Modeling the Effects of Gender on Cardiovascular Disease Survival

Arthur Hla, Timothy Nguyen, Daniel Xi

December 10, 2018

1 Introduction

Our team set out to apply survival analysis to learn what we can about the variables affecting the time until cardiovascular disease related death. To this end, we have selected a dataset from containing time until death for 453 individuals, as well as their cause of death and other anatomical features.

In our initial analysis, we excluded subjects who died due to causes other than cardiovascular disease (CVD). For those subjects who experienced CVD related deaths, we sought to identify factors which significantly impact patients' survival rates. Specifically, we were interested in any differences between males and females regarding CVD survival.

We performed a second analysis with our complete dataset, including subjects who died due to causes other than CVD. Here, we apply competing risk methods to accommodate the additional causes of death as we aimed to answer the same questions posed in our initial analysis.

Ultimately, we found that BMI is highly significant in estimating the likelihood of cardiovascular disease.

2 Exploratory Data Analysis

2.1 Metadata

We begin with a preliminary analysis to provide a summary understanding of our dataset. Our data is comprised of the times until death for 453 individuals, measured from the start of the study. For each of these subjects, we also had additional bodily information as described in Table 1. Table 2 provides a sample of the data.

Table 1: Features available for each subject

Variable	Description	Codes..Values
age	Age	Years
bmi	Body Mass Index	kg/mˆ2
time	Time Until Death	Days
gender	Gender	0 = Male
		1 = Female
ev_typ	Cause of Death	1 = CVD
		2 = Other Cause
		0 = Censored

Table 2: First 3 observations from the data

id	age	gender	bmi	time	ev_typ
1	20	male	27.85998	1923	2
2	23	male	18.92089	1353	2
3	26	male	29.64232	1266	2

Table 3: Event frequencies

Event Type	Frequency
Censor	116
CVD	167

Event Type	Frequency
Other Cause	170

2.2 Data Missingness and Redundancy

Our data has no missing values nor duplicated records.

```
sapply(data, function(x) sum(is.na(x)))
```

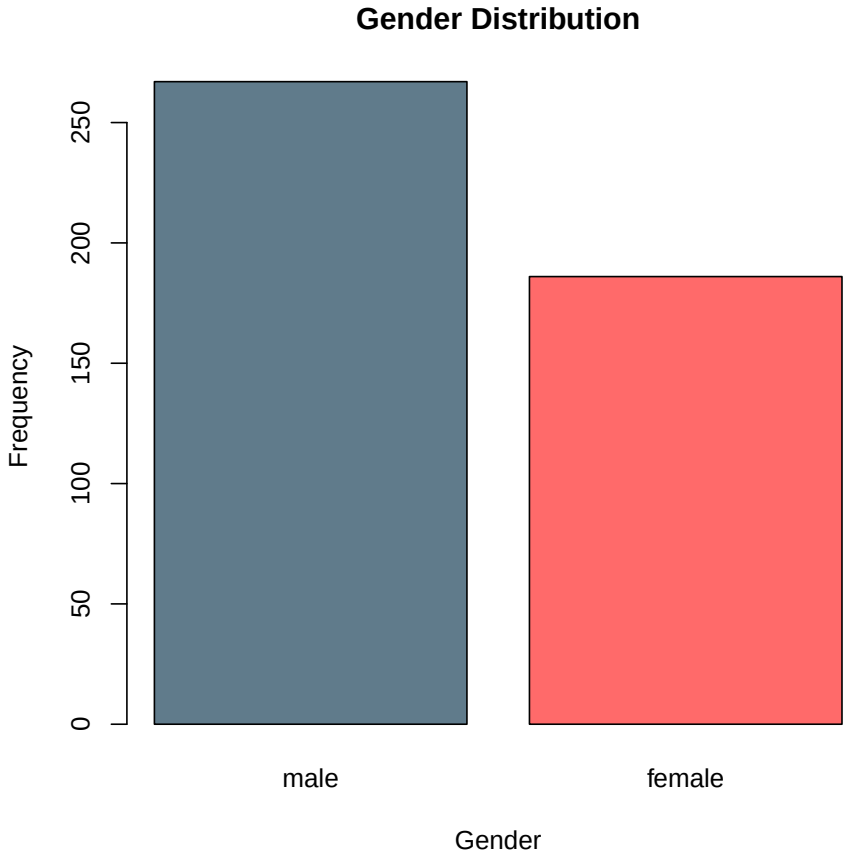
```
##     id    age gender    bmi   time ev_typ
##      0      0      0      0      0      0
```

```
sum(duplicated(data))
```

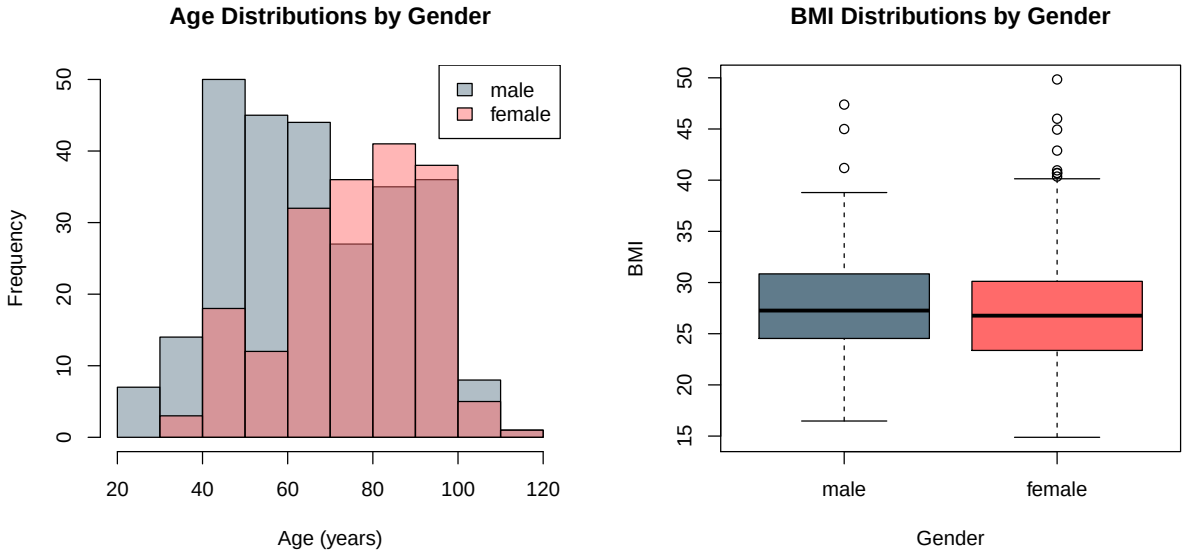
```
## [1] 0
```

2.3 Variable Distributions

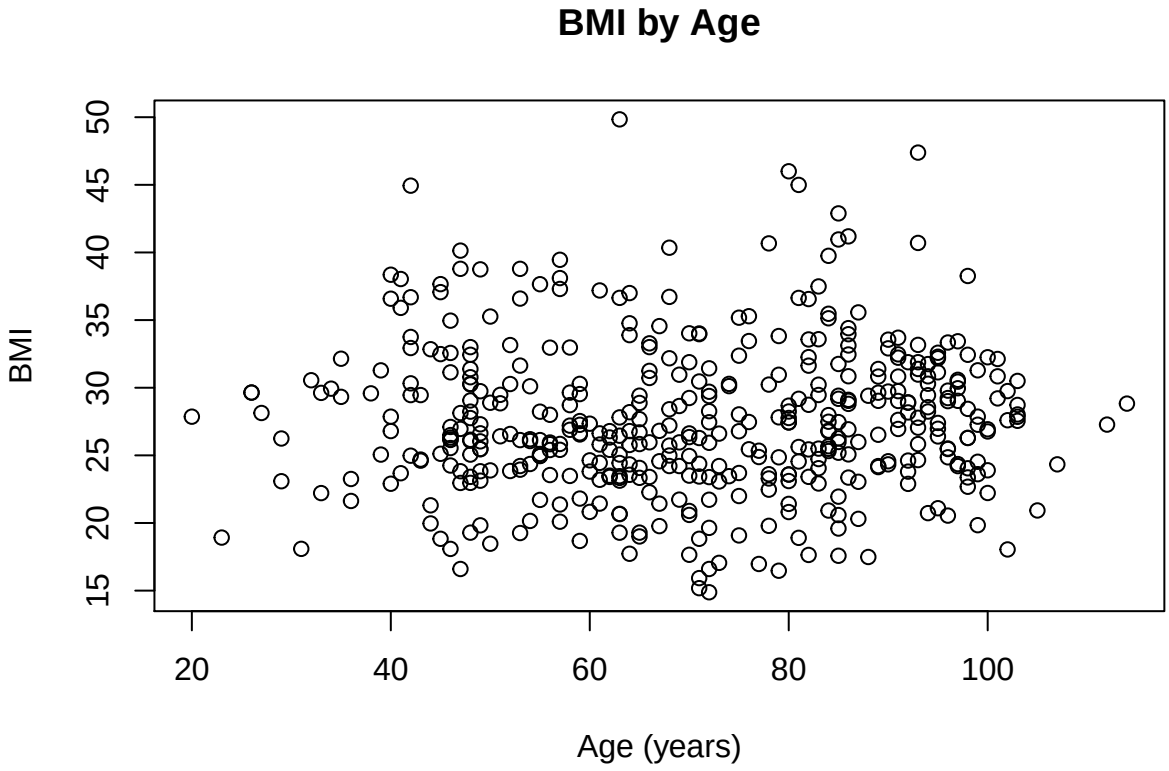
Here, we examine the distributions of various variables in our dataset, starting with gender. We found that males comprised 267 (59%) of our subjects, with females making up 186 (41%).



Furthermore, we explored distributions of our subjects age and Body Mass Index (BMI), each broken out by gender. From plotting the age distributions for males and females, we observe that most of our female subjects are older than most of our male subjects. From our boxplot, we see that more females are outliers on the high end of the BMI spectrum than males, and males produce a slightly tighter BMI distribution. Otherwise, we see little difference in BMI between males and females.



Finally, we plot a scatter of age against BMI and observe little to no correlation between the two variables.

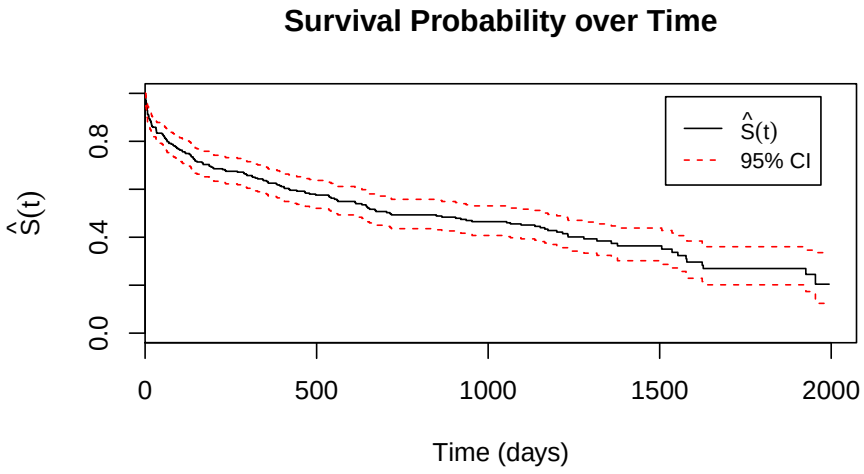


2.4 Kaplan-Meier Estimation of Survival Function

We now use Kaplan-Meier (KM) estimation to estimate the survival function, which represents an individual's survival rates over time. We fit our KM estimator on a subset of our total data that excludes subjects who died from diseases other than cardiovascular disease.

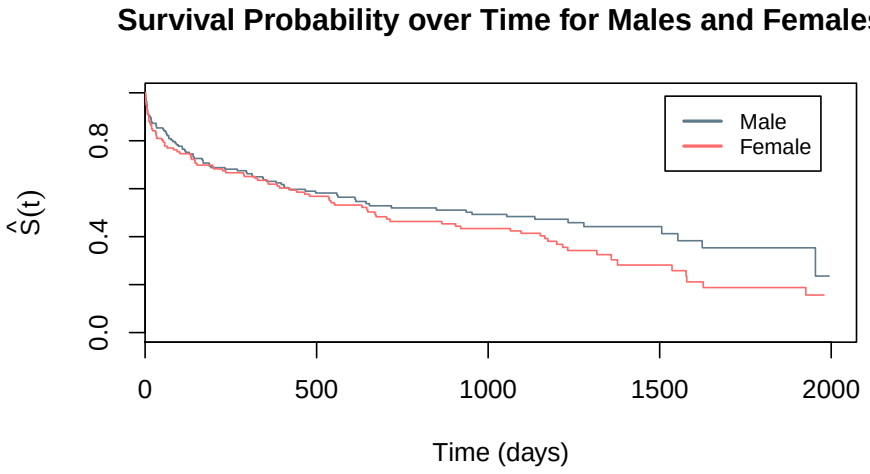
```
data.cvd <- data[!data$ev_typ == 2,] #excluding event type 2 (death from other diseases)
surv.cvd <- Surv(data.cvd$time, data.cvd$ev_typ)

KMsurv.1 <- survfit(surv.cvd ~ 1)
```

The resulting KM plot shows that survival rates decline sharply for over the first 200 days, and steadily thereafter. We omit code used to plot these estimates because it is repetitive and reduces readability. We leave the survival function estimation code, which serves as the main focus.



Next, we wanted to see if gender had an impact on survival rates. Here, our KM plot seemingly indicates that males have consistently higher survival rates over time. Furthermore, the survival rates seem converge beyond 2000 days.

```
KMsurv.gender <- survfit(surv.cvd ~ gender, data = data.cvd)
```



To test whether males exhibit significantly higher survival rates than females, we apply a log-rank test. The hypotheses for this test are as follows:

$$H_0 : \hat{S_{male}}(t) = \hat{S_{female}}(t)$$
$$H_a : \hat{S_{male}}(t) \neq \hat{S_{female}}(t)$$

```
survdiff(surv.cvd ~ data.cvd$gender)
```

```
## Call:
## survdiff(formula = surv.cvd ~ data.cvd$gender)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## data.cvd$gender=male   157       83     92.7      1.02       2.3
## data.cvd$gender=female 126       84     74.3      1.27       2.3
##
##  Chisq= 2.3  on 1 degrees of freedom, p= 0.1
```

The p-value is 0.1 is greater than our alpha level of 0.05. Thus, we fail to reject our null hypothesis and conclude that there is no significant difference between the survival rates for men and women.

3 Modeling

3.1 Cox Proportional Hazards Model

3.1.1 Model Fitting

We began fitting univariate models to determine the most significant covariates that would optimize our model. Univariate fitting suggests that the gender covariate is not significant and that age and bmi covariates are significant in the univariate mode.

```
PH.Model.A <- coxph(surv.cvd ~ gender,
                    data = data.cvd)
PH.Model.B <- coxph(surv.cvd ~ bmi,
                    data = data.cvd)
PH.Model.C <- coxph(surv.cvd ~ age,
                    data = data.cvd)
```

Then we tried all possible additive models.

```
PH.Model.D <- coxph(surv.cvd ~ bmi + age,
                     data = data.cvd)
PH.Model.E <- coxph(surv.cvd ~ bmi + gender,
                    data = data.cvd)
PH.Model.F <- coxph(surv.cvd ~ age + gender,
                    data = data.cvd)
PH.Model.G <- coxph(surv.cvd ~ bmi + age + gender,
                    data = data.cvd)
```

Lastly, we tried interactions and stratified models. We thought that age might be related to BMI, so we fitted a model to better understand the relationship between the two covariates. The results from adding the gender covariate to the model still suggested that it was still insignificant, so we stratified on gender since it was slightly unproportional.

```
PH.Model.H <- coxph(surv.cvd ~ bmi*age,
                    data = data.cvd)
PH.Model.I <- coxph(surv.cvd ~ bmi*age + gender,
                    data = data.cvd)
PH.Model.J <- coxph(surv.cvd ~ bmi*age + strata(gender),
                    data = data.cvd)
```

We fitted a few other stratified models and interactions with other covariates, but none seemed to be significant.

3.1.2 Model Selection

According to our table of the AIC of different models, Model J has the lowest AIC. We chose this to be our final CoxPH model.

```
##        A        B        C        D        E        F        G        H
## 1687.789 1686.328 1543.113 1536.299 1686.002 1544.140 1537.340 1532.615
##        I        J
## 1533.671 1303.705
```

```
summary(PH.Model.J)
```

```
## Call:
## coxph(formula = surv.cvd ~ bmi * age + strata(gender), data = data.cvd)
##
##   n= 283, number of events= 167
##
##               coef  exp(coef)   se(coef)      z Pr(>|z|)
## bmi      0.2592807  1.2959975  0.0818040  3.170  0.00153 **
## age      0.1444541  1.1554087  0.0295218  4.893 9.92e-07 ***
## bmi:age -0.0025116  0.9974916  0.0009554 -2.629  0.00856 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## bmi        1.2960     0.7716    1.1040    1.5214
## age        1.1554     0.8655    1.0905    1.2242
## bmi:age    0.9975     1.0025    0.9956    0.9994
##
## Concordance= 0.763  (se = 0.035 )
## Rsquare= 0.428   (max possible= 0.994 )
## Likelihood ratio test= 158.1  on 3 df,   p=<2e-16
## Wald test            = 123.9  on 3 df,   p=<2e-16
## Score (logrank) test = 137.8  on 3 df,   p=<2e-16
```

The hazard rate for BMI (1.295997) suggests that for every unit increase in BMI, the risk of dying increases by 29.60%; specifically, a unit increase in BMI increases the risk of death by a percentage between 10.4% and 15.2% with 95% certainty.

The hazard rate for age (1.155409) suggests that the risk of dying increases by 15.54% for every passing year. The risk of death increases by a percentage between 9% and 22.4% with 95% certainty.

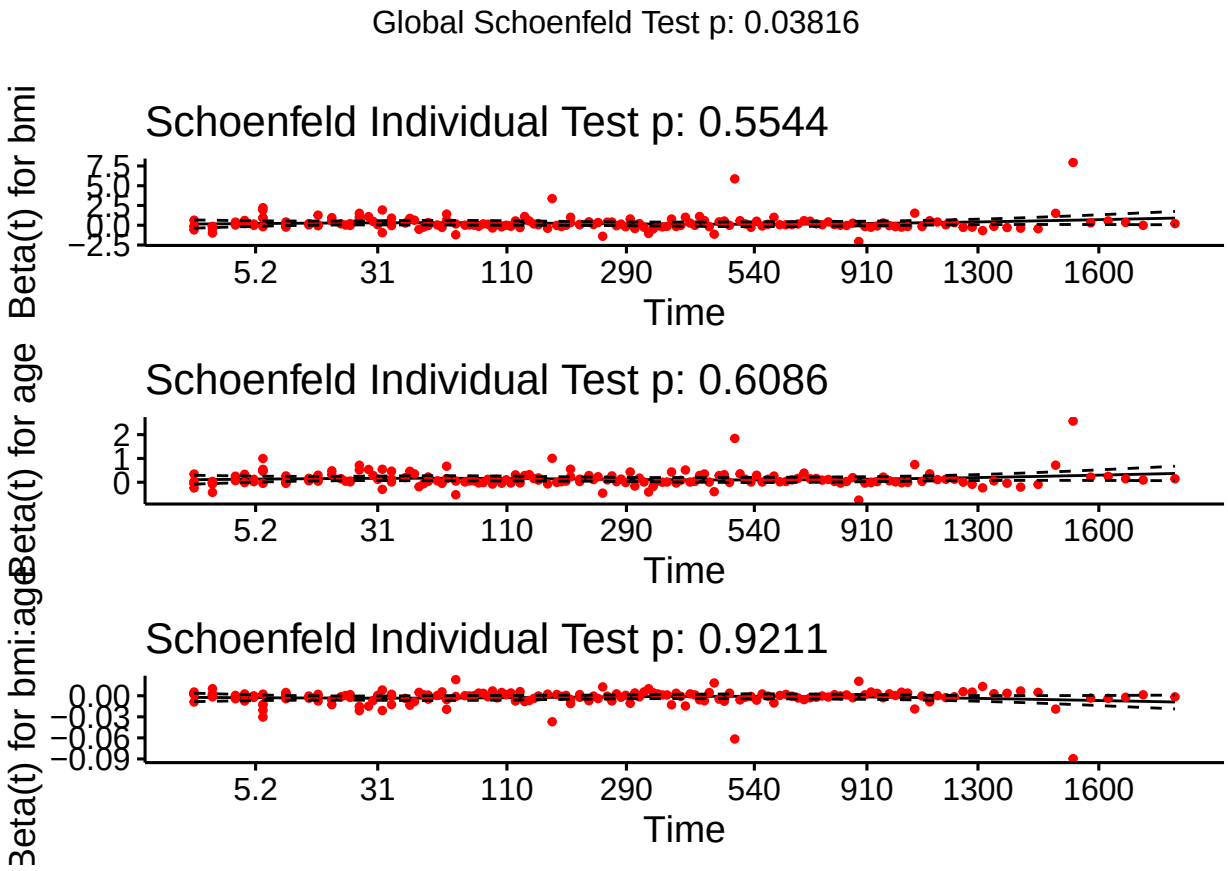
The hazard rate for the interaction between age and BMI (0.9975) suggests that the risk of dying decreases by 0.25% for every unit interaction bewteen age and BMI. The risk of death decreases by a percentage between 0.06% and 0.44% with 95% certainty.

3.1.3 Model Diagnostics

After deciding on a model to use, we began checking diagnostic testing for the assumptions attributed to an appropriate Cox PH model, namely the proportional hazards assumption, detecting non-linearity, and examining any potential outliers or influential observations.

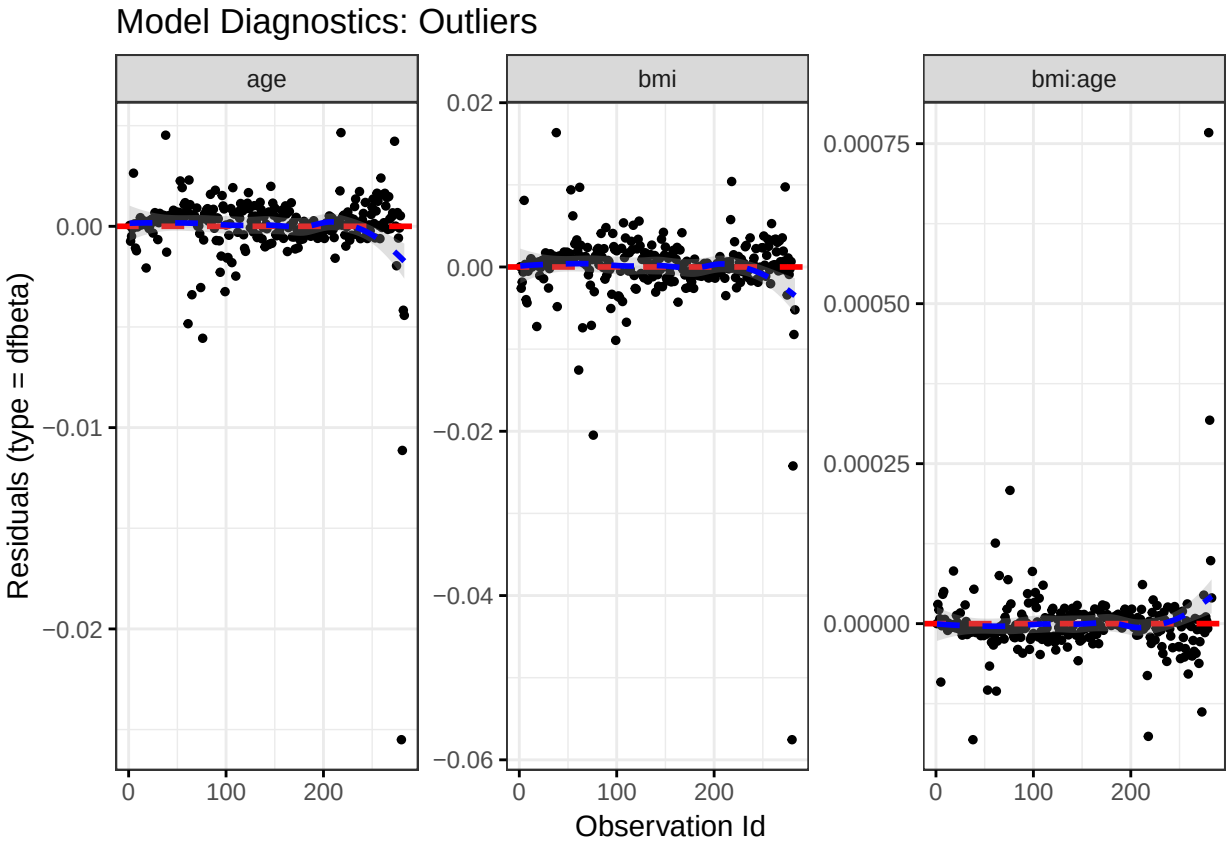
The most integral assumption for a Cox model, the PH (proportional hazards) assumption, suggests that the covariate effects on "survival" have to be independent of time; to test this, we observed our model's Schoenfeld Residuals (SR). In principle, the SR is independent of time, so a plot showing any trends or non-random patterns would suggest a violation of the PH assumption. Specifically, we were looking for a random distribution of observations with a constant mean centered around zero, and set up our hypothesis test as follows:

```
H0: Covariate's effect is independent of time
HA: Covariate's effect is not independent of time
```



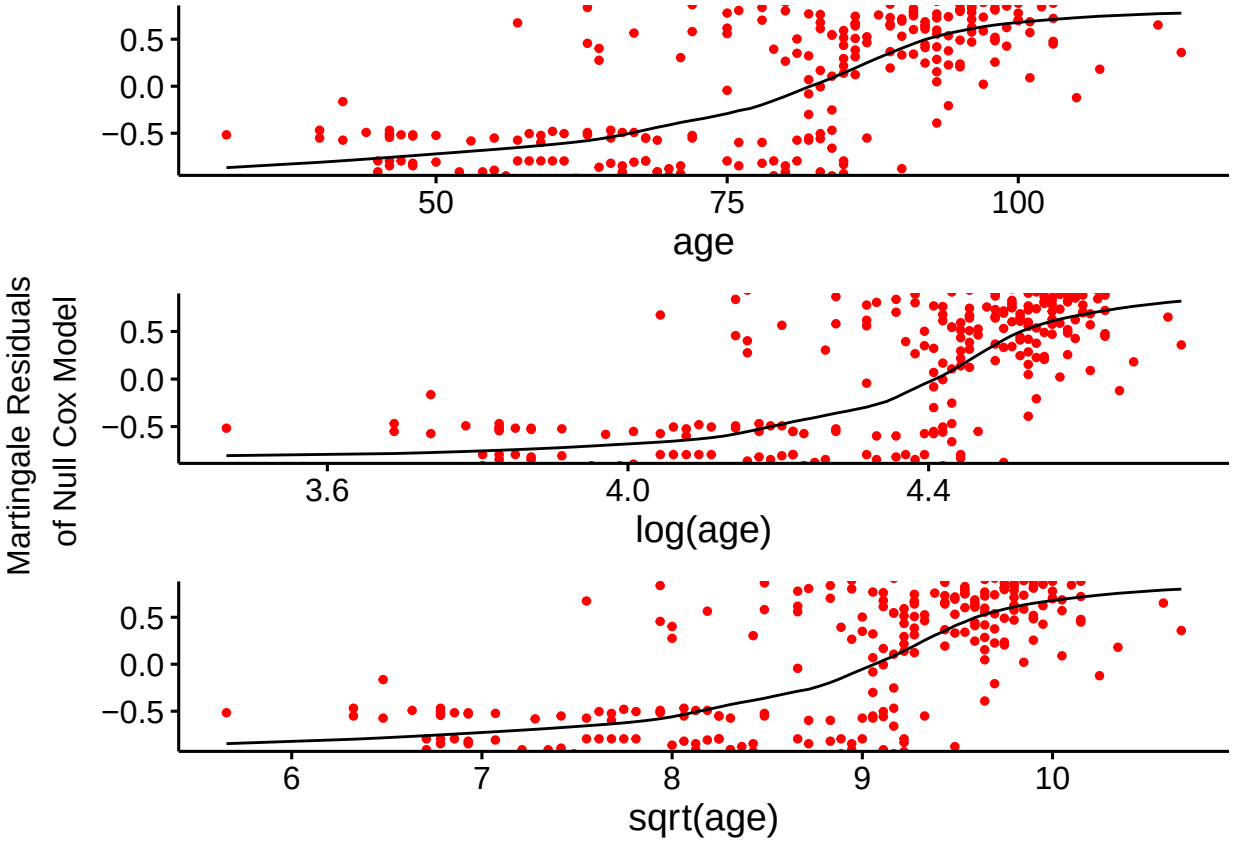
The output above aligns itself well with the conditions of our test: the mean is centered just above zero throughout the entire duration of time, except after time 1300 where it begins fanning slightly; however, we decided that the fanning at the tail end can be attributed to a very small sample size of observations, and thusly isn't enough evidence to reject the PH assumption entirely. Additionally, the p-value associated with our covariates far exceeds the significance level (.05), further supporting our conclusion that the model we chose satisfies the Proportional Hazards assumption.

Assumption: Outliers



We tested for outliers by plotting the residuals of our covariates. Most of the observations were centered around zero, but there were clearly a few outliers in our data; we decided against removing the outliers from our data since it could provide additional insight, coupled with the fact that our dataset was small to begin with.

Assumption: Linearity



Following the diagnostic testing for the PH assumption, we observed the Martingale residuals (MR) against the covariate age to detect any non-linearity. To assess the functional form of our covariate, we plotted age, sqrt(age), and log(age) against MR's of null Cox PH models using the function ggcoxfunctional(); the resulting plots appear to be slightly nonlinear, suggesting that our model may not have been properly fitted. The pattern is not symmetric around 0, which violates the assumption of nonlinearity.

After considering the results of our diagnostic testing, we decided against the model we had chosen since it violated assumptions to be an appropriate Cox PH model. Additionally, we realized that the standard Cox PH model wasn't adequate for our dataset due to the presence of competing risk, since the Cox PH model would treat the competing event of interest (Other Diseases) as censored observations.

3.2 Competing Risks Model

For our new model, we aimed to model the time until "failure" as a function of the most statistically significant covariates. Since we altered our dataset to convert BMI into a categorical variable with four groups, we created three indicator variables with the normal BMI group as the baseline.

```
# Overall P-value for BMI; results conclude BMI's significance as a covariate
wald.test(mod1$var, mod1$coef, Terms = 3:5)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 12.9, df = 3, P(> X2) = 0.0048
```

We built our first regression model (mod1) using every variable as predictors in order to recognize the most statistically influential covariates; by doing so, we concluded that age was the most significant covariate, followed by the marginally significant obese and overweight BMI groups. Additionally, we found the relative risk of the overweight and obese levels of BMI to be around 1/8 (.167 & .125, respectively); the normal BMI group was statistically insignificant with a p-value of 0.180, and yielded a relative risk of around 1/4 (0.252). To further verify the statistical significance of BMI, we utilized the Wald test to determine the overall p-value associated with the bmi covariate as a whole; the p-value yielded (.0048) was significantly less than the level of significance, justifying our decision to keep it in our model.

```
#BMI
mod2 = crr(data$time,data$ev_typ,x[,3:5])
#BMI + age
mod3 = crr(data$time,data$ev_typ,x[,c(3:5,1)])
#BMI + gender
mod4 = crr(data$time,data$ev_typ,x[,c(3:5,2)])

modsel.crr(mod1,mod2,mod3,mod4)
```

```
## Model selection table
##
## Model 0: Null model
## Model 1: crr(ftime = data.bmi$time, fstatus = data.bmi$ev_typ, cov1 = x)
## Model 2: crr(ftime = data$time, fstatus = data$ev_typ, cov1 = x[, 3:5])
## Model 3: crr(ftime = data$time, fstatus = data$ev_typ, cov1 = x[, c(3:5, 1)])
## Model 4: crr(ftime = data$time, fstatus = data$ev_typ, cov1 = x[, c(3:5, 2)])
##   Num.obs  logLik Df.fit    BIC BIC diff
## 0     453 -969.27      0 1938.5  313.385
## 1     453 -800.23      5 1631.0    5.894
## 2     453 -955.48      3 1929.3  304.157
## 3     453 -800.34      4 1625.2    0.000
## 4     453 -950.22      4 1924.9  299.741
```

```
mod.chosen = mod3
```

After building the first model, we began fitting additional models with different covariates in order to determine the most appropriate fit; ultimately, we used the BIC criterion to compare and conclusively decide on a working model: the model including BMI and age covariates (Model 3) yielded the lowest BIC & BIC difference, so we decided on Model 3 as our working model moving forward.

```
summary(mod3)
```

```
## Competing Risks Regression
##
## Call:
## crr(ftime = data$time, fstatus = data$ev_typ, cov1 = x[, c(3:5,
##     1)])
##
##                    coef exp(coef) se(coef)     z p-value
## bmi.normal       0.7110     2.036  0.22716  3.13  0.0017
## bmi.overweight   0.4101     1.507  0.23837  1.72  0.0850
## bmi.obese       -1.3903     0.249  1.03459 -1.34  0.1800
## age              0.0889     1.093  0.00693 12.83  0.0000
##
##                 exp(coef) exp(-coef)   2.5% 97.5%
## bmi.normal          2.036      0.491 1.3045  3.18
## bmi.overweight      1.507      0.664 0.9445  2.40
```

```
## bmi.obese          0.249      4.016 0.0328  1.89
## age                1.093      0.915 1.0782  1.11
##
## Num. cases = 453
## Pseudo Log-likelihood = -800
## Pseudo likelihood ratio test = 338  on 4 df,
```

Checking Model Assumptions:

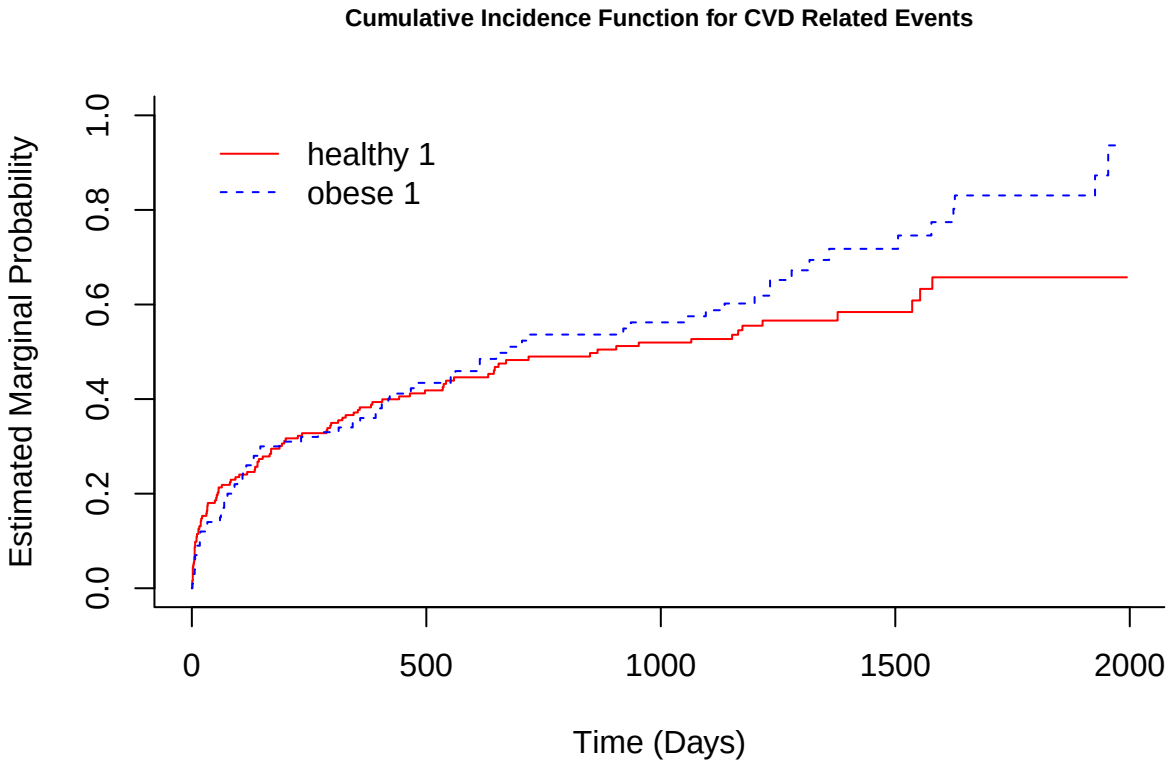
Plotting our Schoenfeld residuals across our covariates yields results supporting the fit of our model. The means of the residuals have a relatively constant mean across time; in most graphs, the mean fluctuates slightly towards the tail end of the time-axis due to a smaller number of observations to record.

```
(ci.overall = cuminc(ftime = data.bmi$time, fstatus = data.bmi$ev_typ))
```

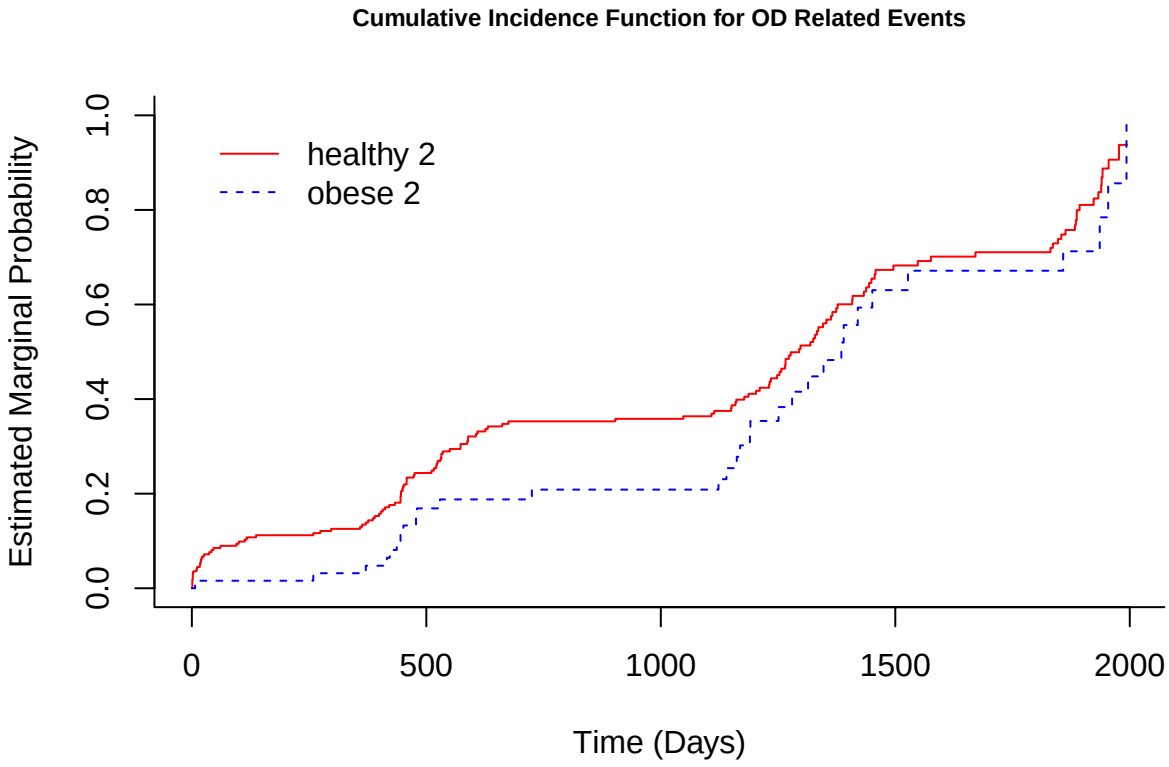
```
## Estimates and Variances:
## $est
##           500      1000      1500
## 1 1 0.2639860 0.3255391 0.3716929
## 1 2 0.1425074 0.2031471 0.4038211
##
## $var
##              500         1000         1500
## 1 1 0.0004328607 0.0005084718 0.0005758851
## 1 2 0.0002778898 0.0003822172 0.0007174177
```

The printed results yield the estimated marginal probability of the possible outcome of death (1 = CVD, 2 = OD) at days 500, 1000, and 1500 after the study began. The resulting figures provided valuable insight on the overall failure trends for CVD and OD induced deaths; for example, the probability of a patient dying of cardiovascular diseases by 500 days is around 26%

Incident of CVD related deaths is higher in obese patients compared to healthy ones; the estimated probability of death between obese and healthy individuals are similar up until 1000 days. Afterward, obese patients have a significantly higher estimated probability of death than their healthy peers.



We were interested in the influence of BMI in the context of CVD induced events; the resulting output confirmed out initial hypothesis that obese patients are more susceptible to CVD mortality than their healthy peers. Additionally, the results suggest that healthy and obese patients are similarly prone to CVD induced death up to 500 days past their last check up, with the risk of death increasing at a much more accelerated pace for obese patients afterwards.



We conducted a similar procedure to see the influence of BMI on non-CVD induced mortalities. The resulting plot confirmed our initial hypothesis that obese patients would be less prone to OD induced deaths than CVD induced; however, we were surprised by the plot's implications for healthy patients – healthy patients seem to be at much higher risk than obese patients for a non-CVD induced death at any point.

Conclusion

Our goal was to apply survival analysis to learn about death related to cardiovascular disease. Furthermore, we sought to understand this considering the competing risks of death for other causes. Overall, we observed that BMI is highly significant in estimating the likelihood of cardiovascular disease. Specifically, overweight people are much more likely to die relating to cardiovascular disease than other reasons. Inversely, underweight people are extremely more likely to die due to some cause other than cardiovascular health than otherwise, with that chance rising over time. This makes sense because heart problems have been proven to directly relate to high body fat content. Additionally, we found that females are more prone to cardiovascular disease. Ultimately, our analysis demonstrated that gender and BMI are significant covariates relative to time until death.

Works cited:

Data Set

1.Gauchospace, https://gauchospace.ucsb.edu/courses/mod/folder/view.php?id=1459647 2. Competitive Risk Dataset, 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

BMI Table

3. Defining Adult Overweight and Obesity, https://www.cdc.gov/obesity/adult/defining.html