Analysis of Drug Relapses (Survival Analysis)

Luis Aragon

December 1, 2018

Description

This data set is a Drug Treatment Dataset from *Applied Survival Analysis: Regression Modeling of Time to Event Data* by David Hosmer and Stanley Lemeshow. The data covers the time since a patient was admitted into a treatment program until they relapsed (failure time). For the censored data, a 1 is given if they relapsed (returned to drugs), and a 0 if they were censored. The rest of the columns are interesting variables which we can explore. By looking at the data, we see it contains 575 records. This dataset originally contained 628 records, but now has 575 with no missing values. Each row refers to one patient. Variables include AGE (age at enrollment), BECK (continuous Beck Depression Score from 0-63), TREAT (short vs long treatment), and SITE (treatment site). The aim for modelling this data is to see the time to relapse for these patients and the probability of relapsing over time. The researchers of the study collected many variables, which can be used to model and predict relapse (failure) probabilities throughout time. The researchers and creators of this study intentionally randomized the treatment variable, TREAT, because they wanted to know if there was an effect of treatment length on relapse time of the patients. This is the variable I will be analyzing.

Research Question

Does the treatment length have an effect on relapse time?

I am curious about the effect of treatment length on the probability of someone relapsing after being released from treatment. We know that the Treatment Length is not observational and was randomly assigned by the researchers, so we can form a strong conclusion on the effects of Treatment using the methods in this paper. For other observational variables, like depression, there may be unknown/unmeasured confounders like genetics or stress tolerance that effect the depression score as well as the probability of relapsing/failing. This makes it difficult to form very strong conclusions about the effect of depression.

Important Variables

Variable	Description	Code
ID	Identfication	1-628
AGE	Age at Enrollment	Years
BECK	Beck Depression Scoref	0.000-54.000
IV	History of IV Drug Use	1=Never 2=Previous 3=Recent
TREAT	Treatment Randomization	0=Short 1=Long
SITE	Treatment Site	0=A 1=B
LEN.T	Length of Treatment Stay	Days
TIME	Time to Drug Relapse	Days
CENSOR	Censor Status	1=Relapsed 0=Censored

Look at the Data

```
# Access Data
data_dir <- "/Users/lmaragon/Documents/STATS/PSTAT175/SurvAnalysis/SurvivalAnalysis/Data"
file <- file.path(data_dir, "uis.csv")
relapseData <- read.csv(file, header = TRUE)
```

```
# Dimensions of dataframe
dim(relapseData)
```

```
## [1] 575  19
```

```
# Look at our Data
head(relapseData)
```

```
##   X ID AGE  BECK HC IV NDT RACE TREAT SITE LEN.T TIME CENSOR        Y
## 1 1  1  39  9.00  4  3   1    0     1    0   123  188      1 5.236442
## 2 2  2  33 34.00  4  2   8    0     1    0    25   26      1 3.258097
## 3 3  3  33 10.00  2  3   3    0     1    0     7  207      1 5.332719
## 4 4  4  32 20.00  4  3   1    0     0    0    66  144      1 4.969813
## 5 5  5  24  5.00  2  1   5    1     1    0   173  551      0 6.311735
## 6 6  6  30 32.55  3  3   1    0     1    0    16   32      1 3.465736
##        ND1        ND2      LNDT       FRAC IV3
## 1 5.000000 -8.0471896 0.6931472 0.68333333   1
## 2 1.111111 -0.1170672 2.1972246 0.13888889   0
## 3 2.500000 -2.2907268 1.3862944 0.03888889   1
## 4 5.000000 -8.0471896 0.6931472 0.73333333   1
## 5 1.666667 -0.8513760 1.7917595 0.96111111   0
## 6 5.000000 -8.0471896 0.6931472 0.08888889   1
```

Time

Since we are looking at the effect of the Beck Depression score on time until relapse, we want to look at the time to relapse from when the patient was released from treatment and not from when they started the treatment. Essentialy the patient goes through this flow chart:

Start –> LEN.T (Length of Time) –> TIME (relapsed/censored)

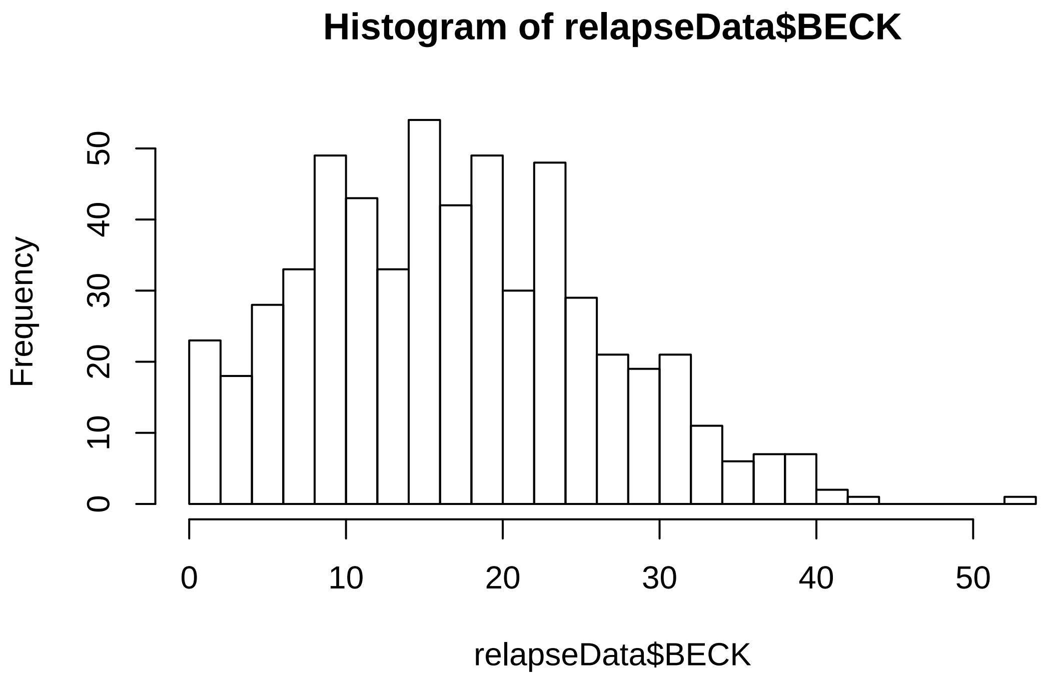
I am going to subtract the length of time in treatment (LEN.T) from the TIME variable. The reason I am doing this is because there may be changing hazard rates throughout the entire study. For example, the hazard of failing (relapsing) is very low during treatment. That means that people assigned a long treatment, have low hazard rates until they are released. Therefore, by being assigned a longer treatment, the patient is does not relapse for a longer time. This is accounted for by adding our new time variale below that measure every patient at the time they are released from treatment.

```
relapseData$TIME2 <- relapseData$TIME - relapseData$LEN.T
```

Beck Depression Score

I am going to look at the distribution of the Beck Depression Score because it is a variable we are interested in.

```
hist(relapseData$BECK, breaks = 20)
```



We can see that the highest point is over 50. The Beck Depression Score, ranges from 0 to 63 and is categorized into 4 categories according to the Psych Congress Network:

Raw Scores	Depression Severity
0-13	Indicated minimal depression
14-19	Indicates mild depression
20-28	Indicates moderate depression
29-63	Indicates severe depression

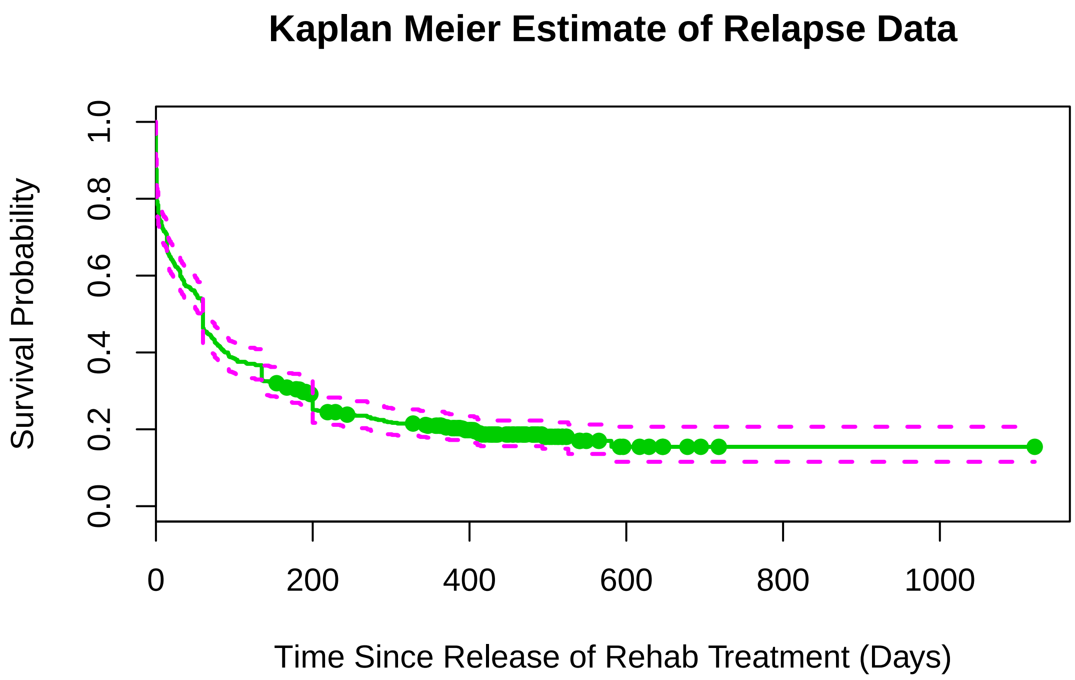
I will discretize this contiuous variable so we can compare the different categories of depression. Because we only have 575 records, I will split the data into 2 groups rather than 4. The reason I am discretizing the Beck Score, is because it is easier to interpret categories of depression than a continuous score. For example, it is more clear discussing the difference between Mild and Severe Depression (discrete) than it is discussing a 10.0 point difference in a depression score (continuous).

```r
# Discretize BECK
relapseData$dep <- "Moderate/Severe"
relapseData$dep[relapseData$BECK <= 19] <- "Minimal/Mild"
```

Basic Estimation

```r
library(survival)
fit.1 <- survfit(Surv(TIME2, CENSOR)~1, data = relapseData)

# Time,Censor ~ 1
plot(fit.1, mark=19, lwd=2,
col=c(3,6,6), xlab="Time Since Release of Rehab Treatment (Days)", ylab="Survival Probability",
main="Kaplan Meier Estimate of Relapse Data", mark.time = TRUE)
```



Above, I plotted the Kaplan Meier Estimate of the relapse data using the `survival` and the `survfit` function.

```
quantile(fit.1, c(.25,.50,.75), conf.int=FALSE)
```

```
##  25  50  75
##   4  60 206
```

We can see that an alarming 25% of patients relapse within 4 days after being released from treatment and 50% of patients relapse within 60 days from being released.

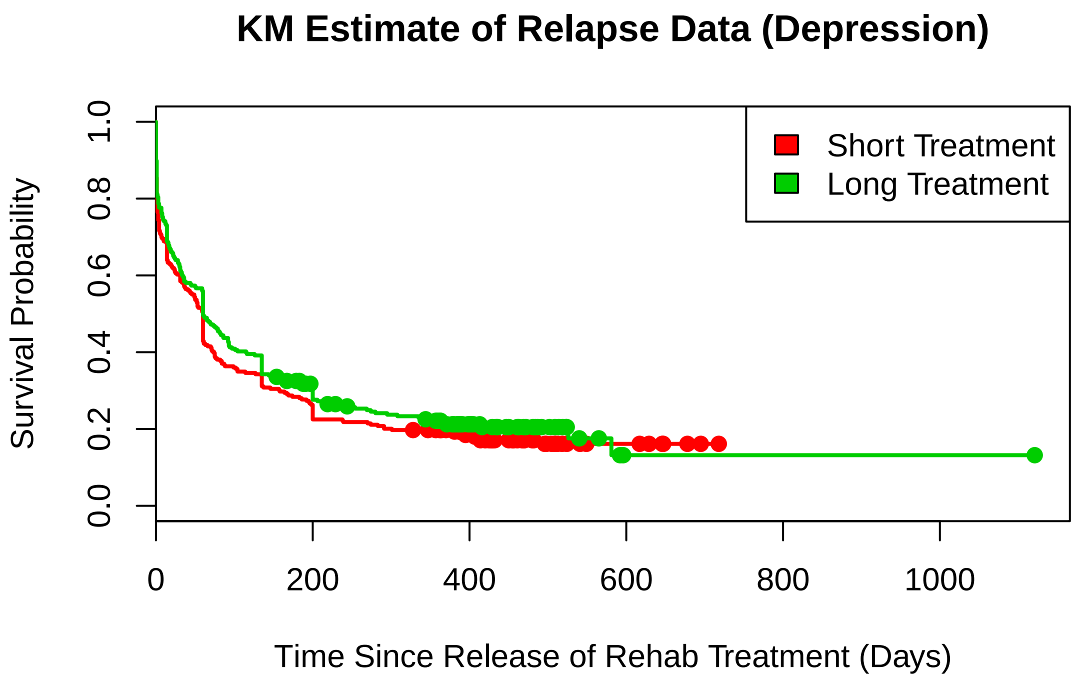
Basic Estimation of Treatment

```
relapseData$TREAT <- as.factor(relapseData$TREAT)
relapseData$SITE <- as.factor(relapseData$SITE)

fit.2.trt <- survfit(Surv(TIME2, CENSOR)~TREAT, data = relapseData)

# Time,Censor ~ Depression
plot(fit.2.trt, mark=19, lwd=2,
col=c(2,3), xlab="Time Since Release of Rehab Treatment (Days)", ylab="Survival Probability",
main="KM Estimate of Relapse Data (Depression)", mark.time = TRUE)

# Legend
legend("topright",c("Short Treatment", "Long Treatment"), fill=c(2,3))
```



Visually, we see that the Long Treament group survives relapse longer than the Short Treatment group by very small amount. We will test to see if this difference is significant.

```
# Summarize above graph
print(fit.2.trt)
```

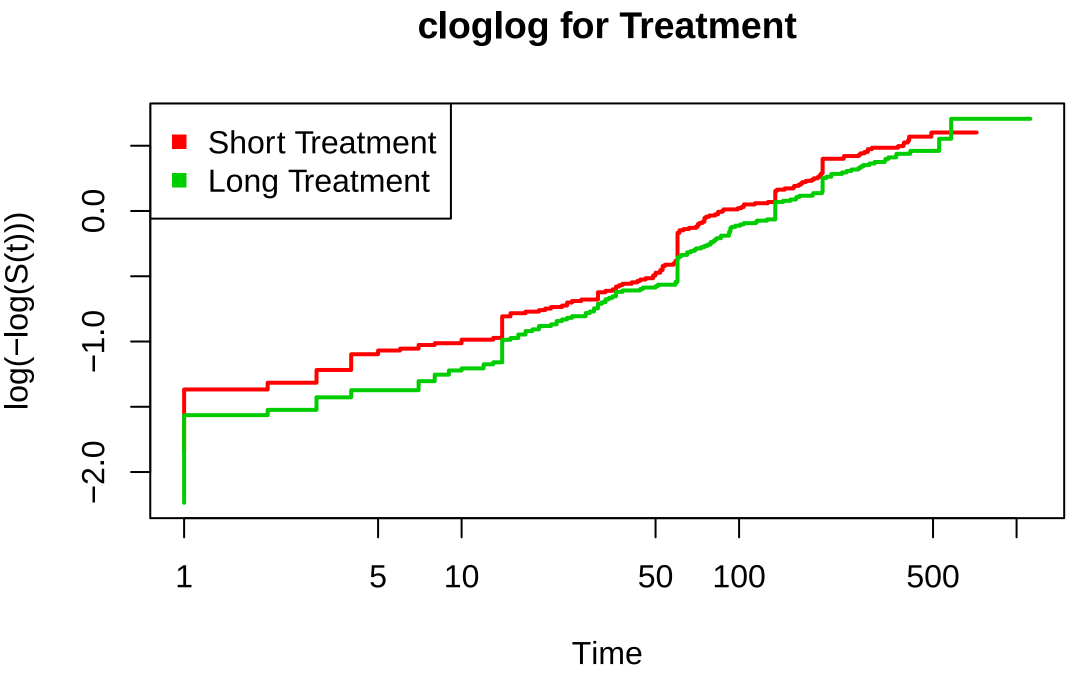
```
## Call: survfit(formula = Surv(TIME2, CENSOR) ~ TREAT, data = relapseData)
##
##            n events median 0.95LCL 0.95UCL
## TREAT=0 289    239     60      44      60
## TREAT=1 286    225     60      60      86
```

We can see that 50% of the people in the short treatment relapse within 60 days of being released from the treatment. This is the same as the people in the long treatment.

Coxph of Treatment Variable

```
cox.1.trt <- coxph(Surv(TIME2, CENSOR)~TREAT, data=relapseData)

plot(fit.2.trt,
fun="cloglog",col=c(2:3),xlab="Time",ylab="log(-log(S(t)))",lwd=2, main="cloglog for Treatment")
legend("topleft",legend=c("Short Treatment","Long Treatment"),
pch = rep(15,4),col=2:3)
```



We see that the two lines are parallel and the distance between them remain the same throughout the graph. There is a cross between the two lines at the end, but this is minor. It is not an obvious cross. The proportional hazards model is reasonable.

```
summary(cox.1.trt)
```

```
## Call:
## coxph(formula = Surv(TIME2, CENSOR) ~ TREAT, data = relapseData)
##
##   n= 575, number of events= 464
##
##            coef exp(coef) se(coef)      z Pr(>|z|)
## TREAT1 -0.11884   0.88795  0.09295 -1.279    0.201
##
##        exp(coef) exp(-coef) lower .95 upper .95
## TREAT1     0.888      1.126    0.7401     1.065
##
## Concordance= 0.52  (se = 0.013 )
## Rsquare= 0.003   (max possible= 1 )
## Likelihood ratio test= 1.64  on 1 df,   p=0.2
## Wald test            = 1.63  on 1 df,   p=0.2
## Score (logrank) test = 1.64  on 1 df,   p=0.2
```

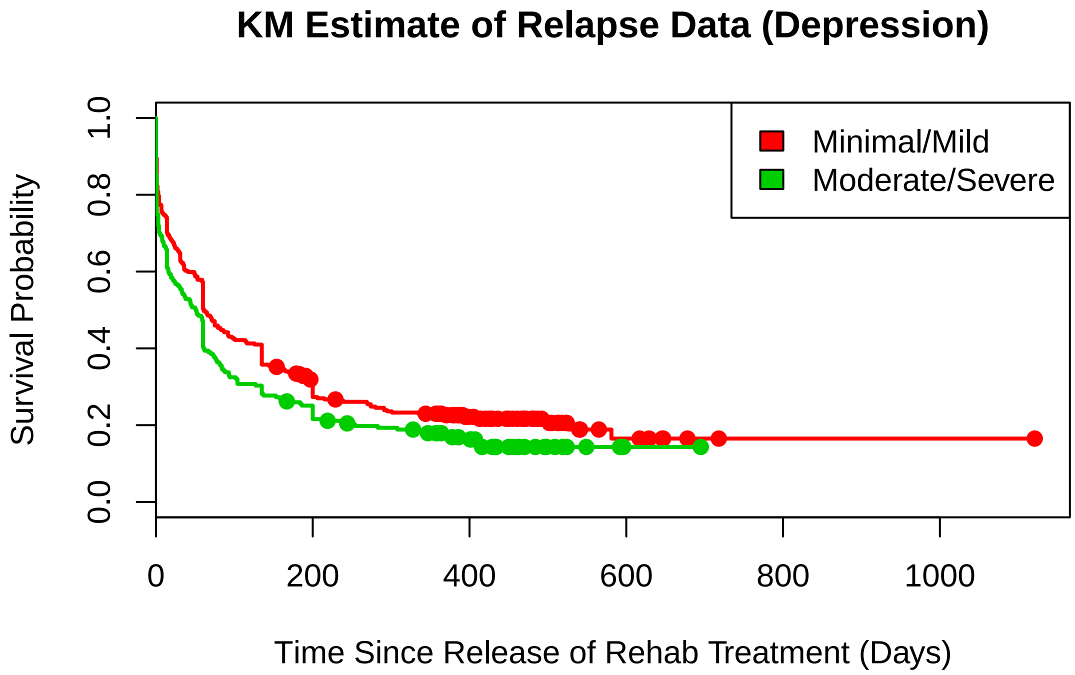
We can see that the likelihood ratio test, p=0.2, that the treatment length is not statistically significant.

Basic Estimation of Categorical Beck Depression

```
fit.2 <- survfit(Surv(TIME2, CENSOR)~dep, data = relapseData)

# Time,Censor ~ Depression
plot(fit.2, mark=19, lwd=2,
col=c(2,3), xlab="Time Since Release of Rehab Treatment (Days)", ylab="Survival Probability",
main="KM Estimate of Relapse Data (Depression)", mark.time = TRUE)

# Legend
legend("topright",c("Minimal/Mild", "Moderate/Severe"), fill=c(2,3))
```



Visually, we can see that people with Moderate/Severe deppression have a smaller survival probability throughoutthe entire study than people with Minimal/Mild Depression.

```
# Summarize above graph
print(fit.2)
```

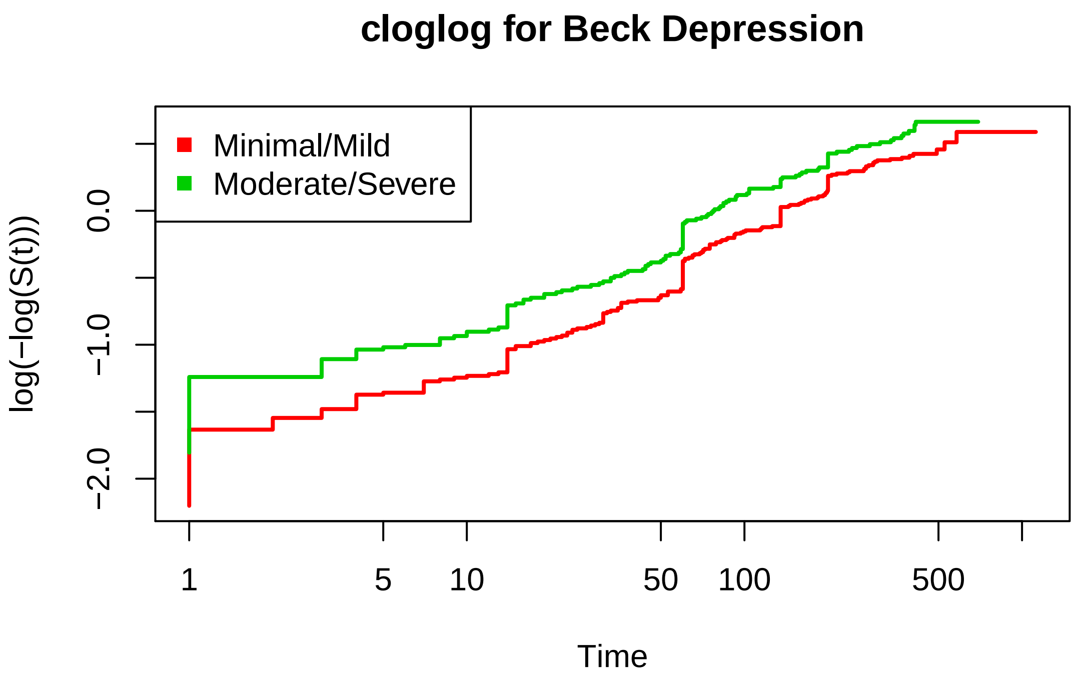
```
## Call: survfit(formula = Surv(TIME2, CENSOR) ~ dep, data = relapseData)
##
##                         n events median 0.95LCL 0.95UCL
## dep=Minimal/Mild      344    269     61      60      87
## dep=Moderate/Severe 231    195     51      30      60
```

Notice that the 50% of people with Moderate/Severe depression relapse (fail) within 51 days from the release of treatment. This is 10 days before people with Minimal/Mild depression.

Coxph for Categorical Beck Score (Discrete)

```
cox.1 <- coxph(Surv(TIME2, CENSOR)~dep, data=relapseData)

plot(fit.2,
fun="cloglog",col=c(2:3),xlab="Time",ylab="log(-log(S(t)))",lwd=2, main="cloglog for Beck Depression")
legend("topleft",legend=c("Minimal/Mild","Moderate/Severe"),
pch = rep(15,4),col=2:3)
```



By looking at the clog-log plot of the Beck Depression, we see that the two lines are parallel, do not cross, and do not diverge. This means that the proportional hazards model is reasonable.

```r
summary(cox.1)
```

```
## Call:
## coxph(formula = Surv(TIME2, CENSOR) ~ dep, data = relapseData)
##
##   n= 575, number of events= 464
##
##                       coef exp(coef) se(coef)     z Pr(>|z|)
## depModerate/Severe 0.22617   1.25379  0.09413 2.403   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## depModerate/Severe     1.254     0.7976     1.043     1.508
##
## Concordance= 0.533  (se = 0.013 )
## Rsquare= 0.01   (max possible= 1 )
## Likelihood ratio test= 5.69  on 1 df,   p=0.02
## Wald test            = 5.77  on 1 df,   p=0.02
## Score (logrank) test = 5.8  on 1 df,   p=0.02
```

We can see from the likelihood ratio test, p=0.02, that the depression categories are significant in the coxph model.

Model Fitting for Treatment

This data set has many variables and possible confounders I can test for, but I will list out a few that I think are important.

AGE: I will see if age should be in the model because age may or may not have an effect relapse time. The AGE variable is recorded at beginning of the study, so it does not change over time.

dep: Depression may have some correlation with time to relapse.

SITE: The sites have different methods of treating their patients and we should be aware of possible variation created by the sites.

IV: Intravenous drug use is often associated with more intense, addictive drugs, and can lead to a shorter survival time.

Possible methods for choosing a model include likelihood tests, AIC, and BIC. I will choose the method of AIC, because it is more interpretable for the general public than likelihood tests.

```
# Smallest Model
coxph(Surv(TIME2, CENSOR)~TREAT, data=relapseData)
```

```
## Call:
## coxph(formula = Surv(TIME2, CENSOR) ~ TREAT, data = relapseData)
##
##            coef exp(coef) se(coef)      z     p
## TREAT1 -0.11884   0.88795  0.09295 -1.279 0.201
##
## Likelihood ratio test=1.64  on 1 df, p=0.2009
## n= 575, number of events= 464
```

```
# Largest Model
coxph(Surv(TIME2, CENSOR)~AGE+dep+SITE+IV+TREAT, data=relapseData)
```

```
## Call:
## coxph(formula = Surv(TIME2, CENSOR) ~ AGE + dep + SITE + IV +
##     TREAT, data = relapseData)
##
##                         coef exp(coef)  se(coef)      z        p
## AGE                -0.020747  0.979466  0.007972 -2.602  0.00926
## depModerate/Severe  0.181948  1.199552  0.094646  1.922  0.05455
## SITE1               0.177769  1.194550  0.105341  1.688  0.09150
## IV                  0.253018  1.287907  0.056545  4.475 7.65e-06
## TREAT1             -0.079552  0.923530  0.093710 -0.849  0.39592
##
## Likelihood ratio test=29.52  on 5 df, p=1.833e-05
## n= 575, number of events= 464
```

We can see that the treatment variable, TREAT, is not statistically significant in the simplest model (likelihood ratio test p=0.2009) or the largest model (p=0.39592 for TREAT). We will now use the AIC method to find a model in between the smallest and largest.

```
aic.1a <- AIC(coxph(Surv(TIME2,CENSOR)~AGE+TREAT,data=relapseData))
aic.1b <- AIC(coxph(Surv(TIME2,CENSOR)~dep+TREAT,data=relapseData))
aic.1c <- AIC(coxph(Surv(TIME2,CENSOR)~SITE+TREAT,data=relapseData))
aic.1d <- AIC(coxph(Surv(TIME2,CENSOR)~IV+TREAT,data=relapseData))
cat(aic.1a, aic.1b, aic.1c, aic.1d)
```

```
## 5311.9 5308.247 5312.989 5300.347
```

It looks like $Survival \sim IV + TREAT$ has the lowest AIC score at 5300.347.

```
aic.2a <- AIC(coxph(Surv(TIME2,CENSOR)~AGE+IV+TREAT,data=relapseData))
aic.2b <- AIC(coxph(Surv(TIME2,CENSOR)~dep+IV+TREAT,data=relapseData))
aic.2c <- AIC(coxph(Surv(TIME2,CENSOR)~SITE+IV+TREAT,data=relapseData))
cat(aic.2a, aic.2b, aic.2c, "current model:", aic.1d)
```

```
## 5294.484 5297.733 5299.27 current model: 5300.347
```

We now update our model to $Survival \sim AGE + IV + TREAT$ because of the lower AIC 5294.484 < 5300.347 (current model).

```
aic.3a <- AIC(coxph(Surv(TIME2,CENSOR)~dep+AGE+IV+TREAT,data=relapseData))
aic.3b <- AIC(coxph(Surv(TIME2,CENSOR)~SITE+AGE+IV+TREAT,data=relapseData))
cat(aic.3a, aic.3b, "current model:", aic.2a)
```

```
## 5292.7 5293.568 current model: 5294.484
```

We will update our model to $Survival \sim dep + AGE + IV + TREAT$ because of the lower AIC 5292.7 < 5294.484 (current model).

```
aic.4a <- AIC(coxph(Surv(TIME2,CENSOR)~SITE+dep+AGE+IV+TREAT,data=relapseData))
cat(aic.4a, "current model:", aic.3a)
```

```
## 5291.912 current model: 5292.7
```

It looks like our largest model has the lowest AIC at 5291.912. Our final model will be $Survival \sim SITE + dep + AGE + IV + TREAT$

```
fit.trt <- coxph(Surv(TIME2,CENSOR)~SITE+dep+AGE+IV+TREAT,data=relapseData)
anova(fit.trt)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(TIME2, CENSOR)
## Terms added sequentially (first to last)
##
##        loglik   Chisq Df Pr(>|Chi|)
## NULL  -2655.7
## SITE  -2655.2  1.0103  1    0.31483
## dep   -2652.4  5.6969  1    0.01699 *
## AGE   -2651.7  1.2809  1    0.25773
## IV    -2641.3 20.8105  1  5.071e-06 ***
## TREAT -2641.0  0.7210  1    0.39580
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After controlling for SITE, depression, AGE, and IV use, the treatment length is still not significant at p=0.39580.

At this point, we have our model, but before I move forward with the analysis, I will check if the proportional hazard model is reasonable.

Diagnostics

Goodness of Fit Test

```
zp <- cox.zph(fit.trt, transform = "rank")
zp
```

```
##                        rho  chisq     p
## SITE1               -0.0595 1.7134 0.191
## depModerate/Severe -0.0420 0.8124 0.367
## AGE                 0.0110 0.0514 0.821
## IV                 -0.0237 0.2547 0.614
## TREAT1              0.0403 0.7701 0.380
## GLOBAL                  NA 3.8759 0.567
```

The proportional hazards model is appropriate according to the Goodness of Fit Test. Every variable has a p-value above $\alpha = 0.05$.

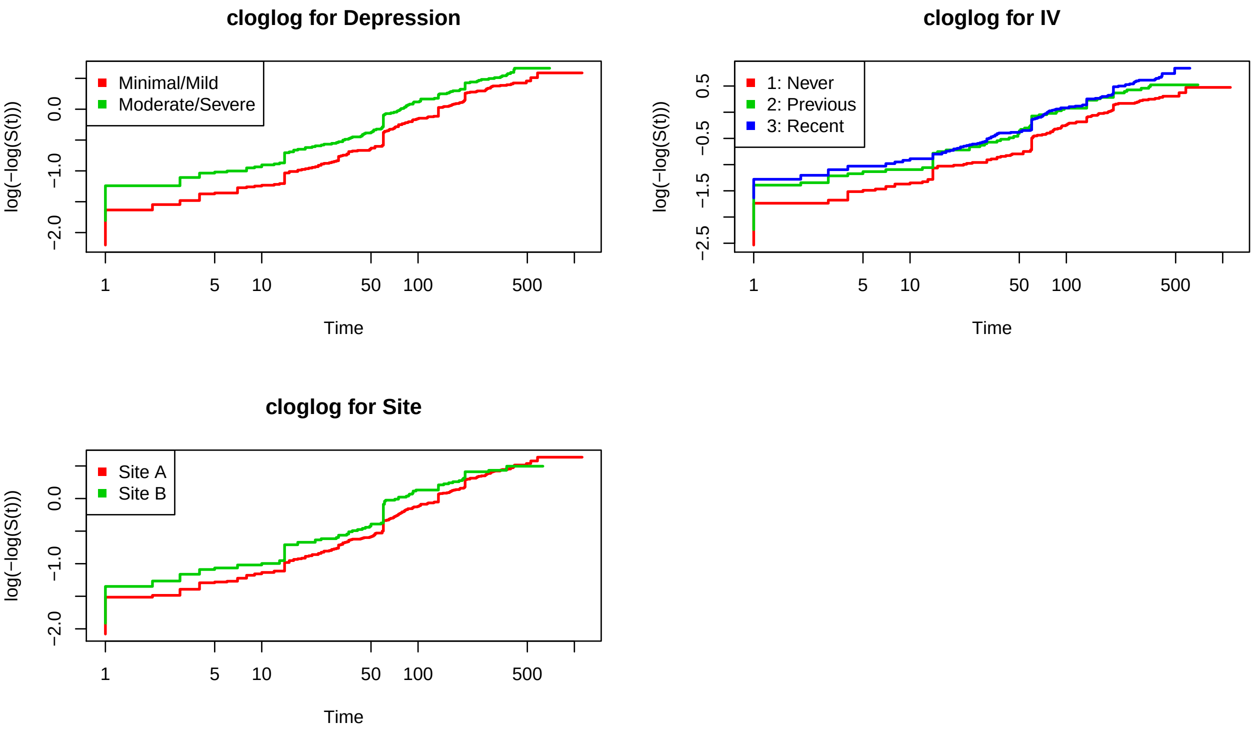
Complementary Log-Log Plot

```
par(mfrow=c(2,2))

plot(survfit(Surv(TIME2,CENSOR)~dep,data=relapseData),
fun="cloglog",col=c(2:3),xlab="Time",ylab="log(-log(S(t)))",lwd=2, main="cloglog for Depression")
legend("topleft",legend=c("Minimal/Mild","Moderate/Severe"),
pch = rep(15,4),col=2:4)

plot(survfit(Surv(TIME2,CENSOR)~IV,data=relapseData),
fun="cloglog",col=c(2:4),xlab="Time",ylab="log(-log(S(t)))",lwd=2, main="cloglog for IV")
legend("topleft",legend=c("1: Never","2: Previous","3: Recent"),
pch = rep(15,4),col=2:4)

plot(survfit(Surv(TIME2,CENSOR)~SITE,data=relapseData),
fun="cloglog",col=c(2:3),xlab="Time",ylab="log(-log(S(t)))",lwd=2, main="cloglog for Site")
legend("topleft",legend=c("Site A","Site B"),
pch = rep(15,4),col=2:4)
```



The clog-log plot for IV, Depression, and Site are plotted above. The lines are parallel for Depression and IV, but the lines cross in the Complementary Log Log Plot for Site. This means that the proportional hazards assumption is not satisfied. I will stratify the SITE variable. Our new, updated model is $Survival \sim Strata(SITE) + +dep + AGE + IV + TREAT$.

Because we are stratifying the SITE variable, the data will be split and we will have different baselines for each Site. They will share the same $\hat{\beta}$, so our hazard ratios will be the same.

```
# Main Analysis of Treatment Variable
fit.trt.new <- coxph(Surv(TIME2,CENSOR)~strata(SITE)+dep+AGE+IV+TREAT,data=relapseData)
summary(fit.trt.new)
```

```
## Call:
## coxph(formula = Surv(TIME2, CENSOR) ~ strata(SITE) + dep + AGE +
##     IV + TREAT, data = relapseData)
##
##   n= 575, number of events= 464
##
##                          coef exp(coef)  se(coef)      z Pr(>|z|)
## depModerate/Severe  0.183163  1.201011  0.094650  1.935  0.05297 .
## AGE                -0.021078  0.979142  0.007982 -2.641  0.00828 **
## IV                  0.252134  1.286769  0.056431  4.468  7.9e-06 ***
## TREAT1             -0.087364  0.916344  0.093707 -0.932  0.35118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## depModerate/Severe    1.2010     0.8326    0.9977    1.4458
## AGE                   0.9791     1.0213    0.9639    0.9946
## IV                    1.2868     0.7771    1.1520    1.4373
## TREAT1                0.9163     1.0913    0.7626    1.1011
##
## Concordance= 0.584  (se = 0.016 )
## Rsquare= 0.049   (max possible= 1 )
## Likelihood ratio test= 28.8  on 4 df,   p=9e-06
## Wald test            = 28.8  on 4 df,   p=9e-06
## Score (logrank) test = 28.98  on 4 df,   p=8e-06
```

H_0 : The short and long treatments are the same.

First, note that the treatment is still not significant (p-value=0.35118 for TREAT). This means that we accept the null that the short and long treatments are the same.

Our β estimation can be seen under `coef`, which is -0.087364. Because this is negative, we know that long treatment (TREAT=1) has a lower hazard rate than short treatment (baseline). Specifically, The hazard ratio for long to short treatment is $e^{\beta} = 0.916344$ (found under `exp(coef)`). This means that a patient who had a long treatment has a hazard rate 8.4% lower than a patient who had a short treatment. In other words, they appear to have a longer survival time and relapse more slowly. Again, remember that this difference is not statistically significant because we accepted the null hypothesis that the short and long treatments are the same and do not have an effect on relapse time.

Confidence Interval for Hazard Probability Ratio

Confidence Interval: $[e^{coef-1.96*s.e.}, e^{coef+1.96*s.e.}]$

```
c(exp(-0.087364 - 1.96 * 0.093707), exp(-0.087364 + 1.96 * 0.093707))
```

```
## [1] 0.7625938 1.1010912
```

The 95% confidence interval for the Hazard Ratio of $\frac{Short.Trt}{Long.Trt}$ is [0.7625938, 1.1010912]. This interval contains 1.00, so this confirms on conclusion that the hazard rate for treatment lengths are the same.

Extension: Parametric Model Comparison (Exponential vs. Weibull)

For this next part, I will create a weibull model and see if it is a better fit for the data. The parametric models do not work when someone fails at time 0, so I added 0.1 to people who have failed on the first day they were released.

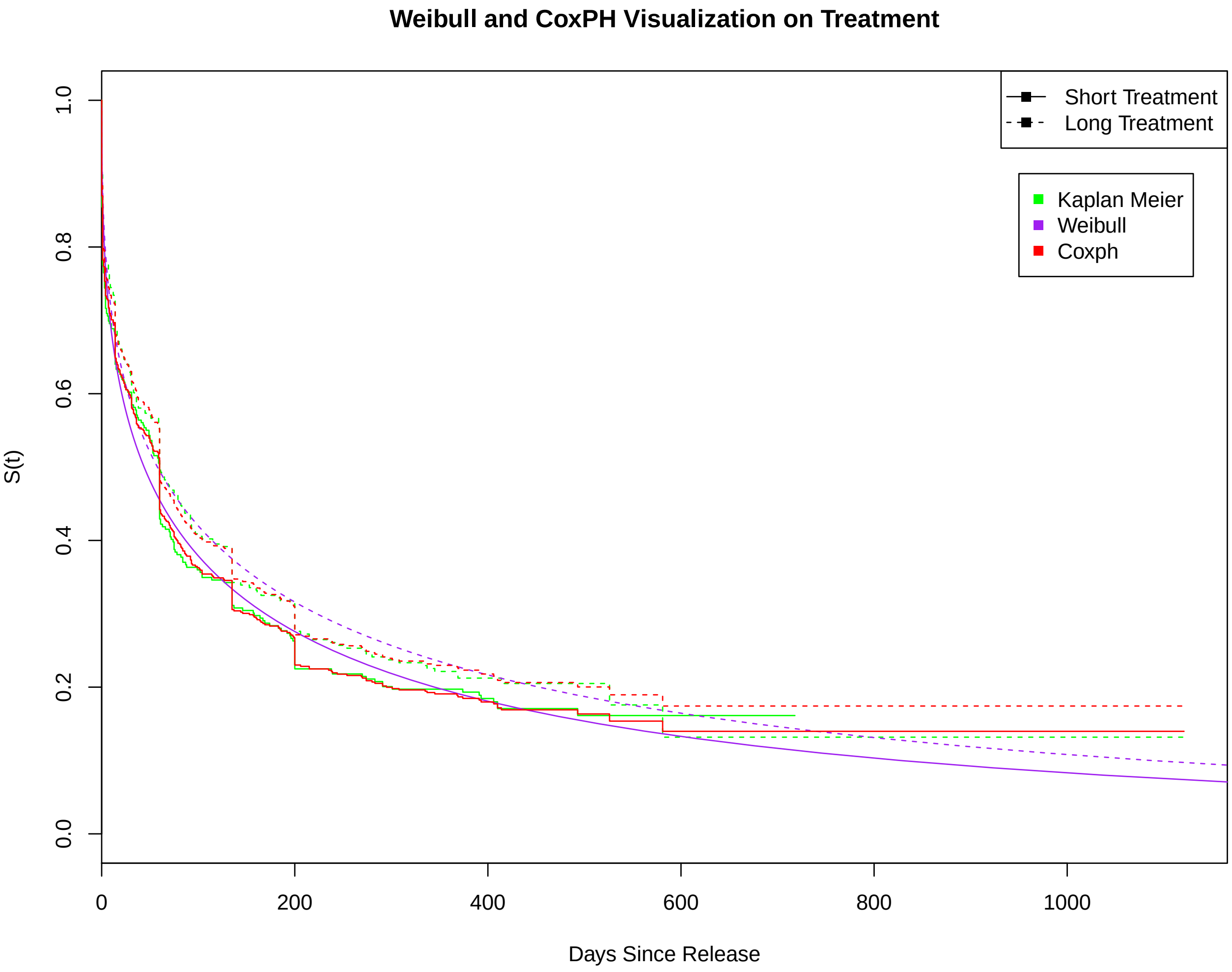
Advantage of Weibull Model: It is great for visualizations because of its smooth curve. Furthermore, since the entire survival function is parameterized, we can extrapolate survival probabilities at whatever time. We are not restricted to the domain given by the time data variable.

```
newdata <- relapseData
newdata$TIME2 <- ifelse(relapseData$TIME2==0, relapseData$TIME2+0.1, relapseData$TIME2)
```

For simplicity and visualization, I will plot the model on the simplest model of $Survival \sim TREAT$ and not include the other variables.

```
fit.wei <- survreg(Surv(TIME2,CENSOR)~TREAT, data = newdata, dist = "weibull")
```

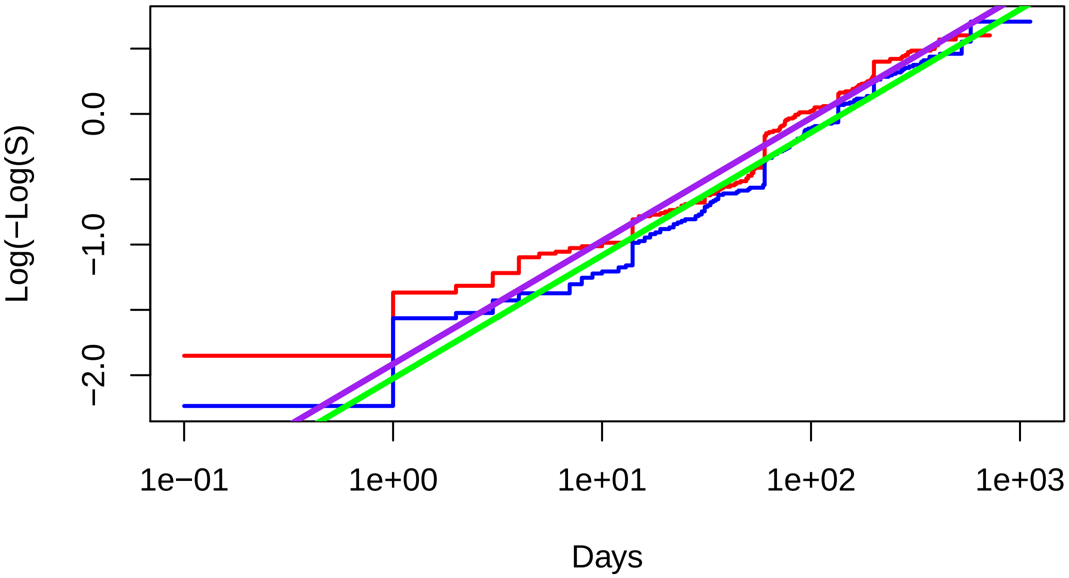
```
ps <- seq(0.01,0.99,by=0.01)
plot(survfit(Surv(TIME2,CENSOR)~TREAT,data=relapseData),lty=1:2,col="green",xlab="Days Since Release",yl
lines(predict(fit.wei,data.frame(TREAT=newdata$TREAT[4]),type="quantile",p=ps),1-ps,col="purple",lty=1)
lines(predict(fit.wei,data.frame(TREAT=newdata$TREAT[1]),type="quantile",p=ps),1-ps,col="purple",lty=2)
lines(survfit(cox.1.trt, newdata = data.frame(TREAT = 0:1)), col = "red", lty = 1:2, mark.time = FALSE)
legend("topright",c("Short Treatment","Long Treatment"),pch = 15,lty=1:2)
legend(950,0.9,c("Kaplan Meier","Weibull","Coxph"),col = c("green","purple","red"),pch = 15)
```



Just visually, we can see that the Weibull model fits the Kaplan Meier estimates very well, but not as well as the coxph.

Clog-log of Weibull

```
ps <- seq(0.01,0.99,by=0.01)
plot(survfit(Surv(TIME2,CENSOR)~TREAT,data=newdata),fun="cloglog",lwd=2,col=c(2,4),xlab="Days",ylab="Log
lines(predict(fit.wei,data.frame(TREAT=newdata$TREAT[4]),type="quantile",p=ps),log(-log(1-ps)),lwd=3,col
lines(predict(fit.wei,data.frame(TREAT=newdata$TREAT[4]),type="quantile",p=ps),log(-log(1-ps)),lwd=3,col
lines(predict(fit.wei,data.frame(TREAT=newdata$TREAT[1]),type="quantile",p=ps),log(-log(1-ps)),lwd=3,col
```



The Complementary Log-Log plot of the weibull works very well. It closely aligns with the clog-log plot from the regular fit.

Comparison of Weibull vs Coxph (Simple Models)

```
# Set Up Predictions
wei.predict.short <- data.frame(x=predict(fit.wei,data.frame(TREAT=newdata$TREAT[4]),type="quantile",p=p
wei.predict.long <- data.frame(x=predict(fit.wei,data.frame(TREAT=newdata$TREAT[1]),type="quantile",p=ps
cox.predict <- survfit(cox.1.trt, newdata = data.frame(TREAT = 0:1))

# Prediction: Survival Probability after 1 Year (Coxph, Short Treatment)
min(cox.predict[1]$surv[cox.predict$time <= 365])
```

```
## [1] 0.1908036
```

```
# Prediction: Survival Probability after 1 Year (Coxph, Long Treatment)
min(cox.predict[2]$surv[cox.predict$time <= 365])
```

```
## [1] 0.2297182
```

```
# Prediction: Survival Probability after 1 Year (Weibull, Short Treatment)
min(wei.predict.short$y[wei.predict.short$x <= 365])
```

```
## [1] 0.2
```

```
# Prediction: Survival Probability after 1 Year (Weibull, Long Treatment)
min(wei.predict.long$y[wei.predict.long$x <= 365])
```

```
## [1] 0.23
```

We can see that the predictions are very close. For Short Treatment patients, the survival probability after 1 year is 0.1908036 for Coxph and 0.2 for Weibull. For Long Treatment Patients, the survival probability after 1 year is 0.2297182 for Coxph and 0.23 for Weibull. These predictions are very close. Unlike the coxph, and kaplan meier estimates, we can find the survival probability and any point. Let's check the survival probability after 5 years.

```r
# Prediction: Survival Probability after 5 Years (Weibull, Long Treatment)
min(wei.predict.short$y[wei.predict.short$x <= 1825])
```

```
## [1] 0.05
```

```r
# Prediction: Survival Probability after 5 Years (Weibull, Long Treatment)
min(wei.predict.long$y[wei.predict.long$x <= 1825])
```

```
## [1] 0.06
```

The chances of surving (not relapsing) past 5 years is 5% for the short treatment group and 6% for the long treatment group. This is very low. This is a prediction that we got using the Weibull method.

```r
# Weibull Full Model
fit.wei.new <- survreg(Surv(TIME2,CENSOR)~strata(SITE)+dep+AGE+IV+TREAT, data = newdata, dist = "weibul
summary(fit.wei.new)
```

```
##
## Call:
## survreg(formula = Surv(TIME2, CENSOR) ~ strata(SITE) + dep +
##     AGE + IV + TREAT, data = newdata, dist = "weibull")
##                        Value Std. Error     z       p
## (Intercept)           4.3227     0.6303  6.86 6.9e-12
## depModerate/Severe -0.4560     0.2270 -2.01  0.0446
## AGE                   0.0529     0.0194  2.73  0.0063
## IV                   -0.5663     0.1318 -4.30 1.7e-05
## TREAT1                0.2373     0.2245  1.06  0.2905
## 0                     0.8513     0.0472 18.04 < 2e-16
## 1                     0.9405     0.0743 12.66 < 2e-16
##
## Scale:
##    0    1
## 2.34 2.56
##
## Weibull distribution
## Loglik(model)= -2476.8   Loglik(intercept only)= -2490.8
##  Chisq= 28.03 on 4 degrees of freedom, p= 1.2e-05
## Number of Newton-Raphson Iterations: 5
## n= 575
```

According to the full Weibull Model, we can see that having a Long Treatment (TREAT=1) makes relapse time longer by 26.78%. We know this by taking the number under `Value`, 0.2373, and exponentiated it: $e^{0.2373=1.2678}$. Note that this is not significant due to the p-value=0.2905.

Time-Varying Variable

The time varying covariate is the time spent in treatment. I will split the data to take into account the time spent in treatment. I will split the data into two episodes: time of treatment (episode 1) and time after treatment to relapse or censor (episode 2). We are doing this because there may be different hazard rates when someone is in treatment vs. when they are released.

```r
# Split Data Manually
newdata2 <- relapseData
relapseData$epi <- 2
newdata2$epi <- 1
```

```r
newdata2$epi <- as.factor(1 *(newdata2$epi == "1") + 2*(newdata2$epi=="2"))
newdata2$LEN.T <- 0
newdata2$CENSOR <- 0
newdata2$TIME <- relapseData$LEN.T

# New Counting Process Dataset
cpmdata <- rbind(relapseData, newdata2)

# Counting Process Model
fit.cpm <- coxph(Surv(LEN.T,TIME,CENSOR)~TREAT+TREAT:epi, data = cpmdata)
fit.cpm
```

```
## Call:
## coxph(formula = Surv(LEN.T, TIME, CENSOR) ~ TREAT + TREAT:epi,
##     data = cpmdata)
##
##                  coef exp(coef)  se(coef)     z     p
## TREAT1      5.627e-01 1.755e+00 3.116e+03 0.000 1.000
## TREAT0:epi2 2.179e+01 2.903e+09 2.553e+03 0.009 0.993
## TREAT1:epi2 2.132e+01 1.807e+09 1.787e+03 0.012 0.990
##
## Likelihood ratio test=586.3  on 3 df, p=< 2.2e-16
## n= 1079, number of events= 393
##    (71 observations deleted due to missingness)
```

H_0 : The treatment lengths are the same.

We can see that after splitting the data, the treatment lengths still do not have a significant effect during and after treatment (p-values are 0.993 and 0.990). We accept the null. Based on this conclusion and all the conclusions above, treatment length does not have a significant affect on relapse time and so people should opt for the shorter treatment length, especially if it is more cost efficient.

Reference Section

Source

Table 1.3 of Hosmer,D.W. and Lemeshow, S. (1998)

References

Hosmer,D.W. and Lemeshow, S. (1998) Applied Survival Analysis: Regression Modeling of Time to Event Data, John Wiley and Sons Inc., New York, NY