PSTAT-175-Final-Project

William Nelson, Cole Spyksma, Alicia Qu

2025-06-01

```
library(readr)
library(tidyverse)
library(survival)
library(survminer)
library(ggplot2)
library(knitr)
```

```
# Read in data
GRACE1000 <- read_table("GRACE1000.dat", col_names = FALSE)
#GRACE1000 <- read_table("~/Downloads/GRACE1000.dat", col_names = FALSE)
GRACE1000 <- GRACE1000 %>% select(-X10)
colnames(GRACE1000) <- c("id", "days", "death", "revasc", "revascdays", "los", "age", "sysbp", "stchang
```

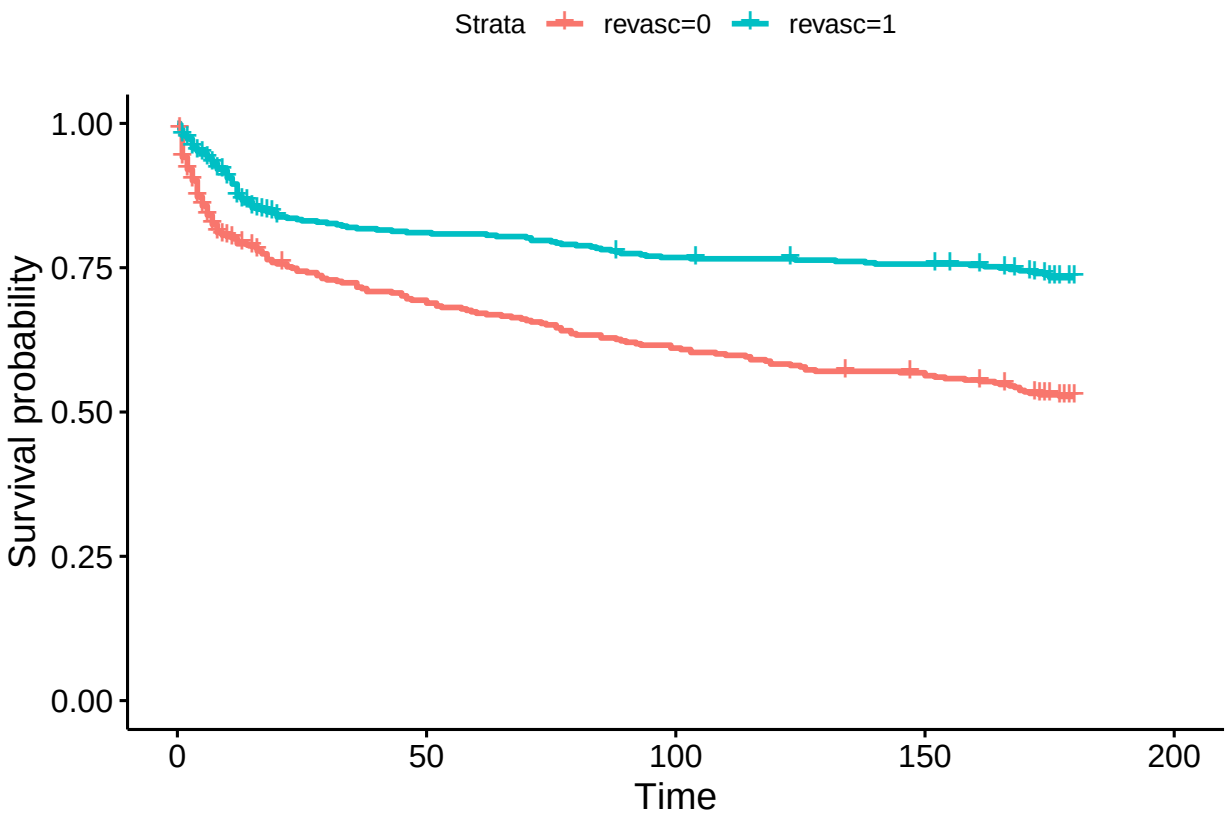
Final Project

1: Introduction

We will be analyzing the "GRACE1000" dataset from Hosmer and Lemenshow and May (2008), which contains data on 1000 patients who were part of a study on the effectiveness of revascularization by the Global Registry of Acute Coronary Events (GRACE). 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. The main research question of interest is whether revascularization is associated with higher survival rates after completing the procedure.

Here is a plot of the survival functions of patients who had the revascularization procedure in blue and patients who did not have the procedure done in red:

```r
ggsurvplot(survfit(Surv(days,death) ~ revasc, data = GRACE1000))
```



We can see that at first glance, patients who had revascularization done appear to have a higher survival probability compared to patients who did not have the procedure done. We will further explore this data with the tools we have learned:

2: Model Fitting

```r
# Forward Stepwise Selection
# Possible covariates are revasc, sysbp, age, stchange

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

kable(AIC(age.mod, sysbp.mod, stchange.mod, revasc.mod))
```

	df	AIC
age.mod	1	4151.309
sysbp.mod	1	4245.999
stchange.mod	1	4252.286

	df	AIC
revasc.mod	1	4231.999

```
kable(BIC(age.mod, sysbp.mod, stchange.mod, revasc.mod))
```

	df	BIC
age.mod	1	4155.090
sysbp.mod	1	4249.780
stchange.mod	1	4256.067
revasc.mod	1	4235.780

Upon analyzing the Cox proportional hazards models, the results indicate that age alone provides the best balance between model fit and simplicity, as evidenced by the lowest AIC value of 4151.309 and BIC value of 4155.090. In comparison, models using systolic blood pressure (AIC = 4245.999), ST changes (AIC = 4252.286), and revascularization status (AIC = 4231.999) each demonstrate higher AIC values, suggesting that these individual predictors are less informative in explaining survival outcomes. The differences between AIC values are considerable, with the age model clearly outperforming others by a margin exceeding 4-10 AIC units, reinforcing the strength of age as a predictor.

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

kable(AIC(age.sysbp.mod, age.stchange.mod, age.revasc.mod))
```

	df	AIC
age.sysbp.mod	2	4131.179
age.stchange.mod	2	4136.352
age.revasc.mod	2	4136.017

```
kable(BIC(age.sysbp.mod, age.stchange.mod, age.revasc.mod))
```

	df	BIC
age.sysbp.mod	2	4138.741
age.stchange.mod	2	4143.913
age.revasc.mod	2	4143.579

In this analysis, we compared three multivariable Cox models incorporating age with additional predictors: systolic blood pressure, ST change, and revascularization status. The model that included age and systolic blood pressure yielded the lowest AIC value (4131.179), indicating the best fit among the multivariable models tested. Models incorporating age with ST change (AIC = 4136.352) and age with revascularization (AIC = 4136.017) had slightly higher AIC values, suggesting marginally inferior fits. The consistently low AIC values across all models reinforce the importance of age as a key predictor while indicating that including systolic blood pressure may offer a slight improvement over the other two variables when combined with age.

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

kable(AIC(sysbp.age.stchange.mod, sysbp.age.revasc.mod))
```

	df	AIC
sysbp.age.stchange.mod	3	4117.788
sysbp.age.revasc.mod	3	4116.536

```
kable(BIC(sysbp.age.stchange.mod, sysbp.age.revasc.mod))
```

	df	BIC
sysbp.age.stchange.mod	3	4129.130
sysbp.age.revasc.mod	3	4127.879

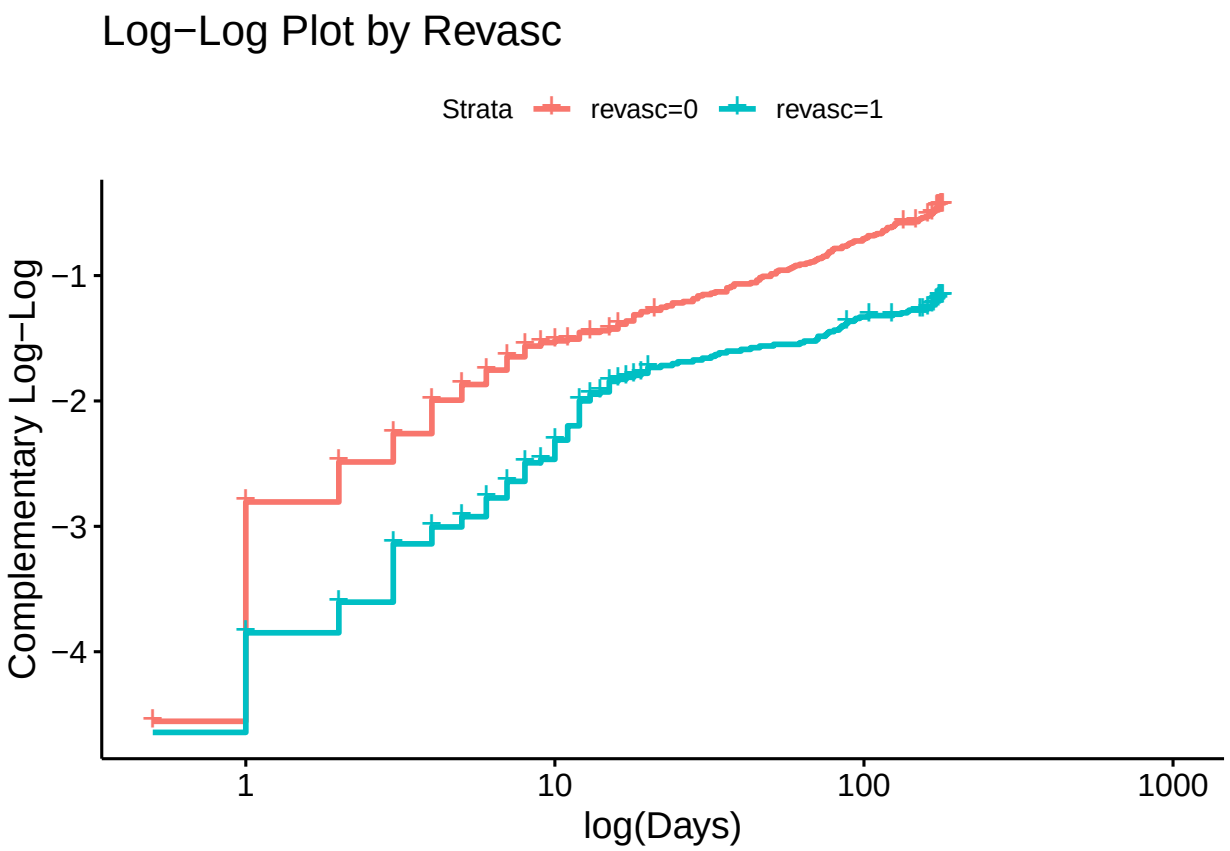
```
# See if stchange lowers AIC and BIC
all.mod <- coxph(Surv(days,death) ~ stchange + sysbp + revasc + age, data = GRACE1000)
kable(data.frame(Metric = c("AIC", "BIC"), Value = c(AIC(all.mod), BIC(all.mod))))
```

Metric	Value
AIC	4100.721
BIC	4115.844

After fitting a comprehensive Cox proportional hazards model incorporating all considered predictors: ST change, systolic blood pressure, revascularization, and age, the resulting AIC value was 4100.721 and BIC was 4115.844, both lower than those from the previous models. This indicates that including all variables yields the best-fitting model based on forward stepwise selection criteria. These results suggest that while age alone and age in combination with other predictors offer substantial predictive power, the full model provides the most accurate explanation of survival probability in this dataset, balancing model complexity with fit.

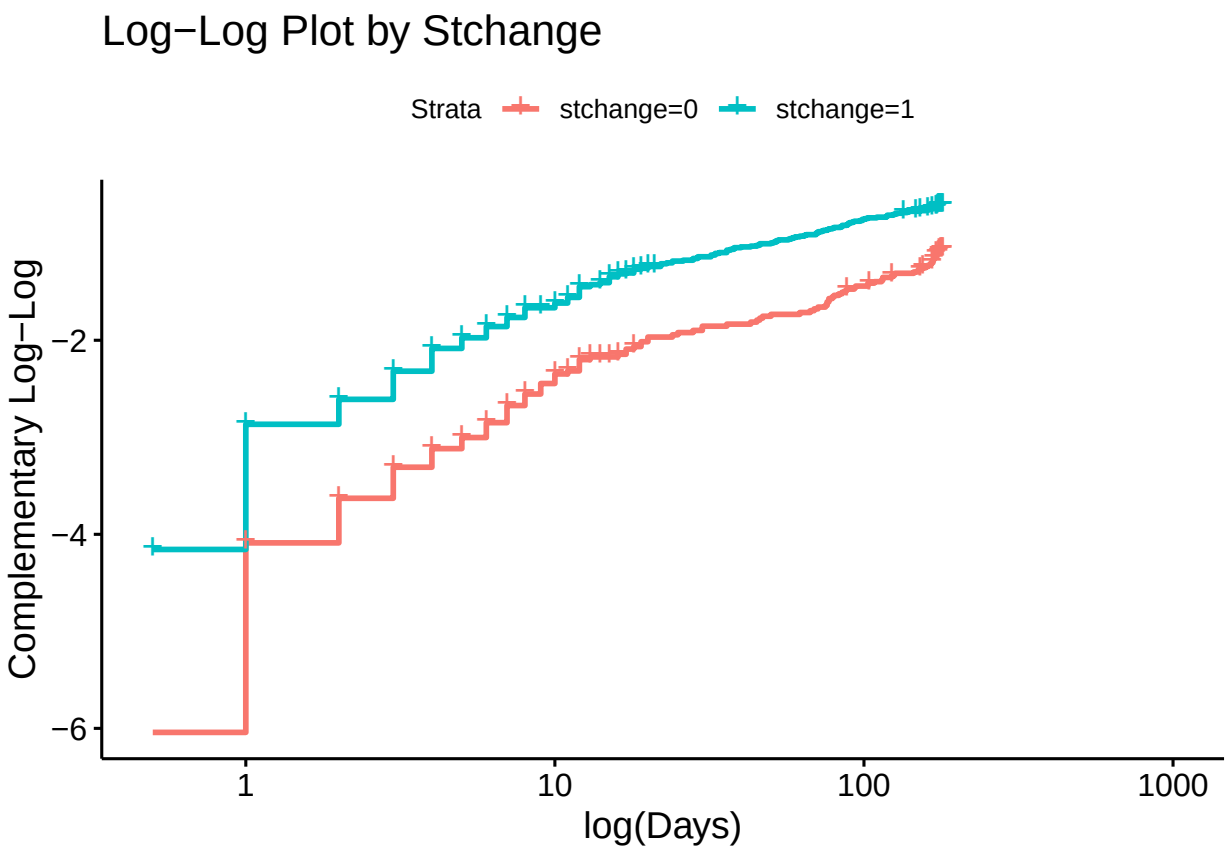
3: Check Proportional Hazards Assumptions

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



In this Complementary Log-Log plot of the treatment and control groups, the curves for patients who underwent revascularization (revasc=1) and those who did not (revasc=0) are presented. The relatively parallel nature of the curves suggests that the proportional hazards assumption holds reasonably well for the revascularization variable, as the log-log curves for the two groups do not cross and maintain a fairly consistent vertical separation throughout the range of time. Overall, the log-log plot supports the use of the Cox proportional hazards model with revascularization as a predictor

```
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
```

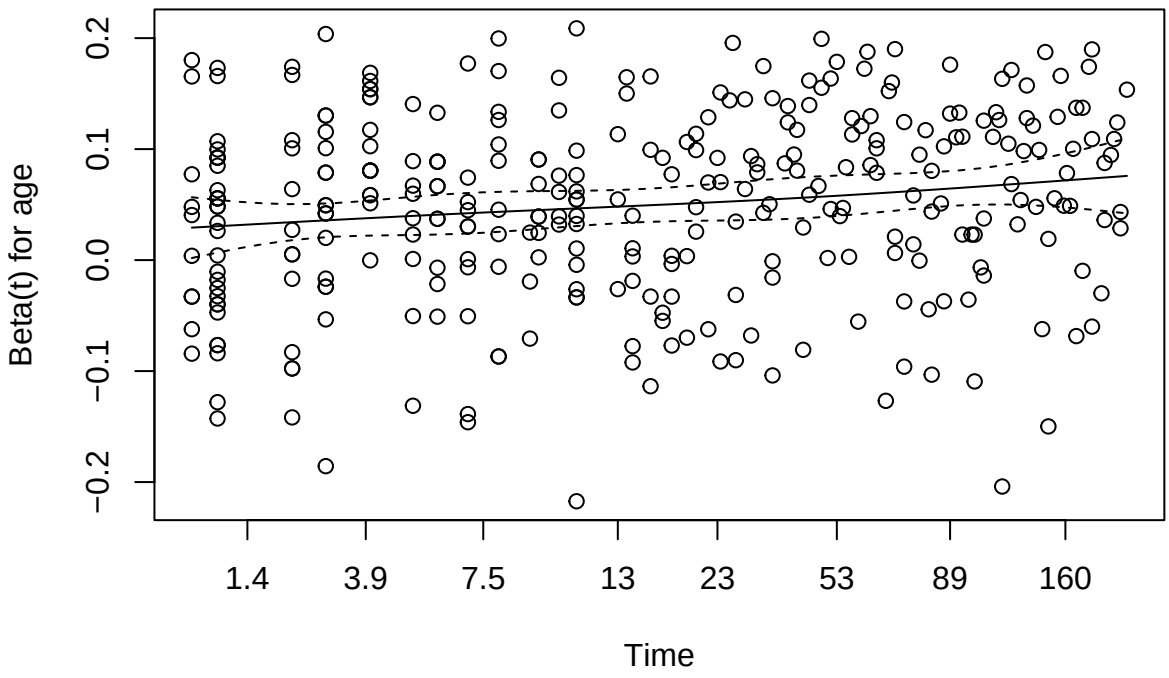
The Complementary Log-Log plot for Stchange also displays curves that do not cross into higher time values, meaning the proportional hazards assumption also holds here with the different Stchange values.

ZPH plot for age:

```
kable(as.data.frame(cox.zph(age.mod)$table))
```

	chisq	df	p
age	7.758003	1	0.0053475
GLOBAL	7.758003	1	0.0053475

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



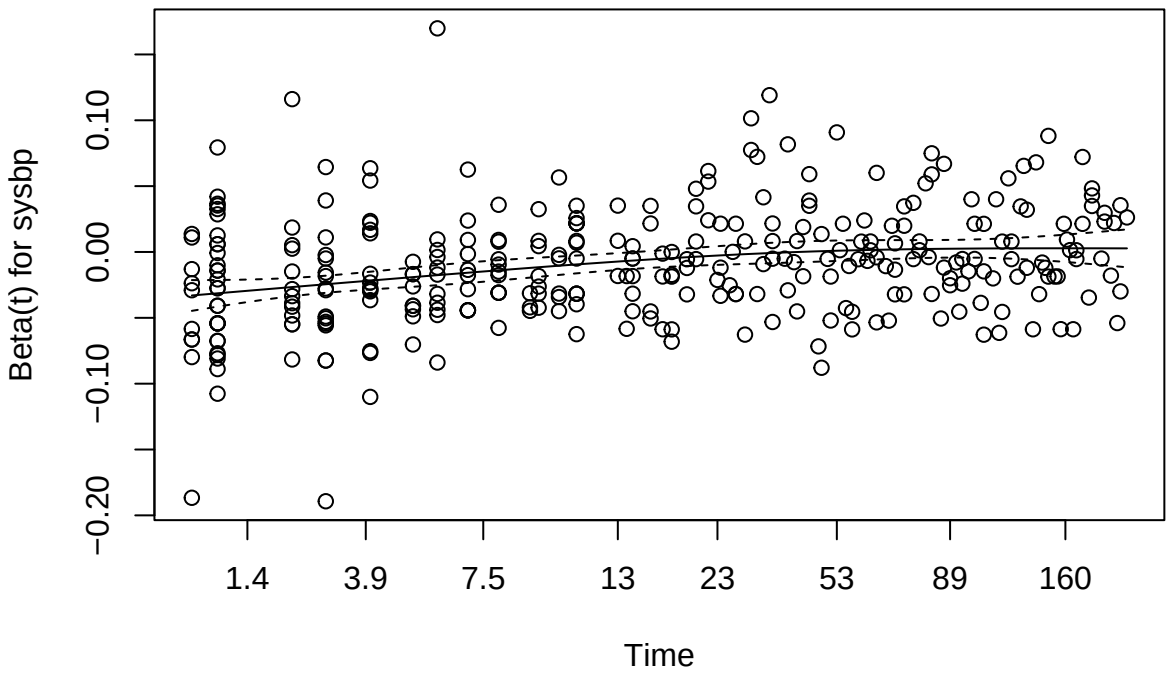
The statistical ouput shows a p-value of 0.0053 for both age and the global test, which is highly significant ($p < 0.05$). This indicates a violation of the proportional hazards assumption for age, meaning that the effect of age on the hazard is not constant over time. This zph plot visualizes Schoenfeld residuals for age over time in the Cox model. The residuals appear randomly scattered around the horizontal axis, and the smoothed curve does not display a clear trend. This pattern suggests that the proportional hazards assumption for age holds reasonably well. Although we can't see it in the graph, with the p-value being that low we will still assume the ph assumptions are being violated.

ZPH Plot for Systolic Blood Pressure

```
kable(as.data.frame(cox.zph(sysbp.mod)$table))
```

	chisq	df	p
sysbp	33.18016	1	0
GLOBAL	33.18016	1	0

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



Here we see Systolic Blood Pressure also violating the proportional hazards assumption with a p_value of 8.4e-09. Once again it doesn't appear to have any clear correlation in the graph but with a p_value that low we can not ignore that there is a clear time-varying effect in the model.

4: Conclusions

Our main scientific question of interest is whether the revascularization procedure significantly increases patients' survival probabilities, and while we have a general idea, we can take a look at the results of our models:

```
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
```

```
exp(coef(revasc.mod) + c(-1.96,1.96) * sqrt(revasc.mod$var[1,1]))
```

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

```
exp(confint(revasc.mod))
```

```
##            2.5 %    97.5 %
## revasc 0.3909664 0.6122071
```

The summary of the revasc.mod Cox model indicates that revascularization is a highly significant predictor of survival ($p < 0.0001$), with a hazard ratio of 0.4892 and a 95% confidence interval of [0.391, 0.612]. This means that patients who underwent revascularization were approximately 51% less likely to die compared to those who did not undergo the procedure. The hazard ratio being less than 1 signifies a protective effect of revascularization.

```
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
```

```
exp(coef(stchange.mod) + c(-1.96, 1.96) * sqrt(stchange.mod$var[1, 1]))
```

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

```
exp(confint(stchange.mod))
```

```
##             2.5 %  97.5 %
## stchange 1.332041 2.11933
```

Patients with an ST change on their initial ECG have a 68% higher hazard of mortality compared to those without such deviations, after controlling for time. This is a clinically meaningful result, as it highlights that ST changes detected early in an ECG are strongly associated with worse survival outcomes. The statistical significance ($p < 0.00001$) and a confidence interval for the hazard ratio of [1.332, 2.119] (an interval greater than 1) suggest that this association is reliable.

5: Advanced Methods

Splitting data on the day the patient got the revascularization

When trying to answer the question of whether the revascularization procedure is associated with higher survival rates, we run into a problem with our data in that not all of our observations had the revascularization procedure on day 0 or the first day they were a part of the study. This means that if a patient had the revascularization done on the 5th day (revascdays = 5), then our models are counting the first 5 days as this patient already having the procedure done, when in reality it has not happened until later. To fix this issue with the data, we will treat "revasc" as a binary time-dependent covariate that changes at some point in the study. To implement the time-varying covariate, we will split every observation that had the revascularization procedure done into 2 observations where the first observation has a start time of 0 and a stop time of the day they had the procedure done, and the second observation will have a start time of the day the procedure was done and a stop time of when the observation was censored. This way our models will properly understand that patients have not had the procedure done until the day given by "revascdays", and therefore are not a part of the treatment group until then.

We will manipulate the data to achieve this:

```
library(dplyr)
library(tidyr)

grace2 <- GRACE1000 %>%
  mutate(
    start_1 = 0,
    stop_1 = pmin(revascdays, days, na.rm = TRUE),
    status_1 = ifelse(days <= revascdays, death, 0),
    revasc_status_1 = 0,  # Renamed for clarity

    start_2 = revascdays,
    stop_2 = ifelse(revasc == 1, days, NA),
    status_2 = ifelse(revasc == 1 & days > revascdays, death, NA),
    revasc_status_2 = 1  # Renamed for clarity
  ) %>%
  pivot_longer(
    cols = c(start_1, stop_1, status_1, revasc_status_1, start_2, stop_2, status_2, revasc_status_2),
    names_to = c("variable", "interval"),
    names_pattern = "([a-z_]+)_([12])"
  ) %>%
  pivot_wider(
    names_from = variable,
    values_from = value
  ) %>%
  filter(!is.na(stop)) %>%
  arrange(id, interval) %>%
  filter(!(start==0 & stop==0))

# Check output
kable(head(grace2))
```

id	days	death	revasc	revascdays	los	age	sysbp	stchange	interval	start	stop	status	revasc_status
1	180	0	0	180	9	28	107	1	1	0	180	0	0
2	5	0	1	0	5	32	121	1	2	0	5	0	1

id	days	death	revasc	revascdays	los	age	sysbp	stchange	interval	start	stop	status	revasc_status
3	2	0	0	2	2	33	150	0	1	0	2	0	0
4	5	0	1	2	5	35	172	0	1	0	2	0	0
4	5	0	1	2	5	35	172	0	2	2	5	0	1
5	180	0	1	9	10	35	106	0	1	0	9	0	0

Cox PH to show the effect of revascularization now that we have accounted for time

```
tvc <- coxph(Surv(start, stop, status) ~ revasc, data = grace2)
summary(tvc)
```

```
## Call:
## coxph(formula = Surv(start, stop, status) ~ revasc, data = grace2)
##
##   n= 1353, number of events= 319
##    (12 observations deleted due to missingness)
##
##           coef exp(coef) se(coef)      z Pr(>|z|)
## revasc -0.6920    0.5006   0.1150 -6.019 1.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## revasc    0.5006      1.998    0.3996    0.6271
##
## Concordance= 0.585  (se = 0.014 )
## Likelihood ratio test= 37.41  on 1 df,   p=1e-09
## Wald test            = 36.23  on 1 df,   p=2e-09
## Score (logrank) test = 37.69  on 1 df,   p=8e-10
```

In this Cox proportional hazards model, we analyzed the effect of revascularization on patient survival, accounting for the time-varying nature of the treatment by splitting the dataset at the point when revascularization was performed. The model revealed a significant association between revascularization and reduced hazard of death, with a hazard ratio of approximately 0.50 (95% CI: 0.40, 0.63), indicating that after revascularization, the risk of death was roughly half that of patients who had not yet undergone the procedure. This approach effectively modeled the time-dependent impact of revascularization by allowing the treatment's hazard to vary based on when it was performed during the follow-up period. Overall, these results suggest that revascularization is associated with a substantial and statistically significant improvement in survival, which becomes apparent when we account for the dynamic nature of treatment over time.

Test to see if we want to use the other covariates:

```
zph_test <- cox.zph(coxph(Surv(start, stop, status) ~ revasc + age + sysbp + stchange, , data = grace2))
kable(as.data.frame(zph_test$table))
```

	chisq	df	p
revasc	1.100055	1	0.2942539
age	7.441267	1	0.0063745
sysbp	24.201243	1	0.0000009
stchange	9.519219	1	0.0020333
GLOBAL	43.802813	4	0.0000000

We would not want to use any of the other covariates in this case as they all fail to meet the PH assumptions even while splitting on the day of revascularization

Following the textbook - "suppose the clinicians on the GRACE study wanted to determine whether the effect depended on the number of days from admission to revascularization."

```
tvc_interaction<- coxph(Surv(start, stop, status) ~ revasc_status * revascdays, data = grace2)
```

```
## Warning in Surv(start, stop, status): Stop time must be > start time, NA
## created
```

```
summary(tvc_interaction)
```

```
## Call:
## coxph(formula = Surv(start, stop, status) ~ revasc_status * revascdays,
##     data = grace2)
##
##   n= 1353, number of events= 319
##    (12 observations deleted due to missingness)
##
##                               coef exp(coef)  se(coef)       z Pr(>|z|)
## revasc_status            -3.119073  0.044198  0.241760 -12.902   <2e-16 ***
## revascdays               -0.023352  0.976919  0.001717 -13.597   <2e-16 ***
## revasc_status:revascdays  0.029756  1.030203  0.021072   1.412    0.158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                           exp(coef) exp(-coef) lower .95 upper .95
## revasc_status                0.0442    22.6254   0.02752   0.07099
## revascdays                   0.9769     1.0236   0.97364   0.98021
## revasc_status:revascdays     1.0302     0.9707   0.98852   1.07364
##
## Concordance= 0.818  (se = 0.014 )
## Likelihood ratio test= 308.3  on 3 df,   p=<2e-16
## Wald test            = 212.5  on 3 df,   p=<2e-16
## Score (logrank) test = 282.6  on 3 df,   p=<2e-16
```

In this interaction Cox proportional hazards model, we examined whether the effect of revascularization on survival depended on the number of days from admission to the revascularization procedure. The model included revasc_status, revascdays, and their interaction term. The analysis revealed that both revasc_status and revascdays were highly significant with p-values less than 2e-16, indicating that both receiving revascularization and the number of days from admission to the procedure independently affect survival. The interaction term (revasc_status:revascdays) was not statistically significant ($p = 0.158$), suggesting that the impact of revascularization on survival does not vary significantly based on how many days after admission

it was performed. The hazard ratio for the interaction term was approximately 1.03 (95% CI: 0.97–1.07), meaning the modification effect is negligible. In summary, while both revascularization and time since admission to revascularization individually influence survival, their interaction does not significantly alter the risk of death in this dataset.

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