PSTAT-175-Final-Project

William Nelson

2025-05-29

You want to write up your project as a report. It should look like something you can read with sentences and paragraphs. It should not be a series of answers to my questions like the homework assignment. You should have sections that are numbered, but those sections should relate to the analysis that you are doing in your model

Final Project

1: Introduction

We will be analyzing the "GRACE1000" dataset from Hosmer and Lemenshow (2008), which contains data on 1000 patients who were part of a study on revascularization. Our failure time of interest is the follow up time when the researchers checked-in with the patients, and our event of interest is whether the patient died during the follow-up period. Our main covariate of interest will be whether the patient had the revascularization procedure performed on them, which is coded by the "revasc" variable with value 1 if the patient had the procedure and 0 if they did not.

1.1: Citations

2: Model Fitting

```r
library(readr)
library(tidyverse)
library(survival)
library(survminer)
library(ggplot2)
GRACE1000 <- read_table("C:/Users/willi/OneDrive/Documents/GitHub/PSTAT175-Final/GRACE1000.dat", col_nar
GRACE1000 <- GRACE1000 %>% select(-X10)
colnames(GRACE1000) <- c("id", "days", "death", "revasc", "revascdays", "los", "age", "sysbp", "stchange
```

```r
# Forward Stepwise Selection
# Possible covariates are revasc, revascdays, sysbp, age, stchange, los
# Included revascdays despite it being equal to days if no revasc happened

revdays.mod <- coxph(Surv(days,death) ~ revascdays, data = GRACE1000)
age.mod <- coxph(Surv(days,death) ~ age, data = GRACE1000)
sysbp.mod <- coxph(Surv(days,death) ~ sysbp, data = GRACE1000)
stchange.mod <- coxph(Surv(days,death) ~ stchange, data = GRACE1000)
los.mod <- coxph(Surv(days,death) ~ los, data = GRACE1000)
```

```r
revasc.mod <- coxph(Surv(days,death) ~ revasc, data = GRACE1000)

AIC(revdays.mod, age.mod, sysbp.mod, stchange.mod, los.mod, revasc.mod)
```

```
##              df      AIC
## revdays.mod   1 4191.468
## age.mod       1 4151.309
## sysbp.mod     1 4245.999
## stchange.mod  1 4252.286
## los.mod       1 4268.978
## revasc.mod    1 4231.999
```

```r
BIC(revdays.mod, age.mod, sysbp.mod, stchange.mod, los.mod, revasc.mod)
```

```
##              df      BIC
## revdays.mod   1 4195.249
## age.mod       1 4155.090
## sysbp.mod     1 4249.780
## stchange.mod  1 4256.067
## los.mod       1 4272.759
## revasc.mod    1 4235.780
```

```r
# Lowest AIC (4151.309) and BIC (4155.090) is age
```

```r
age.revdays.mod <- coxph(Surv(days,death) ~ age + revascdays, data = GRACE1000)
age.sysbp.mod <- coxph(Surv(days,death) ~ age + sysbp, data = GRACE1000)
age.stchange.mod <- coxph(Surv(days,death) ~ age + stchange, data = GRACE1000)
age.los.mod <- coxph(Surv(days,death) ~ age + los, data = GRACE1000)
age.revasc.mod <- coxph(Surv(days,death) ~ age + revasc, data = GRACE1000)

AIC(age.revdays.mod, age.sysbp.mod, age.stchange.mod, age.los.mod, age.revasc.mod)
```

```
##                  df      AIC
## age.revdays.mod   2 4043.699
## age.sysbp.mod     2 4131.179
## age.stchange.mod  2 4136.352
## age.los.mod       2 4146.685
## age.revasc.mod    2 4136.017
```

```r
BIC(age.revdays.mod, age.sysbp.mod, age.stchange.mod, age.los.mod, age.revasc.mod)
```

```
##                  df      BIC
## age.revdays.mod   2 4051.261
## age.sysbp.mod     2 4138.741
## age.stchange.mod  2 4143.913
## age.los.mod       2 4154.247
## age.revasc.mod    2 4143.579
```

```
# Lowest AIC (4043.699) and BIC (4051.261) is revascdays + age
```

```
revdays.age.sysbp.mod <- coxph(Surv(days,death) ~ revascdays + age + sysbp, data = GRACE1000)
revdays.age.stchange.mod <- coxph(Surv(days,death) ~ revascdays + age + stchange, data = GRACE1000)
revdays.age.los.mod <- coxph(Surv(days,death) ~ revascdays + age + los, data = GRACE1000)
revdays.age.revasc.mod <- coxph(Surv(days,death) ~ revascdays + age + revasc, data = GRACE1000)

AIC(revdays.age.sysbp.mod, revdays.age.stchange.mod, revdays.age.los.mod, revdays.age.revasc.mod)
```

```
##                          df      AIC
## revdays.age.sysbp.mod     3 4026.107
## revdays.age.stchange.mod  3 4036.515
## revdays.age.los.mod       3 4020.497
## revdays.age.revasc.mod    3 3701.919
```

```
BIC(revdays.age.sysbp.mod, revdays.age.stchange.mod, revdays.age.los.mod, revdays.age.revasc.mod)
```

```
##                          df      BIC
## revdays.age.sysbp.mod     3 4037.450
## revdays.age.stchange.mod  3 4047.857
## revdays.age.los.mod       3 4031.839
## revdays.age.revasc.mod    3 3713.261
```

```
# Lowest AIC (3701.919) and BIC (3713.261) is revasc + revascdays + age
# drops so much because revasc + revascdays together means if revasc = 0, it predicts
# the days outcome as 180 perfectly
```

```
revasc.revdays.age.sysbp.mod <- coxph(Surv(days,death) ~ revasc + revascdays + age + sysbp, data = GRAC
revasc.revdays.age.stchange.mod <- coxph(Surv(days,death) ~ revasc + revascdays + age + stchange, data =
revasc.revdays.age.los.mod <- coxph(Surv(days,death) ~ revasc + revascdays + age + los, data = GRACE100

AIC(revasc.revdays.age.sysbp.mod, revasc.revdays.age.stchange.mod, revasc.revdays.age.los.mod)
```

```
##                                 df      AIC
## revasc.revdays.age.sysbp.mod     4 3697.444
## revasc.revdays.age.stchange.mod  4 3700.495
## revasc.revdays.age.los.mod       4 3697.495
```

```
BIC(revasc.revdays.age.sysbp.mod, revasc.revdays.age.stchange.mod, revasc.revdays.age.los.mod)
```

```
##                                 df      BIC
## revasc.revdays.age.sysbp.mod     4 3712.567
## revasc.revdays.age.stchange.mod  4 3715.618
## revasc.revdays.age.los.mod       4 3712.618
```

```
# Lowest AIC (3697.444) and BIC (3712.567) is sysbp + revasc + revascdays + age
# very close to los with AIC of 3697.495 and BIC of 3712.618
```

```
sysbp.revasc.revdays.age.stchange.mod <- coxph(Surv(days,death) ~ sysbp + revasc + revascdays + age + st
sysbp.revasc.revdays.age.los.mod <- coxph(Surv(days,death) ~ sysbp + revasc + revascdays + age + los, da

AIC(sysbp.revasc.revdays.age.stchange.mod, sysbp.revasc.revdays.age.los.mod)
```

```
##                                       df      AIC
## sysbp.revasc.revdays.age.stchange.mod  5 3696.180
## sysbp.revasc.revdays.age.los.mod       5 3693.021
```

```
BIC(sysbp.revasc.revdays.age.stchange.mod, sysbp.revasc.revdays.age.los.mod)
```

```
##                                       df      BIC
## sysbp.revasc.revdays.age.stchange.mod  5 3715.083
## sysbp.revasc.revdays.age.los.mod       5 3711.925
```

```
# Lowest AIC (3696.180) and BIC (3711.925) is los + sysbp + revasc + revascdays + age
```

```
# See if stchange lowers AIC and BIC
all.mod <- coxph(Surv(days,death) ~ stchange + sysbp + los + revasc + revascdays + age, data = GRACE1000
AIC(all.mod) # 3691.309 lowers
```

```
## [1] 3691.309
```

```
BIC(all.mod) # 3713.994 raises, stchange not needed by BIC criterion
```

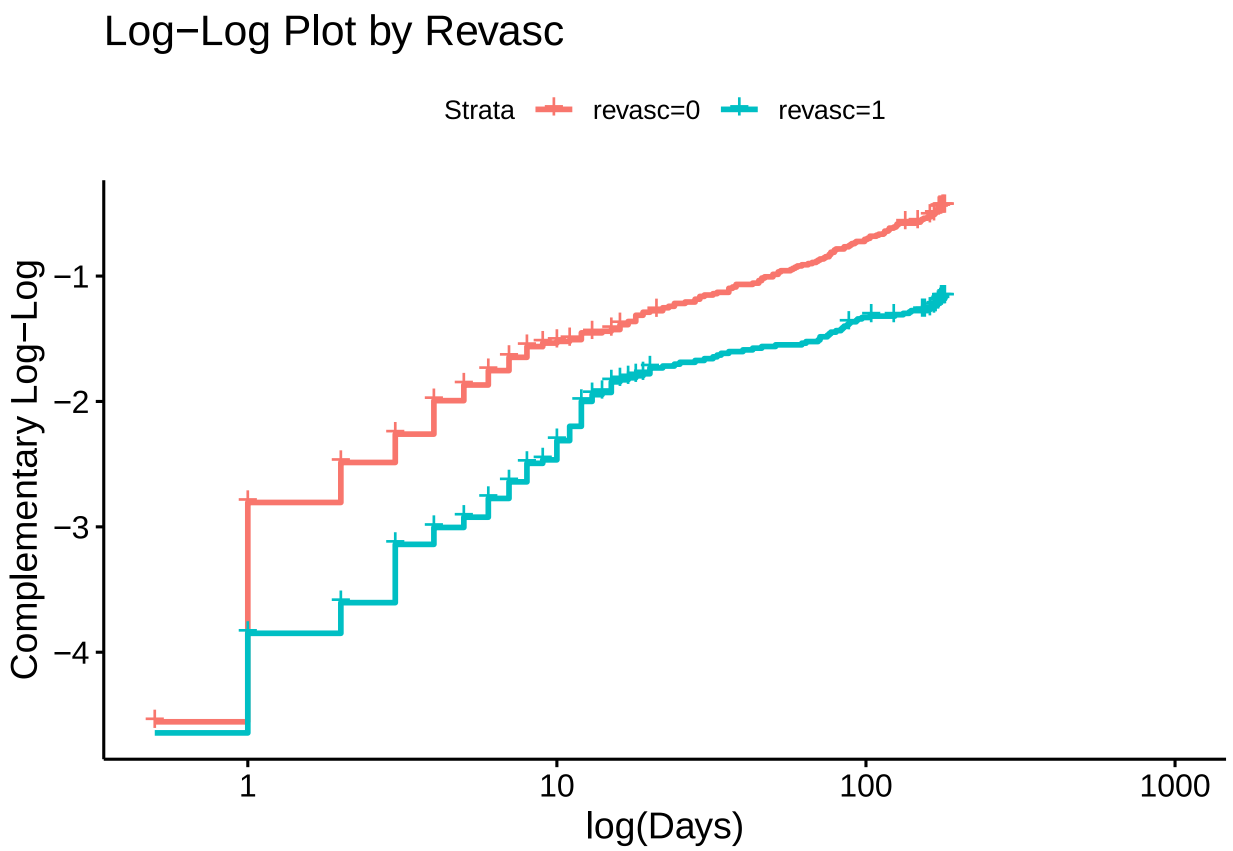
```
## [1] 3713.994
```

```
# According to forward stepwise selection by AIC, the full model is the best model
# According to BIC, the model with all covariates except stchange is the best model
# Maybe go with BIC to not have all covariates?
```

3: Check Proportional Hazards Assumptions

Log-Log plot for revasc:

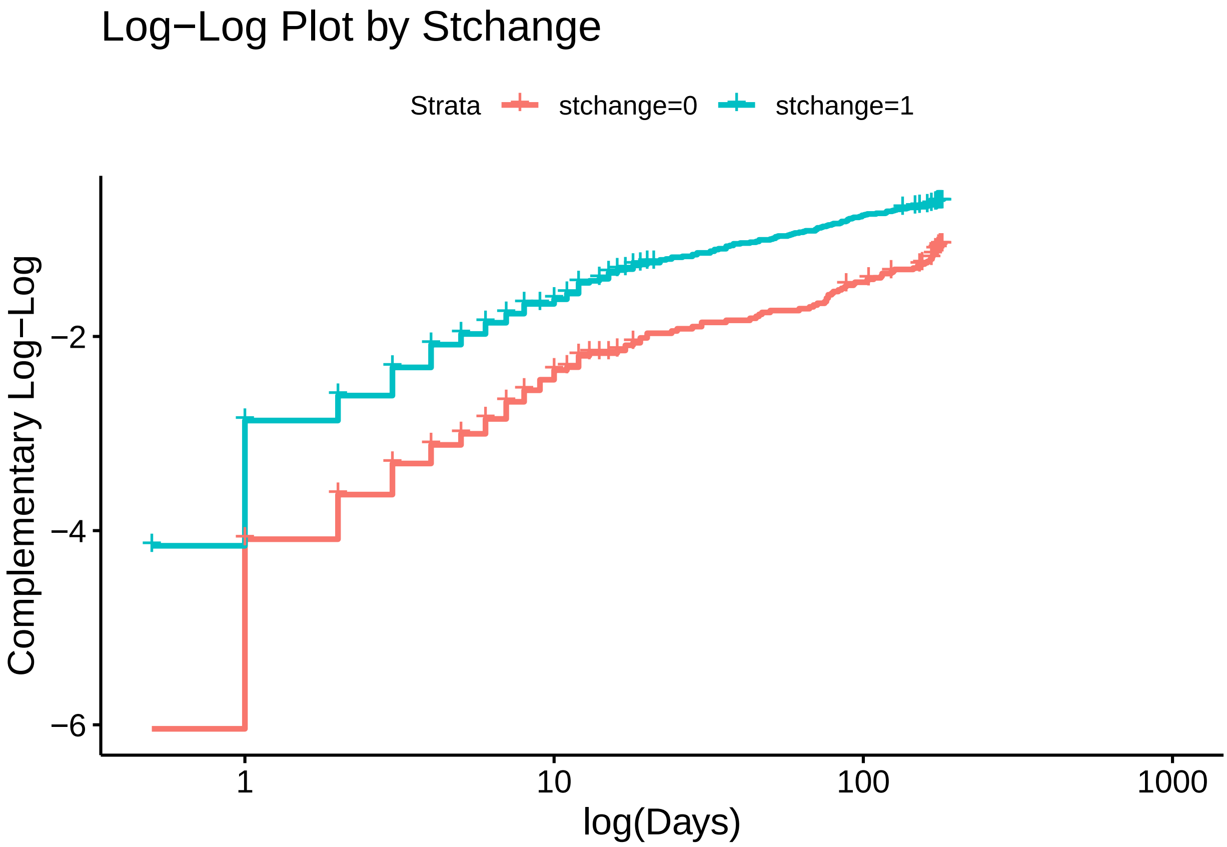
```
ggsurvplot(survfit(Surv(days,death) ~ revasc, data = GRACE1000),
           fun = "cloglog") +
  labs(x = "log(Days)", y = "Complementary Log-Log",
       title = "Log-Log Plot by Revasc")
```



```
# Seems very parallel, no assumptions violated
```

Log-Log plot for stchange? might not include because BIC says not to

```
ggsurvplot(survfit(Surv(days,death) ~ stchange, data = GRACE1000),
           fun = "cloglog") +
  labs(x = "log(Days)", y = "Complementary Log-Log",
       title = "Log-Log Plot by Stchange")
```



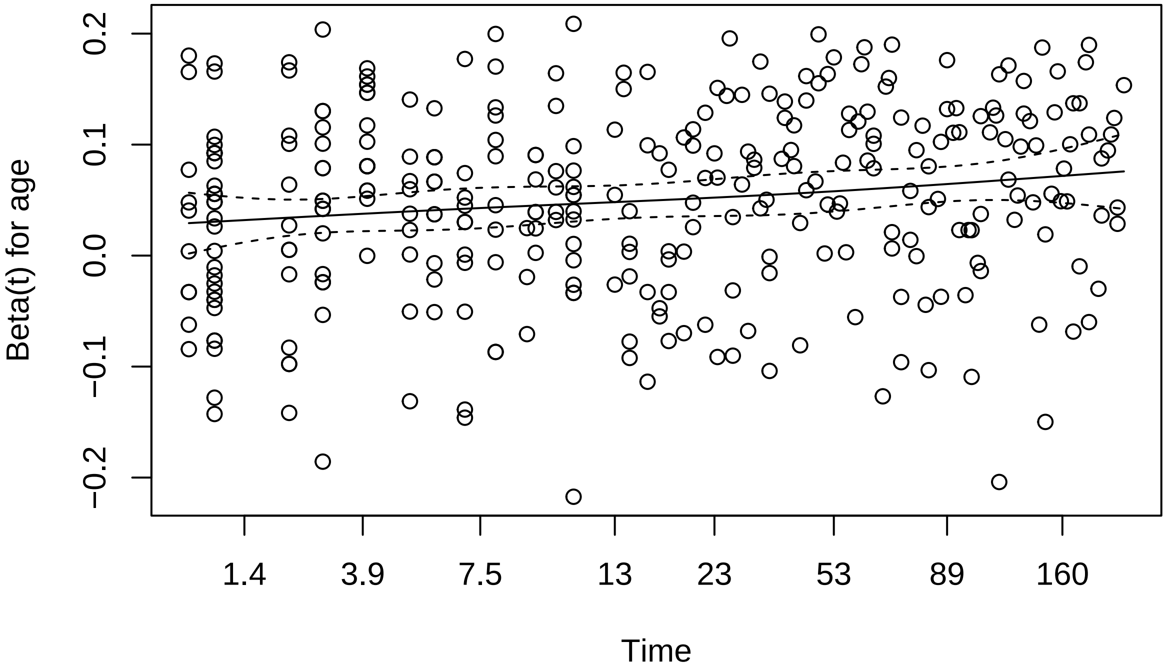
```
# also very parallel lines, no assumptions violated
```

ZPH plot for age:

```
cox.zph(age.mod)
```

```
##          chisq df      p
## age       7.76  1 0.0053
## GLOBAL    7.76  1 0.0053
```

```
plot(cox.zph(age.mod))
```

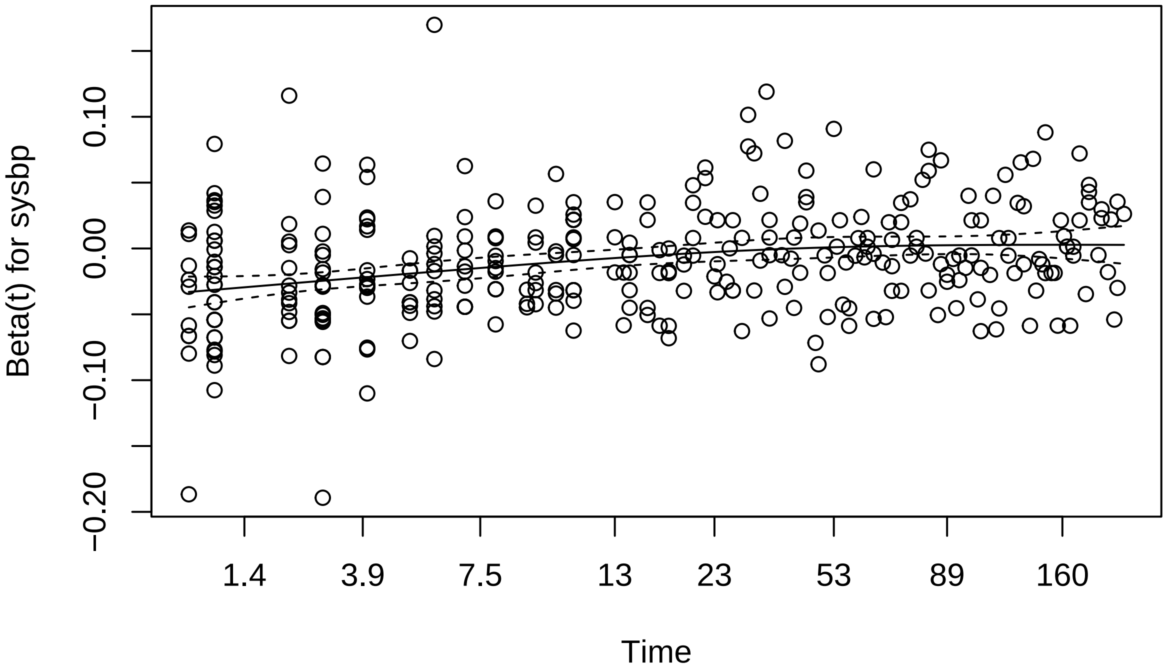


```
# has significant p-value but plot does
# not appear to have strong trend
```

```
cox.zph(sysbp.mod)
```

```
##        chisq df       p
## sysbp   33.2  1 8.4e-09
## GLOBAL  33.2  1 8.4e-09
```

```
plot(cox.zph(sysbp.mod))
```

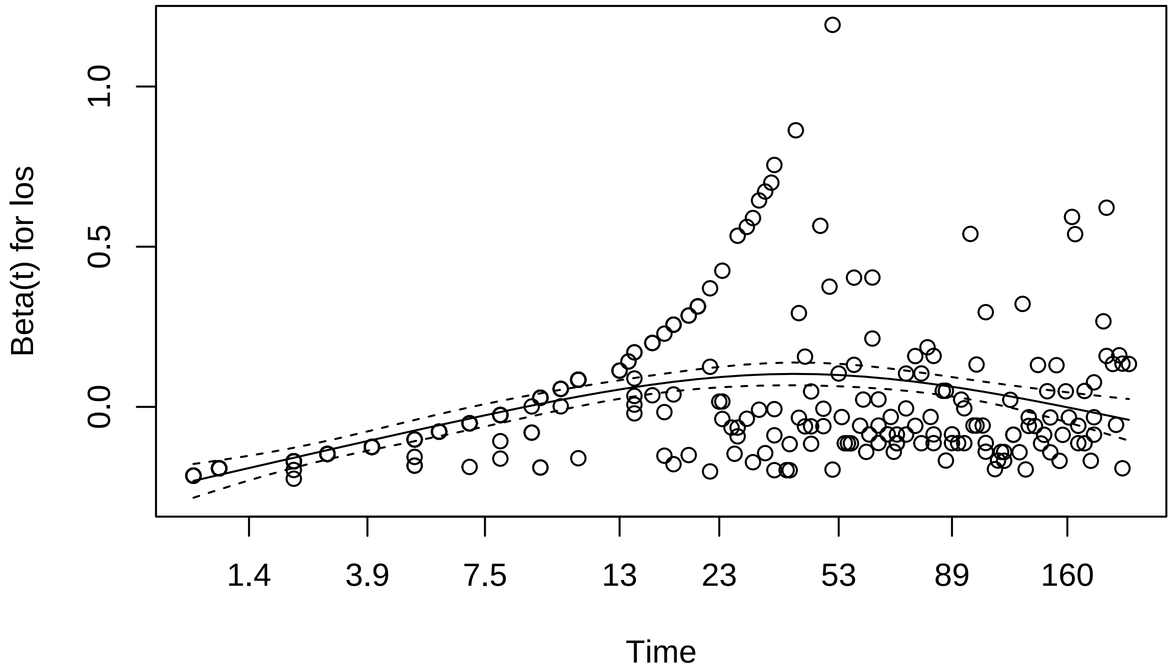


```
# The zph test gives a significant p-value which means sysbp violates
# proportional hazards assumption due to time dependency
# but plot does not seem to have a clear correlation
```

```
cox.zph(los.mod)
```

```
##         chisq df      p
## los      82.7  1 <2e-16
## GLOBAL   82.7  1 <2e-16
```

```
plot(cox.zph(los.mod))
```

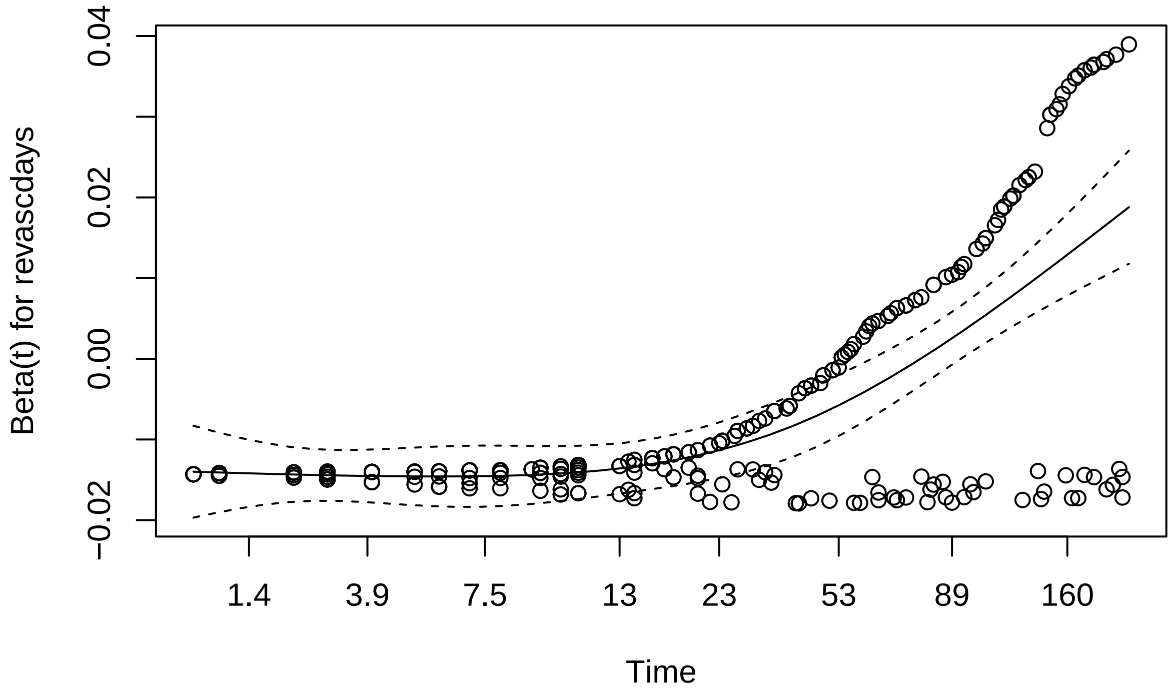


```
# time variable so makes sense it has very small p-value
```

```
cox.zph(revdays.mod)
```

```
##              chisq df       p
## revascdays  71.7  1 <2e-16
## GLOBAL      71.7  1 <2e-16
```

```
plot(cox.zph(revdays.mod))
```



```
# Definitely violates ph assumptions due to way it is coded
```

4: Conclusions

Hazard Ratios

95% confidence intervals for hazard ratios

Our main scientific question of interest is whether the revascularization procedure significantly increases patients' survival probabilities

```
# Revasc:

summary(revasc.mod)
```

```
## Call:
## coxph(formula = Surv(days, death) ~ revasc, data = GRACE1000)
##
##   n= 1000, number of events= 324
##
##            coef exp(coef) se(coef)      z Pr(>|z|)
## revasc -0.7149    0.4892   0.1144 -6.249 4.13e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##        exp(coef) exp(-coef) lower .95 upper .95
## revasc    0.4892      2.044     0.391    0.6122
##
## Concordance= 0.589  (se = 0.014 )
## Likelihood ratio test= 40.46  on 1 df,   p=2e-10
## Wald test            = 39.05  on 1 df,   p=4e-10
## Score (logrank) test = 40.73  on 1 df,   p=2e-10
```

```
# Hazard ratio is exp(coef) = 0.4892
# means almost 50% less likely to die with revasc?
exp(coef(revasc.mod) + c(-1.96,1.96) * sqrt(revasc.mod$var[1,1]))
```

```
## [1] 0.3909648 0.6122096
```

```
# 95% confidence interval is [0.3909648, 0.6122096]
```

```
# stchange

summary(stchange.mod)
```

```
## Call:
## coxph(formula = Surv(days, death) ~ stchange, data = GRACE1000)
##
##   n= 1000, number of events= 324
##
##            coef exp(coef) se(coef)    z Pr(>|z|)
## stchange 0.5189    1.6802   0.1185 4.38 1.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## stchange      1.68     0.5952     1.332     2.119
##
## Concordance= 0.57  (se = 0.013 )
## Likelihood ratio test= 20.17  on 1 df,   p=7e-06
## Wald test            = 19.19  on 1 df,   p=1e-05
## Score (logrank) test = 19.62  on 1 df,   p=9e-06
```

```
# Hazard ratio = 1.6802
# means almost 70% mroe likelt to die with stchange?
exp(coef(stchange.mod) + c(-1.96,1.96) * sqrt(stchange.mod$var[1,1]))
```

```
## [1] 1.332036 2.119339
```

```
# 95% confidence interval is [1.332036, 2.119339]
```

5: Advanced Methods

- Modifiying to have start time as revascdays
  - or split to have a second observation per patient?

- Stratify on stchange
- Gap model with time between revascularization and follow up time?

```
revasc.strata.mod <- coxph(Surv(days,death) ~ strata(stchange) + revasc, data = GRACE1000)
summary(revasc.strata.mod)
```

```
## Call:
## coxph(formula = Surv(days, death) ~ strata(stchange) + revasc,
##     data = GRACE1000)
##
##   n= 1000, number of events= 324
##
##           coef exp(coef) se(coef)      z Pr(>|z|)
## revasc -0.7300    0.4819   0.1144 -6.378 1.79e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## revasc    0.4819      2.075    0.3851    0.6031
##
## Concordance= 0.591  (se = 0.014 )
## Likelihood ratio test= 42.16  on 1 df,   p=8e-11
## Wald test            = 40.68  on 1 df,   p=2e-10
## Score (logrank) test = 42.51  on 1 df,   p=7e-11
```

```
# about the same as without strata
```

References

Hosmer, D.W. and Lemeshow, S. and May, S. (2008) Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition, John Wiley and Sons Inc., New York, NY