 + 1.96 * 0.1269) = (0.578176, 1.075624)$. Therefore, 95% CI TR) for cd4scaled = (exp(0.578176), exp(1.075624)) = (1.78, 2.93).

$ln(\hat{TR}_k) = -0.2405$ for age and 95% CI $ln(TR) = (-0.2405 - 1.96 * 0.0986, -0.2405 + 1.96 * 0.0986) = (-0.433756, -0.047244)$. Therefore, 95% CI TR) for cd4scaled = (exp(-0.433756), exp(-0.047244)) = (0.65, 0.95).

Interpret the estimated time ratios.

$\hat{TR}$ for tx = $exp(0.6803) = 1.97447$. Median survival time for treatment group is 1.97 times that of control group, adjusting for CD4 count and age.
$\hat{TR}$ for cd4scaled = $exp(0.8269) = 2.28622$. With every 50-unit increase in CD4 count, meadian survival time of patients increases by 2.29, adjusting for treatment group and age.
$\hat{TR}$ for age = $exp(-0.2405) = 0.7862346$. With a 1-standard deviation increase in age (SD = 8.81 years), meadian survival time decreases 21 % than before, adjusting for treatment group and CD4 count.

We are 95% confident that the time ratio between treatment group and control group in the population of patients in this trail is between 1.30 and 3.01,adjusting for CD4 count and age. We are 95% confident that every 50-unit increase in CD4 count in the population of patients in this trail is between 1.78 and 2.93, adjusting for treatment group and age. We are 95% confident that the age in the population of patients in this trail is between 0.65 and 0.95,adjusting for treatment group and CD4 count.

b. Using the fitted model, estimate the hazard ratio with a 95% confidence interval comparing treatment to control. Estimate the hazard ratio with a 95% confidence interval for a 50-unit increase in CD4 count. Estimate the hazard ratio with a 95% confidence interval for a 1-unit increase in age.

For tx, HR CI = $ (exp(-1.1),exp(-0.2589)) = (0.3328711, 0.7719002)$.
For cd4scaled, HR CI = $(exp(-1.075624), exp(-0.578176)) = (0.3410849.0.5609206)$.
For age, HR CI = $(exp(0.047244), exp(0.433756)) = (1.048378, 1.543042)$.

Interpret the estimated hazard ratios.

$\hat{HR}$ for tx = $exp(-0.6803) = 0.506465$ Hazard in treatment group is 0.506465 times that of control group.
$\hat{HR}$ for cd4scaled = $exp(-0.8269) = 0.44$. With cd4scaled increasing by 1 unit, hazards increases 0.44 many times than before.
$\hat{HR}$ for age = $exp(0.2405) = 1.27$. a 1-standard deviation increase in age (SD = 8.81 years), hazards increases 27.6%.

Notes to myself: with covariate increase by 1 unit, hazards increase x many times than before/increase x-1%.

c. Compare the interpretation of the time and hazard ratios computed in (a), (b).

In (a), (b), for tx, the survival time for treatment group is longer times that of control group, and the hazard ratio for the effect of treatment was 0.506465 which $1/0.506465 = 1.97 > 1$ implies the exposure is beneficial to survival. The two results are consensus.
In (a), (b), for cd4scaled, the increasing cd4scaled is associated with longer survival time, and the hazard ratio for the effect of treatment was 0.44 which $1/0.44 = 2.29 > 1$ implies the exposure of increasing cd4scaled is beneficial to survival. The two results are consensus.
In (a), (b), for age, the increasing age is associated with shorter survival time, and the hazard ratio for the effect of treatment was 1.27. which $1/1.27 = 0.78 < 1$ implies the increasing age is not beneficial to survival. The two results are consensus.

d. Using the fitted model, calculate the median survival time for each treatment group when CD4 count = 86.5 cells/milliliter and age = 38.6 years.

cd4scaled = cd4/50 = 86.5/50 = 1.73 agez = -0.005363878

```
(38.6 - mean(dt$age))/(sd(dt$age))
```

```
## [1] -0.005363878
```

$t_{50}(X) = -\ln(0.5)\exp(\beta_0 + \beta X) = -\ln(0.5)\exp(6.7148 + 0.6803 * tx + 0.8269 * cd4scaled - 0.2405 * agez)$

Treatment group:

```
-log(0.5)*exp(6.7148 + 0.6803*1 + 0.8269*1.73-0.2405*(-0.005363878))
```

```
## [1] 4724.014
```

Control group:

```
-log(0.5)*exp(6.7148 + 0.6803*0 + 0.8269*1.73-0.2405*(-0.005363878))
```

```
## [1] 2392.548
```

Which group has the shorter median survival time? Does this make sense given your results in part (a) and part (b)? Note: Although the median survival times will be outside the range of the data, please still report the quantities.

The treatment group median survival time is 4724 days and the control group median survival time is 2393 days. Control group has the shorter median survival time.

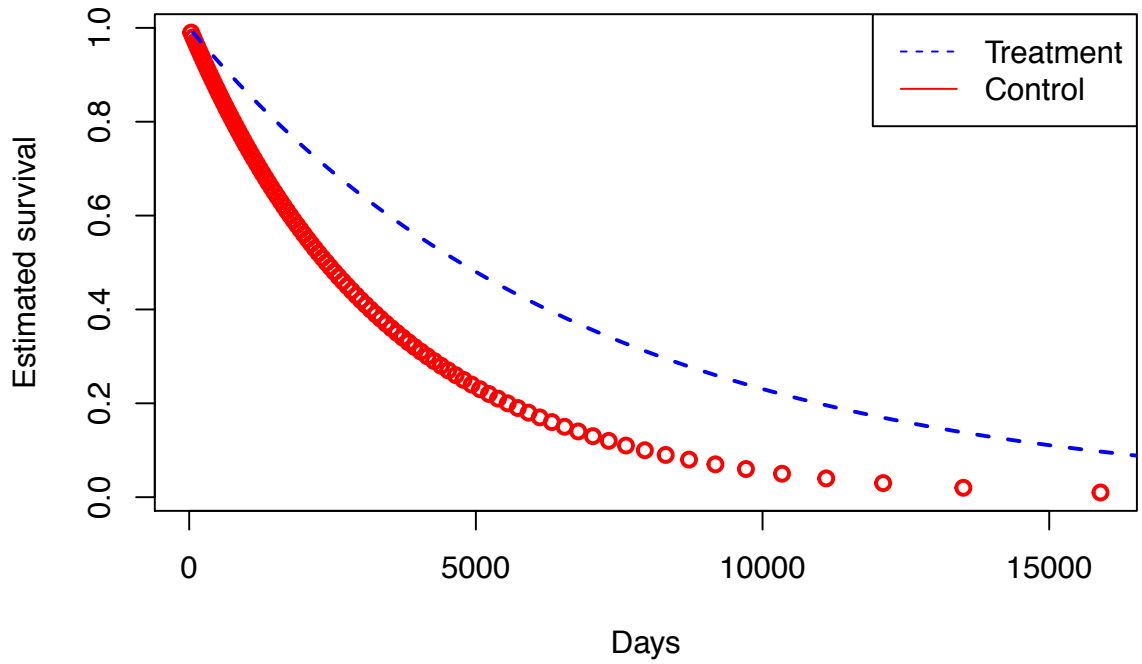
γ = TR = 4723.683/2392.38 = 1.97447 which falls in 95% CI TR for tx = (1.30, 3.01) in (a).

We also know that hazard larger, survival time less. Therefore the result corresponds to (b) as well that treatment harzard is less than control, therefore treatment group median survival time is longer than control group.

e. Plot the covariate-adjusted survival curve for each treatment group when CD4 count = 86.5 cells/milliliter and age = 38.6 years. Does the relative positioning of the control group curve and the treatment group curve make sense given your results in part (a) and (b)?

The blue line is for treatment group, and the red line is for control group.

```
#Exponential survival curves for each treatmeng group
plot(predict(fitex, newdata=list(tx=0, cd4scaled =  1.73, agez = -0.005363878), type="quantile", p=seq(
lines(predict(fitex, newdata=list(tx=1, cd4scaled =  1.73, agez = -0.005363878), type="quantile", p=seq
legend("topright",legend=c("Treatment","Control"),col = c("blue","red"),lty = c(2,1))
```



This makes sense given my results in part (a) and (b) because patients in the treatment group have higher median survival time over survival experience than the control group.

2. [25] Fit the same model assuming a Weibull regression model.

```
fitwei = survreg(Surv(time, censor) ~ tx + cd4scaled + agez,  data=dt, dist="weibull")
summary(fitwei)
```

```
##
## Call:
```

```
## survreg(formula = Surv(time, censor) ~ tx + cd4scaled + agez,
##     data = dt, dist = "weibull")
##               Value Std. Error     z       p
## (Intercept)  7.0613     0.2541 27.79 < 2e-16
## tx           0.8587     0.2870  2.99  0.0028
## cd4scaled    1.0621     0.1928  5.51 3.6e-08
## agez        -0.3119     0.1303 -2.39  0.0167
## Log(scale)   0.2505     0.0972  2.58  0.0099
##
## Scale= 1.28
##
## Weibull distribution
## Loglik(model)= -813.4   Loglik(intercept only)= -852.9
##  Chisq= 79.05 on 3 degrees of freedom, p= 4.9e-17
## Number of Newton-Raphson Iterations: 9
## n= 1151
```

a. Using the fitted model, estimate the time ratio with a 95% confidence interval comparing treatment to control. Estimate the time ratio with a 95% confidence interval for a 50-unit increase in CD4 count. Estimate the time ratio with a 95% confidence interval for a 1-unit increase in age.

We know that

$$\hat{TR}_k = e^{\hat{\beta}_k} \quad 95\%\text{CITR}_k : \left(e^{\text{lcl}(\hat{\beta}_k)}, e^{\text{ucl}(\hat{\beta}_k)}\right)$$

We know that $ln(\hat{TR}_k) = 0.8587$ for tx and 95% CI $ln(TR) = (0.8587 - 1.96 * 0.287, 0.8587 + 1.96 * 0.287) = (0.29618, 1.42122)$. Therefore, 95% CI TR) for tx = (exp(0.29618), exp(1.42122)) = (1.344712, 4.142171).

Similary, $ln(\hat{TR}_k) = 1.0621$ for cd4scaled and 95% CI $ln(TR) = (1.0621 - 1.96 * 0.1928, 1.0621 + 1.96 * 0.1928) = (0.684212, 1.439988)$. Therefore, 95% CI TR) for cd4scaled = (exp(0.684212), exp(1.439988)) = (1.982209, 4.220645).

$ln(\hat{TR}_k) = -0.3119$ for age and 95% CI $ln(TR) = (-0.3119 - 1.96 * 0.1303, -0.3119 + 1.96 * 0.1303) = (-0.567288, -0.056512)$. Therefore, 95% CI TR) for cd4scaled = (exp(-0.567288), exp(-0.056512)) = (0.5670612, 0.9450551).

Interpret the estimated time ratios.

$\hat{TR}$ for tx = $exp(0.8587) = 2.360091$. Median survival time for treatment group is 2.360091 times that of control group.
$\hat{TR}$ for cd4scaled = $exp(1.0621) = 2.892439$. With cd4scaled increasing by 1 unit, meadian survival time increases by 2.892439.
$\hat{TR}$ for age = $exp(-0.3119) = 0.7320547$ With a 1-standard deviation increase, meadian survival time increases by 0.9730693.

b. Using the fitted model, compute point estimates of the hazard ratio comparing treatment to control, the hazard ratio for a 50-unit increase of in CD4 count, and the hazard ratio for a 1-unit increase in age. Interpret the estimated hazard ratios.

$$\hat{HR} = exp(-\hat{\beta}/\hat{\sigma})$$

```
exp(0.2505)
```

```
## [1] 1.284668
```

$\hat{TR}$ for tx = $exp(-0.8587/1.284668) = 0.5125168$. Hazard of death in treatment group is 0.5125168 times that of control group.
$\hat{TR}$ for cd4scaled = $exp(-1.0621/1.284668) = 0.4374685$. With cd4scaled increasing by 1 unit, hazards of death increases 0.4374685 many times than before.
$\hat{TR}$ for age = $exp(0.3119/1.284668) = 1.274796$. With a 1-standard deviation increase in age, hazards of death increases 27.4%.

c. Using the fitted model, calculate the median survival time for each treatment group when CD4 count = 86.5 cells/milliliter and age = 38.6 years.

cd4scaled = cd4/50 = 86.5/50 = 1.73 agez = -0.005363878

$$t_{50}(X) = (-\ln(0.5))^{\sigma} \exp(\beta_0 + \beta_1 X)$$

```
(-log(0.5))^1.284668 * exp(7.0613+0.8587*1+1.0621*1.73-0.3119*(-0.005363878)) # Treatmeent group
```

```
## [1] 10810.34
```

```
(-log(0.5))^1.284668 * exp(7.0613+0.8587*0+1.0621*1.73-0.3119*(-0.005363878)) # Placebo group
```

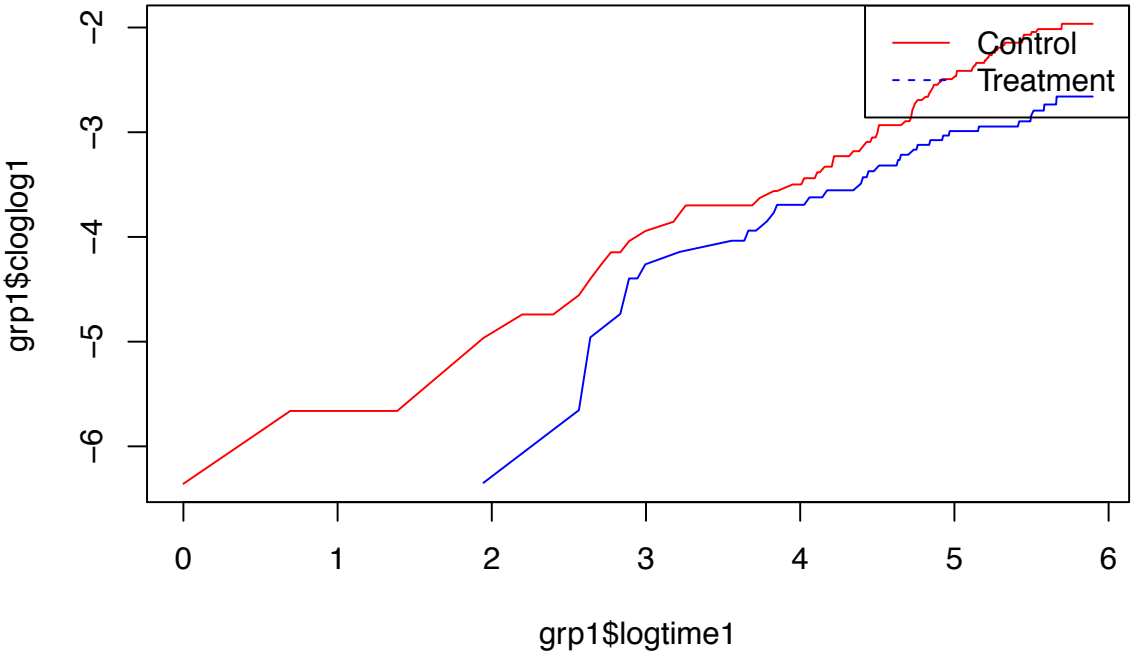
```
## [1] 4580.478
```

d. What does the estimated Weibull scale parameter indicate about the shape of the hazard function in this model?

The scale parameter is $1.284668 > 1$, so we conclude that the hazard is decreasing over time.

e. Graphically check if these data support the proportional hazards and AFT assumptions for the treatment group predictor variable.

```
km1 = survfit(Surv(time, censor) ~ 1, data = dt[dt$tx==1,]) #KM by group
km0 = survfit(Surv(time, censor) ~ 1, data = dt[dt$tx==0,])

time1 = km1$time; logtime1 = log(time1) #log(time)
surv1 = km1$surv; cloglog1 = log(-log(surv1)) #log(-log(S_km(t)))

grp1 = data.frame(time1, logtime1, surv1, cloglog1)
grp1 = grp1[grp1$cloglog1!=Inf,] #In case survival curve ends at 0
grp1 = grp1[-c(1,2),]
time0 = km0$time; logtime0 = log(time0) #log(time)
surv0 = km0$surv; cloglog0 = log(-log(surv0)) #log(-log(S_km(t)))
grp0 = data.frame(time0, logtime0, surv0, cloglog0)
grp0 = grp0[grp0$cloglog0!=Inf,] #In case survival curve ends at 0
plot(grp1$logtime1, grp1$cloglog1, col="blue", type="l", xlim=c(min(grp0$logtime0,grp1$logtime1),max(gr
lines(grp0$logtime0, grp0$cloglog0, col="red", type="l")
legend("topright",legend=c("Control","Treatment"),col = c("red","blue"),lty=1:2)
```



Two lines are roughlly parallel suggesting that these data support the proportional hazards and AFT assumptions for the treatment group predictor variable.

3. [15] Fit the same model assuming a log-logistic regression model.

```
fitloglog = survreg(Surv(time, censor) ~ tx + cd4scaled + agez, data=dt, dist="loglogistic")
summary(fitloglog)
```

```
##
## Call:
## survreg(formula = Surv(time, censor) ~ tx + cd4scaled + agez,
##     data = dt, dist = "loglogistic")
##                Value Std. Error     z       p
## (Intercept)  6.8193     0.2559 26.65 < 2e-16
## tx           0.8663     0.2916  2.97   0.003
## cd4scaled    1.0724     0.1909  5.62 1.9e-08
## agez        -0.3229     0.1351 -2.39   0.017
## Log(scale)   0.2026     0.0956  2.12   0.034
##
## Scale= 1.22
##
## Log logistic distribution
## Loglik(model)= -812.9   Loglik(intercept only)= -852.7
##  Chisq= 79.61 on 3 degrees of freedom, p= 3.7e-17
## Number of Newton-Raphson Iterations: 6
## n= 1151
```

a. Using the fitted model, estimate the time ratio with a 95% confidence interval comparing treatment to control. Estimate the time ratio with a 95% confidence interval for a 50-unit increase in CD4 count. Estimate the time ratio with a 95% confidence interval for a 1-unit increase in age. Interpret the estimated time ratios.

We know that $ln(\hat{TR}_k) = 0.8663$ for tx and 95% CI $ln(TR) = (0.8663 - 1.96 * 0.2916, 0.8663 + 1.96 * 0.2916) = (0.294764, 1.437836)$. Therefore, 95% CI TR) for tx = (exp(0.294764), exp(1.437836)) = (1.342809, 4.211572).

Similary, $ln(\hat{TR}_k) = 1.0724$ for cd4scaled and 95% CI $ln(TR) = (1.0724 - 1.96 * 0.1909, 1.0724 + 1.96 * 0.1909) = (0.698236, 1.446564$. Therefore, 95% CI TR) for cd4scaled = (exp(0.698236), exp(1.446564)) = (2.010204, 4.248492).

$ln(\hat{TR}_k) = -0.3229$ for age and 95% CI $ln(TR) = (-0.3229 - 1.96 * 0.1351, -0.3229 + 1.96 * 0.1351) = (-0.587696, -0.058104)$. Therefore, 95% CI TR) for cd4scaled = (exp(-0.587696), exp(-0.058104)) = (0.5556059, 0.9435518).

Interpret the estimated time ratios.

$\hat{TR}$ for tx = $exp(0.8587) = 2.360091$. Median survival time for treatment group is 2.360091 times that of control group.
$\hat{TR}$ for cd4scaled = $exp(1.0724) = 2.922385$. With cd4scaled increasing by 1 unit, meadian survival time increases by 2.922385.
$\hat{TR}$ for age = $exp(-0.3229) = 0.7240463$. With a 1-standard deviation increase in age, meadian survival time increases by 0.7240463.

b. Using the fitted model, estimate the failure odds ratio comparing treatment to control, the failure odds ratio for a 50-unit increase in CD4 count, and the failure odds ratio for a 1-unit increase in age.

$$\text{OR}_F(t) = \frac{(1 - S(t|x = 1))/(S(t|x = 1))}{(1 - S(t|x = 0))/(S(t|x = 0))} = e^{\beta_1^{*'}}$$

where

$$\beta_k^{*'} = -\beta_k/\sigma$$

```
exp(-0.8663/1.22 ) #comparing treatment to control
```

```
## [1] 0.4916039
```

```
exp(-1.0724/1.22) #for a 50-unit increase in CD4 count
```

```
## [1] 0.4151911
```

```
exp(0.3229/1.22) #for a 1-unit increase in age
```

```
## [1] 1.303004
```

Interpretation: The odds of failure for subjects before time t in treatment group is 0.49 times that for subjects in control group , and this holds for all t. The odds of failure before time t decreases by 58.5%, and this holds for all t for every 50-unit increase in CD4 count. The odds of failure before time t increases by 30.3% for 1-standard deviation increase in age.

Also estimate the survival odds ratio comparing treatment to control, the survival odds ratio for a 50-unit increase in CD4 count, and the survival odds ratio for a 1-unit increase in age. Interpret the estimated failure odds and survival odds ratios.

$$\text{OR}_S(t) = \frac{S(t|x = 1)/(1 - S(t|x = 1))}{S(t|x = 0)/(1 - S(t|x = 0))} = e^{\beta_1^*}$$

where

$$\beta_k^{*'} = \beta_k/\sigma$$

```r
exp(0.8663/1.22) #comparing treatment to control
```

```
## [1] 2.034158
```

```r
exp(1.0724/1.22) #for a 50-unit increase in CD4 count
```

```
## [1] 2.408529
```

```r
exp(-0.3229/1.22) #for a 1-unit increase in age
```

```
## [1] 0.7674575
```

The odds of survival beyond time t among subjects in the treatment group is 2.034 times of that of subjects in the control group, and this holds for all t. The odds of survival before time t increases by 141%, and this holds for all t for every 50-unit increase in CD4 count. The odds of survival before time t decrease by 23% for 1-standard deviation increase in age.

c. Compare the interpretation of the time, failure odds, and survival odds ratios computed in (a), (b).

The decrease and increase are opposite in (a) and (b) which makes sense - because survival and failure are opposite. The longer median survival times corresponds to higher survival OR and lower failure OR.

4. [10] Fit the same model assuming a log-normal regression model.

```r
fitlnor = survreg(Surv(time, censor) ~  tx + cd4scaled + agez, data=dt, dist="lognormal")
summary(fitlnor)
```

```
##
## Call:
## survreg(formula = Surv(time, censor) ~ tx + cd4scaled + agez,
##     data = dt, dist = "lognormal")
##                Value Std. Error     z       p
## (Intercept)  7.3591     0.3172 23.20 < 2e-16
## tx           0.8576     0.3047  2.81  0.0049
## cd4scaled    1.0942     0.1782  6.14 8.2e-10
## agez        -0.3169     0.1435 -2.21  0.0273
## Log(scale)   0.9540     0.0861 11.08 < 2e-16
##
## Scale= 2.6
##
## Log Normal distribution
## Loglik(model)= -811.9   Loglik(intercept only)= -851.4
##  Chisq= 78.94 on 3 degrees of freedom, p= 5.2e-17
## Number of Newton-Raphson Iterations: 5
## n= 1151
```

a. Using the fitted model, estimate the time ratio with a 95% confidence interval comparing treatment to control. Estimate the time ratio with a 95% confidence interval for a 50-unit increase in CD4 count. Estimate the time ratio with a 95% confidence interval for a 1-unit increase in age.

For tx,

$$\hat{\beta}_k \pm 1.96\hat{SE}\left(\hat{\beta}_k\right) = (0.8576 - 1.96 * 0.3047, 0.8576 + 1.96 * 0.3047) = (0.260388, 1.454812)$$

$$(exp(0.260388), exp(1.454812)) = (1.297433, 4.283678)$$

For cd4scaled,

$$\hat{\beta}_k \pm 1.96\hat{S}E\left(\hat{\beta}_k\right) = (1.0942 - 1.96 * 0.1782, 1.0942 + 1.96 * 0.1782) = (0.744928, 1.443472)$$

$$(exp(0.744928), exp(1.443472)) = (2.106290, 4.235376)$$

For agez,

$$\hat{\beta}_k \pm 1.96\hat{S}E\left(\hat{\beta}_k\right) = (-0.3169 - 1.96 * 0.1435, -0.3169 + 1.96 * 0.1435) = (-0.59816, -0.03564)$$

$$(exp(-0.59816), exp(-0.03564)) = (0.5498224, 0.9649876)$$

Interpret the estimated time ratios.

```
exp( 0.8576)
```

```
## [1] 2.357496
```

```
exp(1.0942)
```

```
## [1] 2.986792
```

```
exp(-0.3169)
```

```
## [1] 0.7284036
```

Median survival time for treatment group is 2.357496 times that of group 0. With cd4scaled increasing by 1 unit, meadian survival time increases by 2.986792.
With a 1-standard deviation increase in age, meadian survival time increase by 0.7284036.

b. What is the estimated probability of surviving past 180 days in each group when the CD4 count = 86.5 cells/milliliter and age = 38.6 years? Do these estimates make sense given your results in part (a)?

$$S(t|X) = 1 - \Phi\left[\frac{\ln(t)}{\sigma} - \frac{\beta_0}{\sigma} - \frac{\beta_1}{\sigma}X\right]$$

For treatment group,

$$S(t = 180|X = 1) = 1 - \Phi\left[\frac{\ln(180)}{2.6} - \frac{7.3591}{2.6} - \frac{0.8576}{2.6} - \frac{1.0942}{2.6}\frac{86.5}{50} - \frac{0.3169}{2.6} * -0.005363878\right]$$

```
1-pnorm((log(180) - 7.3591 - 0.8576 - 1.0942*1.73 + 0.3169*0.005363878)/2.6)
```

```
## [1] 0.970647
```

For control group,

```
1-pnorm((log(180) - 7.3591- 1.0942*1.73 + 0.3169*0.005363878)/2.6)
```

```
## [1] 0.9406841
```

The result shows that, 97.1% of patients in the treatment are estimated to survive past 180 day and 94.1% of patients in the control are estimated to survive past 180 day. This makes sense given the results in part (a) because the median survival time in the treatment group is longer than the control group.

5. [5] Fit the same model assuming a gamma regression model.

```
fit.gamma = flexsurvreg(Surv(time,censor)~tx+cd4scaled+agez, data = dt, dist = "gengamma")
fit.gamma
```

```
## Call:
## flexsurvreg(formula = Surv(time, censor) ~ tx + cd4scaled + agez,
##     data = dt, dist = "gengamma")
##
## Estimates:
##             data mean  est        L95%       U95%       se         exp(est)
## mu                 NA   7.34e+00   6.71e+00   7.98e+00   3.24e-01         NA
## sigma              NA   2.45e+00   1.37e+00   4.38e+00   7.25e-01         NA
## Q                  NA   1.03e-01  -8.57e-01   1.06e+00   4.90e-01         NA
## tx           4.99e-01   8.60e-01   2.65e-01   1.45e+00   3.03e-01   2.36e+00
## cd4scaled    1.73e+00   1.09e+00   7.40e-01   1.45e+00   1.80e-01   2.98e+00
## agez         1.42e-16  -3.18e-01  -5.98e-01  -3.86e-02   1.43e-01   7.27e-01
##             L95%       U95%
## mu                 NA         NA
## sigma              NA         NA
## Q                  NA         NA
## tx           1.30e+00   4.28e+00
## cd4scaled    2.10e+00   4.25e+00
## agez         5.50e-01   9.62e-01
##
## N = 1151,  Events: 96,  Censored: 1055
## Total time at risk: 264941
## Log-likelihood = -811.9101, df = 6
## AIC = 1635.82
```

a. Using the fitted model, estimate the time ratio with a 95% confidence interval comparing treatment to control. Estimate the time ratio with a 95% confidence interval for a 50-unit increase in CD4 count. Estimate the time ratio with a 95% confidence interval for a 1-unit increase in age.

For treatment, (exp(2.65e-01),exp(1.45e+00)) = (1.303431 4.263115)
For cd4scale, (exp(7.40e-01),exp(1.45e+00)) = (2.095936, 4.263115)
For agez, (exp(-5.98e-01),exp(-3.86e-02)) = (0.5499104, 0.9621355)
Interpret these estimates. We are 95% confident that the time ratio between treatment group and control group in the population of patients in this trail is between 1.30 and 4.263, adjusting for CD4 count and age. We are 95% confident that every 50-unit increase in CD4 count in the population of patients in this trail is between 2.10 and 4.26, adjusting for treatment group and age. We are 95% confident that the age in the population of patients in this trail is between 0.55 and 0.96,adjusting for treatment group and CD4 count.

6. [20] Which of the five models (exponential, Weibull, log-logistic, log-normal, or gamma regression model fit in questions 1-5) is the best model for these data? Justify your response using plots, statistical tests (when appropriate), and AIC. Note: The choice will not be clear-cut, but justify your pick. Be sure to state if any of the models are clearly not appropriate and why.

Likelihood ratio test

- H_0: the more complicated model is not providing a significantly better fit than the simpler model (the two models are equivalent). Therefore, we prefer the simpler model.
  H_1: the more complicated model is providing a significantly better fit than the simpler model. Prefer the more complicated model.

- Significane level: two sided, $\alpha = 0.05$

- Test statistics:
$$G = -2ll(\text{reduced}) - (-2ll(full)) \tilde{\sim} \chi^2_{\Delta s}$$

- Decision rule: At $\alpha = 0.05$, reject H_0 if $W^2 \geq \chi_{1-0.05,df}$. 3.84 for 1df, 5.99 for 2df.

We know that

$$\begin{aligned}
&\text{Gamma } (\delta, \sigma) \supset \text{ exponential } (\delta = \sigma = 1) \\
&\text{Gamma } (\delta, \sigma) \supset \text{ Weibull } (\delta = 1, \sigma) \supset \text{ exponential } (\delta = \sigma = 1) \\
&\text{Gamma } (\delta, \sigma) \supset \text{ log-normal } (\delta = 0, \sigma)
\end{aligned}$$

```
#(1) Compare Exp to Gamma, 2 dof
lrt1 = -2*(fitex$loglik[2] - fit.gamma$loglik)
lrt1
```

```
## [1] 10.29908
```

```
pchisq(lrt1, 2, lower.tail=FALSE)
```

```
## [1] 0.005802082
```

- Statistical conclusion: Since $G = 10.29908 > 5.99 with p = 0.0058$, reject H_0, p = 2.348e-05

- Interpretation: Reject the null hypothesis and we conclude than the full model Gamma is better than the reduced model Exp.

```
#(2) Compare Exp to Weibull, 1 dof
anova(fitex,fitwei) #  p = 0.0069
```

```
##                   Terms Resid. Df    -2*LL Test Df Deviance    Pr(>Chi)
## 1 tx + cd4scaled + agez      1147 1634.119      NA       NA          NA
## 2 tx + cd4scaled + agez      1146 1626.825    =  1 7.294516 0.006916539
```

- Statistical conclusion: Since $p = 0.0069$, reject H_0

- Interpretation: Reject the null hypothesis and we conclude than the full model Weibull is better than the reduced model Exp.

```
#(3) Compare Weibull to Gamma, 1 dof
lrt2 = -2*(fitwei$loglik[2] - fit.gamma$loglik)
lrt2
```

```
## [1] 3.004561
```

```
pchisq(lrt2, 1, lower.tail=FALSE)
```

```
## [1] 0.08303049
```

- Statistical conclusion: Since $G = 3.00 < 3.84 with p = 0.083$, fail to reject H_0

- Interpretation: We fail to reject the null hypothesis and more complicated model Gamma is not providing a significantly better fit than the simpler model Weibull. We need futher testing (e.g. comparing AIC)

```
#(4) Compare log-normal to Gamma, 1dof
lrt3 = -2*(fitlnor$loglik[2] - fit.gamma$loglik)
lrt3
```

```
## [1] 0.04249292
```

```
pchisq(lrt3, 1, lower.tail=FALSE)
```

```
## [1] 0.836683
```

- Statistical conclusion: Since $G = 0.04249292 < 3.84 with p = 0.083$, fail to reject H_0
- Interpretation: We fail to reject the null hypothesis and more complicated model Gamma is not providing a significantly better fit than the simpler model log-normal We need futher testing (e.g. comparing AIC)

Overall, Gamma and Weibull is better than Exp.

Comparing AIC

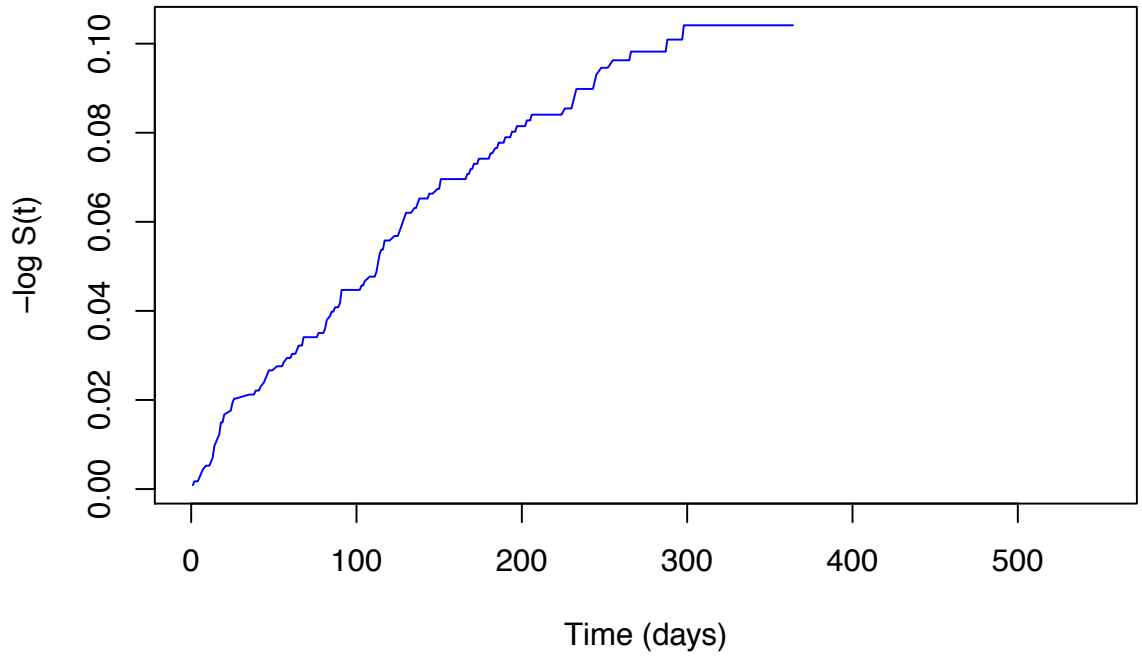
```
c(AIC(fitex), AIC(fit.gamma),AIC(fitlnor), AIC(fitloglog), AIC(fitwei))
```

```
## [1] 1642.119 1635.820 1633.863 1635.781 1636.825
```

Based on the AIC, the log-normal gives the lowest AIC, therefore, it's the best model based on AIC comparison.

Graphical methods

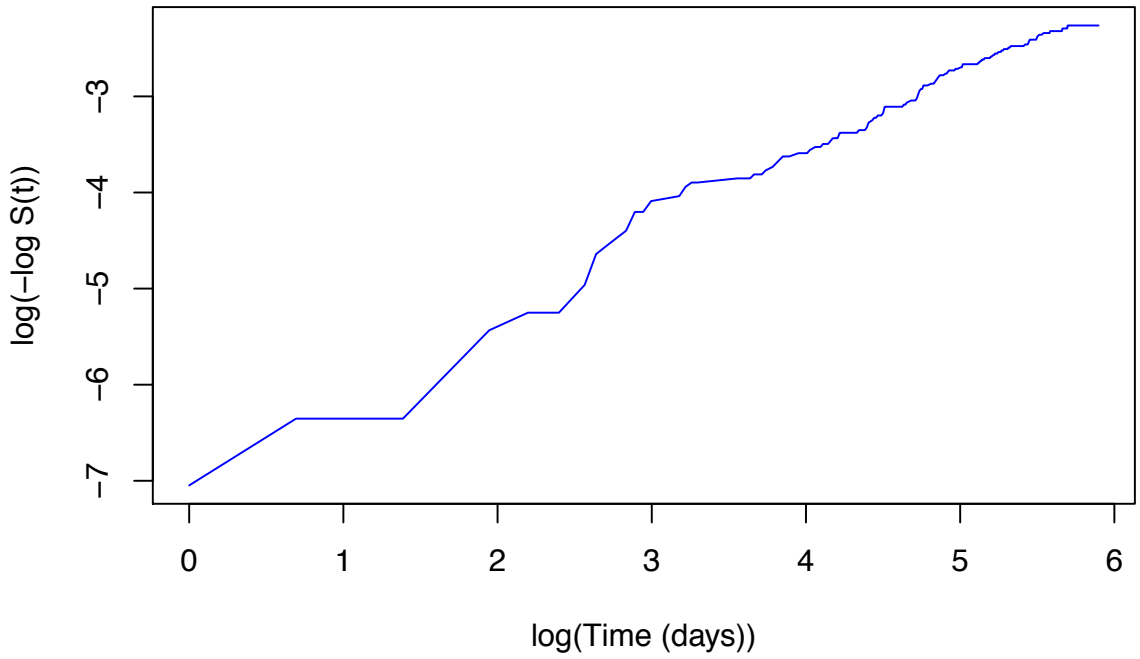
```
#Manual -logS(t) plot  for Exponential Distribution
km = survfit(Surv(time, censor) ~ 1, data=dt) # KM survival probabilities
time = km$time
logS = -log(km$surv) # -log(S_km(t))
forplot = data.frame(time, logS)
forplot = forplot[forplot$logS != -Inf,] # In case survival curve ends at 0
plot(forplot$time, forplot$logS, col="blue", type="l", xlim=c(0,550), xlab="Time (days)", ylab="-log S(
```



This

graph looks kind of linear expect the end part. Therefore, the exponential model is a good choice.

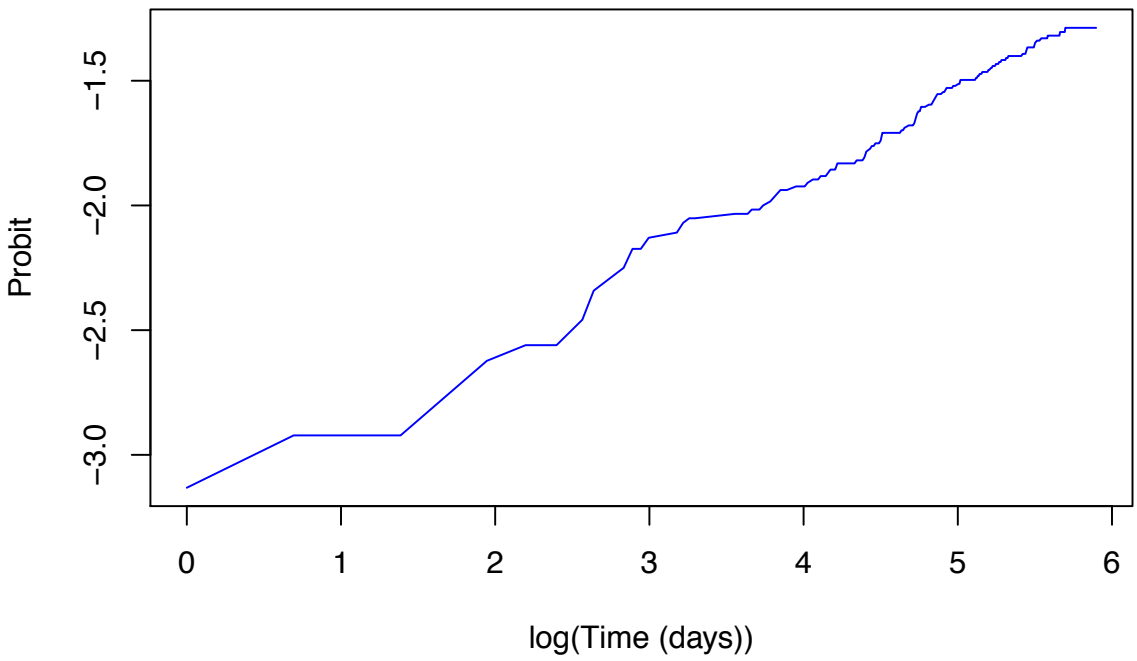
```
#Weibull
km = survfit(Surv(time, censor) ~ 1, data=dt) # KM survival probabilities
time = km$time
logtime = log(time)
surv = km$surv; cloglog = log(-log(surv))
forplot = data.frame(time, logtime, surv, cloglog)
forplot = forplot[forplot$cloglog != Inf,] # In case survival curve ends at 0
plot(forplot$logtime, forplot$cloglog, col="blue", type="l", xlab="log(Time (days))", ylab="log(-log S(
```



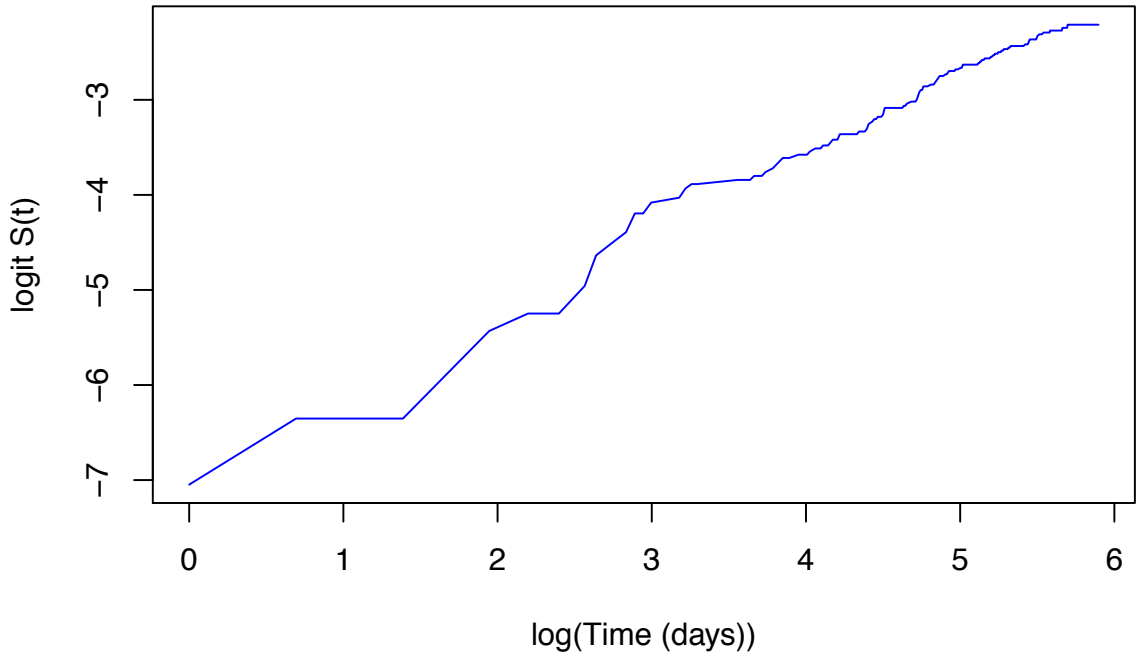
This graph looks roughly linear. Therefore, the Weibull model is a good option.

log-log and log-normal

```
km = survfit(Surv(time, censor) ~ 1, data=dt)
time = km$time;
surv = km$surv
logS = -log(surv)
cloglog = log(-log(surv))
lnorm = qnorm(1-surv)
logit = log((1-surv)/surv)
logtime = log(time)
# S_km(t)
# -log(S_km(t))
# log(-log(S_km(t)))
# probit(1-S_km(t))
# log((1-S_km(t))/S_km(t))
forplot = data.frame(time, logtime, surv, logS, cloglog, logit, lnorm)
forplot = forplot[forplot$surv != 0,] # In case survival curve ends at 0
plot(forplot$logtime, forplot$lnorm, col="blue", type="l", xlab="log(Time (days))", ylab="Probit")
```




```
plot(forplot$logtime, forplot$logit, col="blue", type="l", xlab="log(Time (days))", ylab="logit S(t)")
```

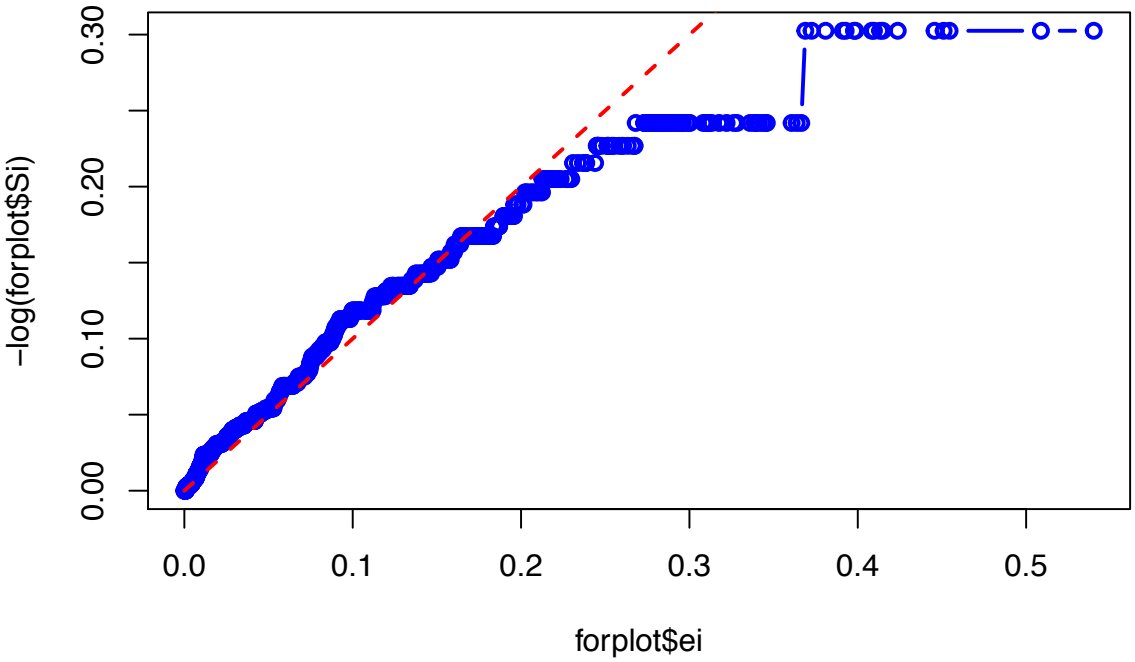


These graph looks roughly linear. Therefore, the lognormall and log-logistic are good options.

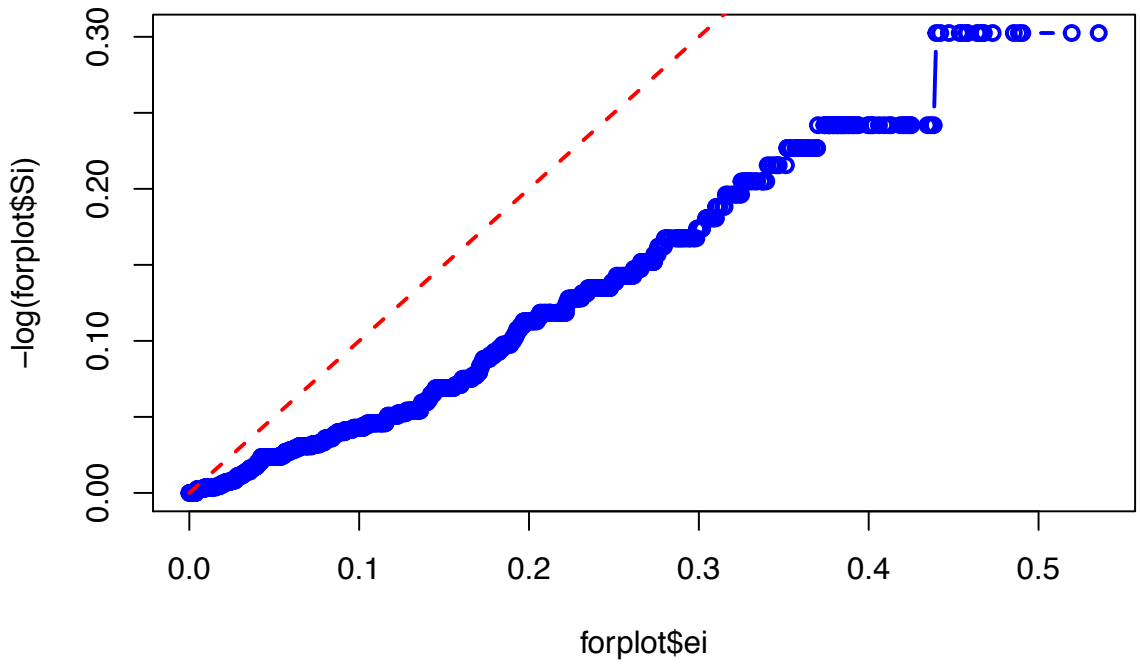
Residual checking

```
# Exponential
linpred = fitex$linear.predictor # ~beta0 + ~beta1X1 + ... from AFT model sig.hat = fitex$scale
sig.hat = fitex$scale
alpha.hat = 1/sig.hat
ti = cbind(Surv(dt$time, dt$censor))[,1]
```

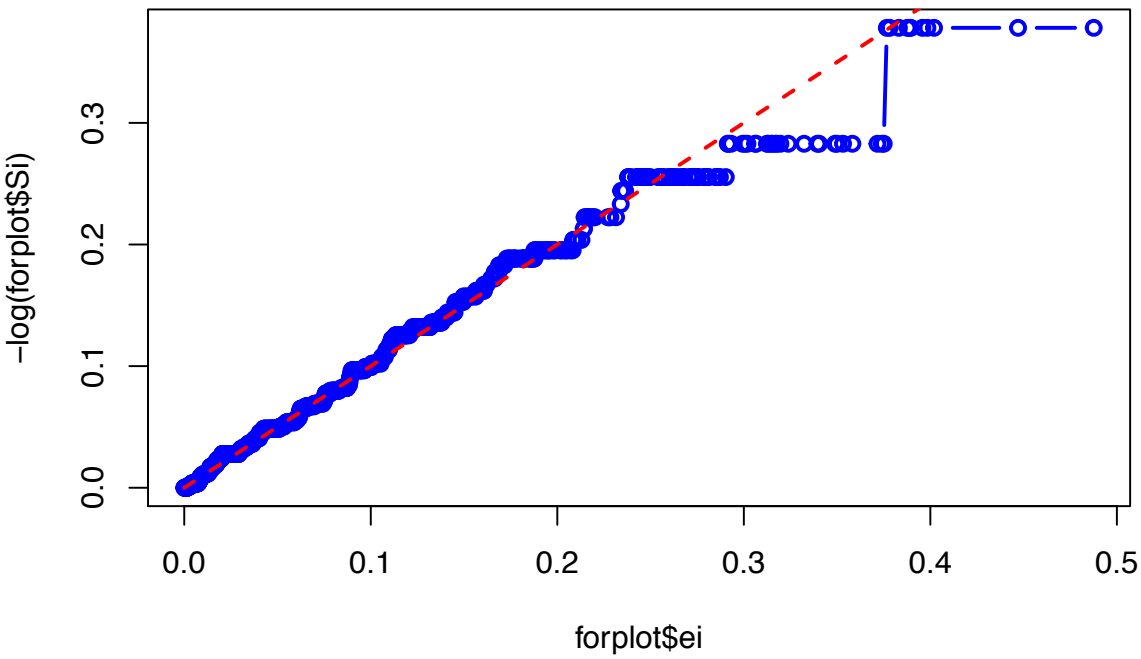
```r
coxsnell.expon = ti*exp(-linpred) # H(t) based on model, H(t)=-ln(S(t))
cs.fit = survfit(Surv(coxsnell.expon, dt$censor) ~ 1)
ei = cs.fit$time # time = Cox-Snell residual
Si = cs.fit$surv
forplot = data.frame(ei, Si)
forplot = forplot[Si != 0,] # In case survival curve ends at 0
plot(forplot$ei, -log(forplot$Si), type="b", lwd=2, col="blue")
lines(c(0,3),c(0,3), lty=2, lwd=2, col="red") # Reference line
```



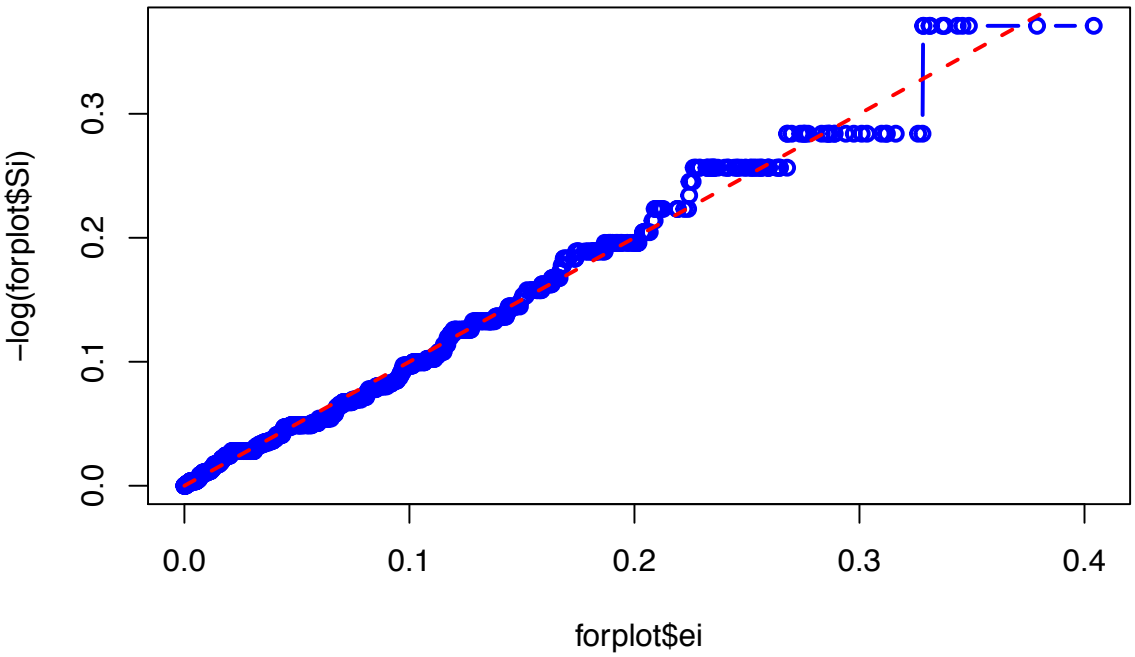

```r
# Gamma
coefficients = fit.gamma$res[,1] # Intercept, scale and shape, estimated slopes betas = fitgg2$coeffici
sig.hat = unname(coefficients[2]) # scale
delta.hat = unname(coefficients[3]) # shape (a.k.a. "delta", "Q", or "theta")
ti = cbind(Surv(dt$time, dt$censor))[,1] # survival times
X = as.matrix(fit.gamma$data$mml$mu) # data including column of 1s for intercept linpred = X %*% betas
vtodelta = (ti*exp(-linpred))^(delta.hat/sig.hat)

library(pracma) # Contains incomplete gamma function, gammainc()
ginc = function(x) gammainc(delta.hat^(-2) * x, delta.hat^(-2)) # for use in sapply
foo = t(sapply(vtodelta, ginc))[,3] # Want the third element that this function
# returns (regularized lower incomplete gamma function)
Si = ifelse(rep(delta.hat,length(ti)) > 0, 1-foo, foo)
coxsnell.gg = -log(Si) # H(t) based on model, H(t)=-ln(S(t))
cs.fit = survfit(Surv(coxsnell.gg, dt$censor) ~ 1)

ei = cs.fit$time # time = Cox-Snell residual
Si = cs.fit$surv
forplot = data.frame(ei, Si)
forplot = forplot[Si != 0,] # In case survival curve ends at 0
plot(forplot$ei, -log(forplot$Si), type="b", lwd=2, col="blue")
lines(c(0,3),c(0,3), lty=2, lwd=2, col="red") # Reference line
```



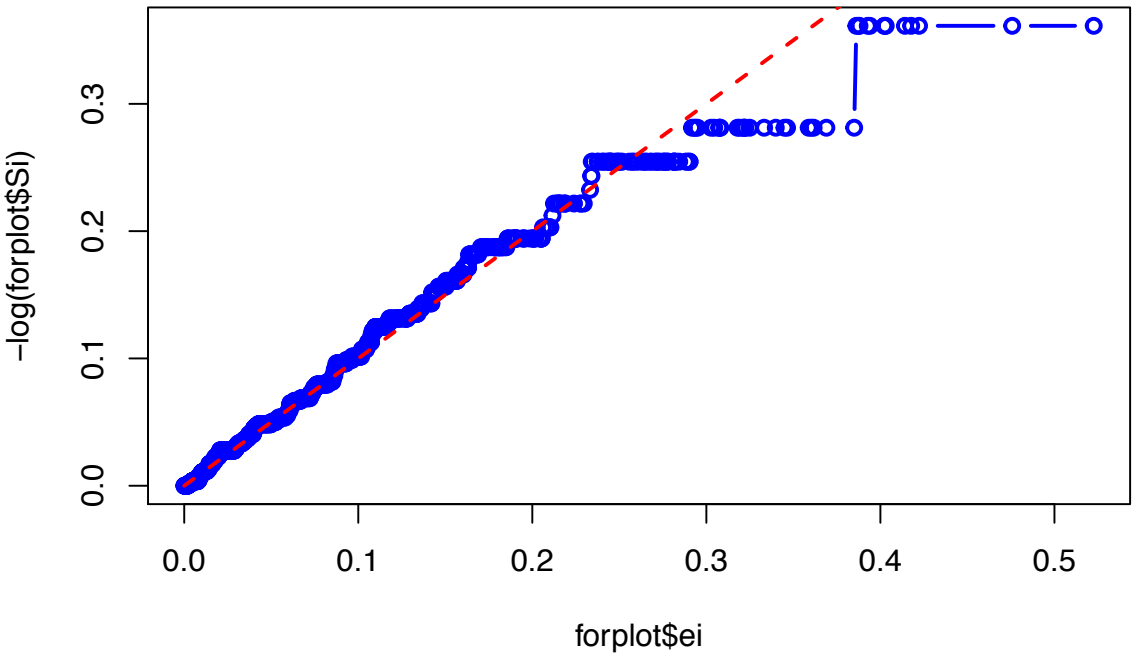

```
# log-logistic normal
linpred = fitloglog$linear.predictor # ~beta0 + ~beta1X1 + ... from AFT model sig.hat = fitex$scale
sig.hat = fitloglog$scale
alpha.hat = 1/sig.hat
ti = cbind(Surv(dt$time, dt$censor))[,1]
coxsnell.expon = -log(1/(1+(ti*exp(-linpred))^alpha.hat))
cs.fit = survfit(Surv(coxsnell.expon, dt$censor) ~ 1)
ei = cs.fit$time # time = Cox-Snell residual
Si = cs.fit$surv
forplot = data.frame(ei, Si)
forplot = forplot[Si != 0,] # In case survival curve ends at 0
plot(forplot$ei, -log(forplot$Si), type="b", lwd=2, col="blue")
lines(c(0,3),c(0,3), lty=2, lwd=2, col="red") # Reference line
```




```
# log-normal
linpred = fitlnor$linear.predictor # ~beta0 + ~beta1X1 + ... from AFT model sig.hat = fitex$scale
sig.hat = fitlnor$scale
alpha.hat = 1/sig.hat
ti = cbind(Surv(dt$time, dt$censor))[,1]
coxsnell.expon = -log(1-pnorm(alpha.hat*log(ti)-linpred*alpha.hat))
cs.fit = survfit(Surv(coxsnell.expon, dt$censor) ~ 1)
ei = cs.fit$time # time = Cox-Snell residual
Si = cs.fit$surv
forplot = data.frame(ei, Si)
forplot = forplot[Si != 0,] # In case survival curve ends at 0
plot(forplot$ei, -log(forplot$Si), type="b", lwd=2, col="blue")
lines(c(0,3),c(0,3), lty=2, lwd=2, col="red") # Reference line
```




```
#Weibull
linpred = fitwei$linear.predictor # ~beta0 + ~beta1X1 + ... from AFT model sig.hat = fitex$scale
sig.hat = fitwei$scale
alpha.hat = 1/sig.hat
ti = cbind(Surv(dt$time, dt$censor))[,1]
coxsnell.expon = (ti*exp(-linpred))^(alpha.hat)
cs.fit = survfit(Surv(coxsnell.expon, dt$censor) ~ 1)
ei = cs.fit$time # time = Cox-Snell residual
Si = cs.fit$surv
forplot = data.frame(ei, Si)
forplot = forplot[Si != 0,] # In case survival curve ends at 0
plot(forplot$ei, -log(forplot$Si), type="b", lwd=2, col="blue")
lines(c(0,3),c(0,3), lty=2, lwd=2, col="red") # Reference line
```



Log-normal is the best by examining residuals and AIC.
log-log and log-normal is the best by graphical method.
LRT gives that Gamma and Weibull is better than Exp.
Therefore, the conclusion from LRT, comparing AIC and residuals all shows that log-normal is the best model.