Survival Analysis of Prostate Cancer Study

Andrew vanderWilden

December 6, 2019

Contents

1. Abstract

Clinical trials are crucial to developing treatments and further understanding disease progressions. Survival analysis can draw insights about survival times and probabilities from these trials to inform treatment. The data in this report comes from the 1980 paper *The Choice of Treatment for Cancer Patients Based on Covariate Information: Application to Prostate Cancer* by D.P. Byar and S.B. Green. We use both a Cox proportional hazard model and the non-parametric Kaplan Meier estimate to analyze survival times. We find the variables `rx` (treatment), `age`, `wt` (weight index), `bm` (bone metastases), `sz` (tumor size), `hx`(history of cardiovascular disease), and `sg` (combined index of stage and hist. grade) to be significant predictors in a Cox proportional hazard model. Additionally we find a treatment of 1.0 mg estrogen to have a statistically significant positive effect on survival probabilities.

2. Introduction

Orientation Material

Survival analysis aims to differentiate survival times or progression rates of different diseases. This report analyzes survival times of patients with prostate cancer. Differentiating survival times between different groups allows us to evaluate the effectiveness of treatments or how traits of patients affect their expected survival times.

The data in this report comes from the 1980 paper *The Choice of Treatment for Cancer Patients Based on Covariate Information: Application to Prostate Cancer* by D.P. Byar and S.B. Green. The data relate to

a study of patients with stage 3 or 4 prostate cancer. The treatments being studied are different doses of estrogen.

The data are cross-sectional and primarily describe the medical characteristics of the patients.

Key Aspects

This report attempts to fit a Cox proportional hazard model to predict survival probability of patients diagnosed with prostate cancer. We find the variables `rx` (treatment), `age`, `wt` (weight index), `bm` (bone metastases), `sz` (tumor size), `hx`(history of cardiovascular disease), and `sg` (combined index of stage and hist. grade) to be significant predictors in our model. Additionally, using a Kaplan Meier estimate we find a treatment of 1.0 mg estrogen to have a significant positive effect on survival probabilities.

Plan for the Rest of the Report

The outline for the remainder of the report is as follows. In section 3, we present the most important characteristics of the data and the relationships between covariates and survival time using a non- parametric Kaplan-Meier estimate. In section 4, we will present a final Cox Proportional Hazard model and discuss the model selection and building processes. A comparison of the two approaches will follow. Concluding remarks can be found in section 5 with final details in the Appendix.

3. Data Characteristics and Kaplan-Meier Estimate

The data are cross-sectional and contain information about 502 patients in the study of patients with stage 3 or 4 prostate cancer. The data set includes information on the following 18 variables:

Item	Variable	Definition
1	patno	Patient Number
2	stage	Stage
3	rx	Treatment
4	dtime	Months of Follow-up
5	status	Follow-up Status
6	age	Age in years
7	wt	Weight Index = wt(kg) - ht(cm) + 200
8	pf	Performance Rating
9	hx	History of Cardiovascular Disease
10	sbp	Systolic Blood Pressure / 10
11	dbp	Diastolic Blood Pressure / 10
12	ekg	Electrocardiogram code
13	hg	Serum Hemoglobin (gr / 100ml)
14	sz	Size of Primary Tumor (cm squared)
15	sg	Combined Index of Stage and Hist. Grade
16	ap	Serum Prostatic Acid Phosphatase
17	bm	Bone Metastases
18	sdate	Date on Study (as days since January 1, 1960)

In survival analysis, the objective of a study is to determine the varying survival times of patients, or put another way, the time until death. The variable `status` indicates if a patient is still alive, or the cause of death. In this case we are not interested in differentiating between causes of death so we create an indicator variable indicating if a patient is either alive or dead. Using this variable we can then examine the variable of interest `dtime` which indicates the time until death or censorship in measured in months. In this report

we are also interested in the effectiveness of the different treatments and their relation to varying survival times.

The below summary information shows the status of the patients at the end of the study (a 0 indicates the patient was still alive at the end of the study while a 1 indicates the patient died):

```
##   0   1
## 148 354
```

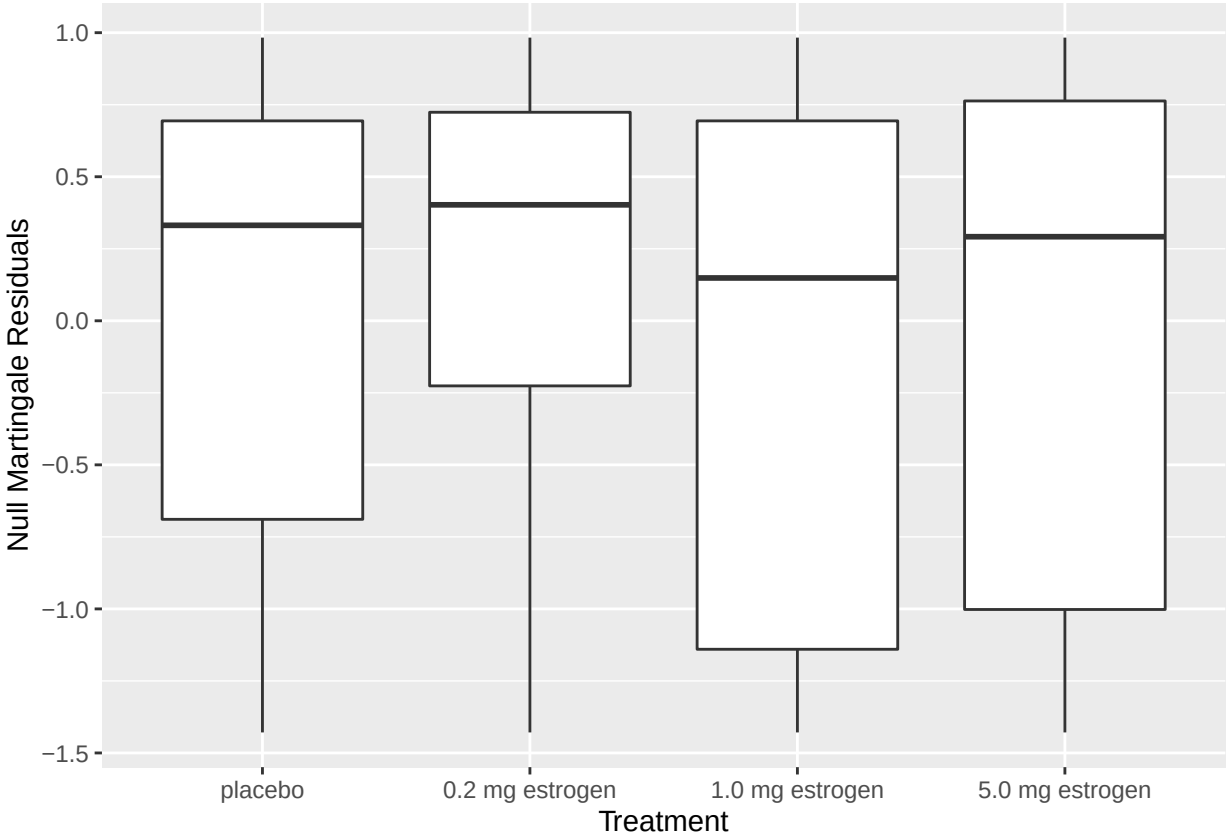
A number of variables had values that were missing. To account for the missing entries, we used recursive partitioning to impute values for the following variables: `sg`, `sz`, `age`, and `wt`.

We are also interested in the effect of the various covariates on the varying survival times. To determine potential effects, the covariates are plotted against the martingale residuals of a null model to reveal possible patterns and relationships that can be captured by adding the covariate to a Cox proportional hazard model.

We can see the summary information for the different treatment groups below:

```
##          placebo 0.2 mg estrogen 1.0 mg estrogen 5.0 mg estrogen
##              127             124             126             125
```

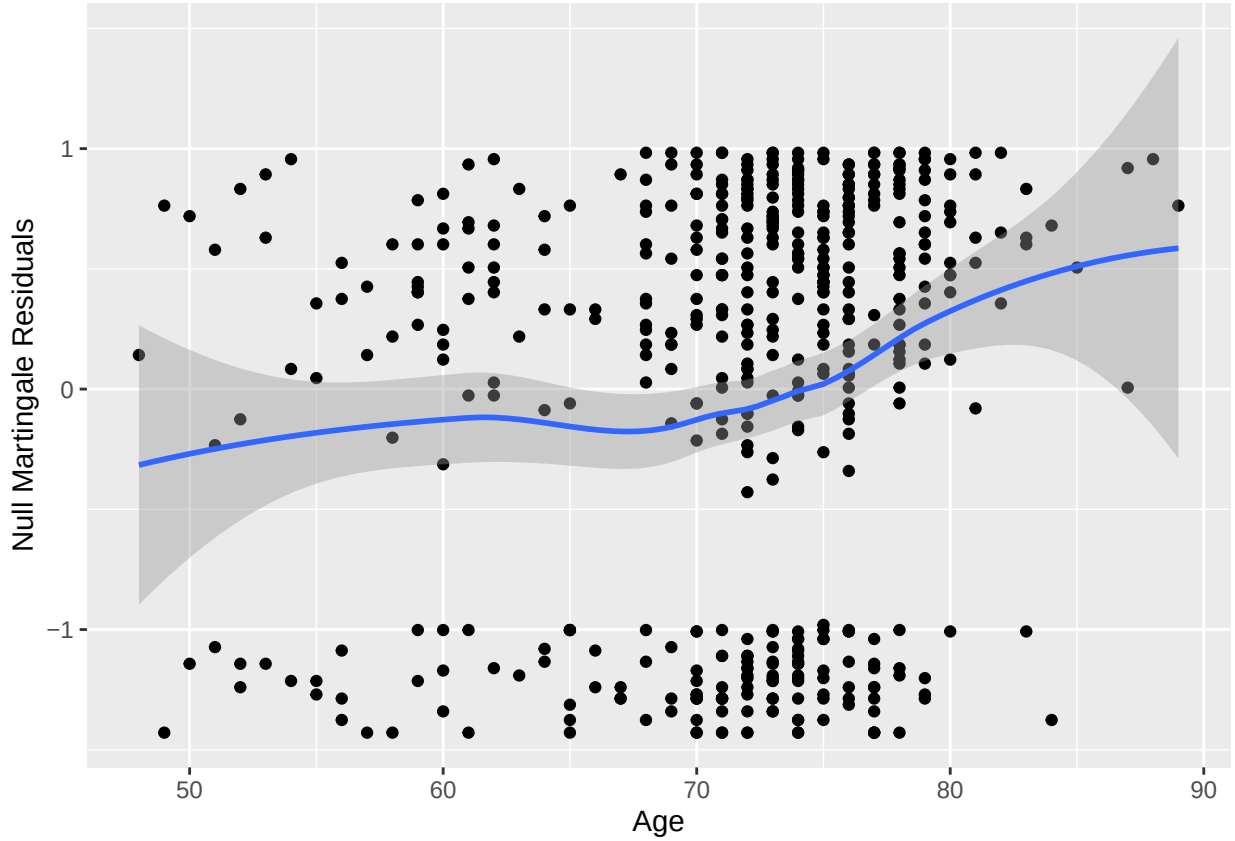
When examining a categorical variable, using boxplots can show differences in survival times. As we can see from the below plot, we can see the average values for each treatment appear to have differing effects on survival times, indicating `rx` should be included in our model.



We can see the summary statistics for the variable `age` after imputing values below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.00   70.00   73.00   71.46   76.00   89.00
```

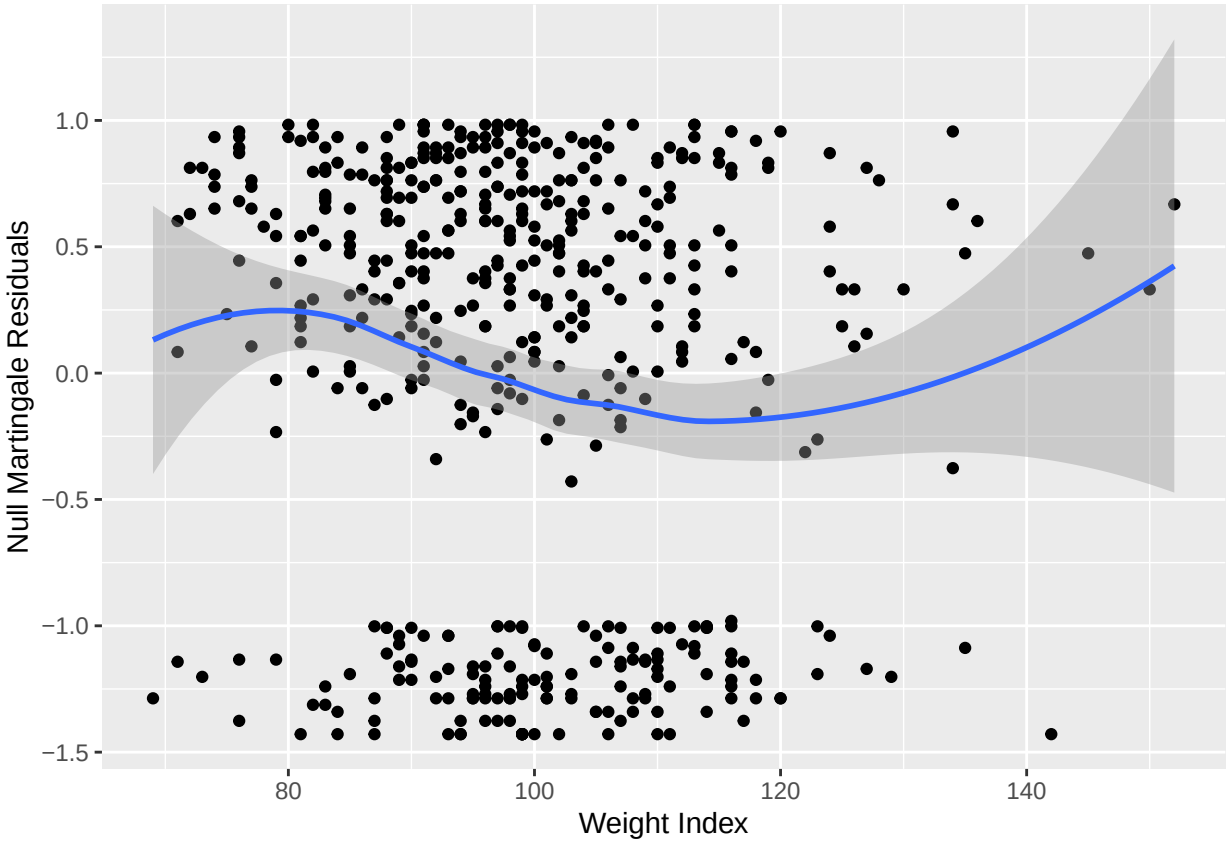
The below graph plots age against the null Martingale residuals. The blue line indicates the general pattern. The line is not flat and centered at 0, indicating age has an effect on survival times and is possibly non-linear.



We can see the summary statistics for the variable `wt` or weight index after imputing values below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   69.00   90.00   98.00   99.05  107.00  152.00
```

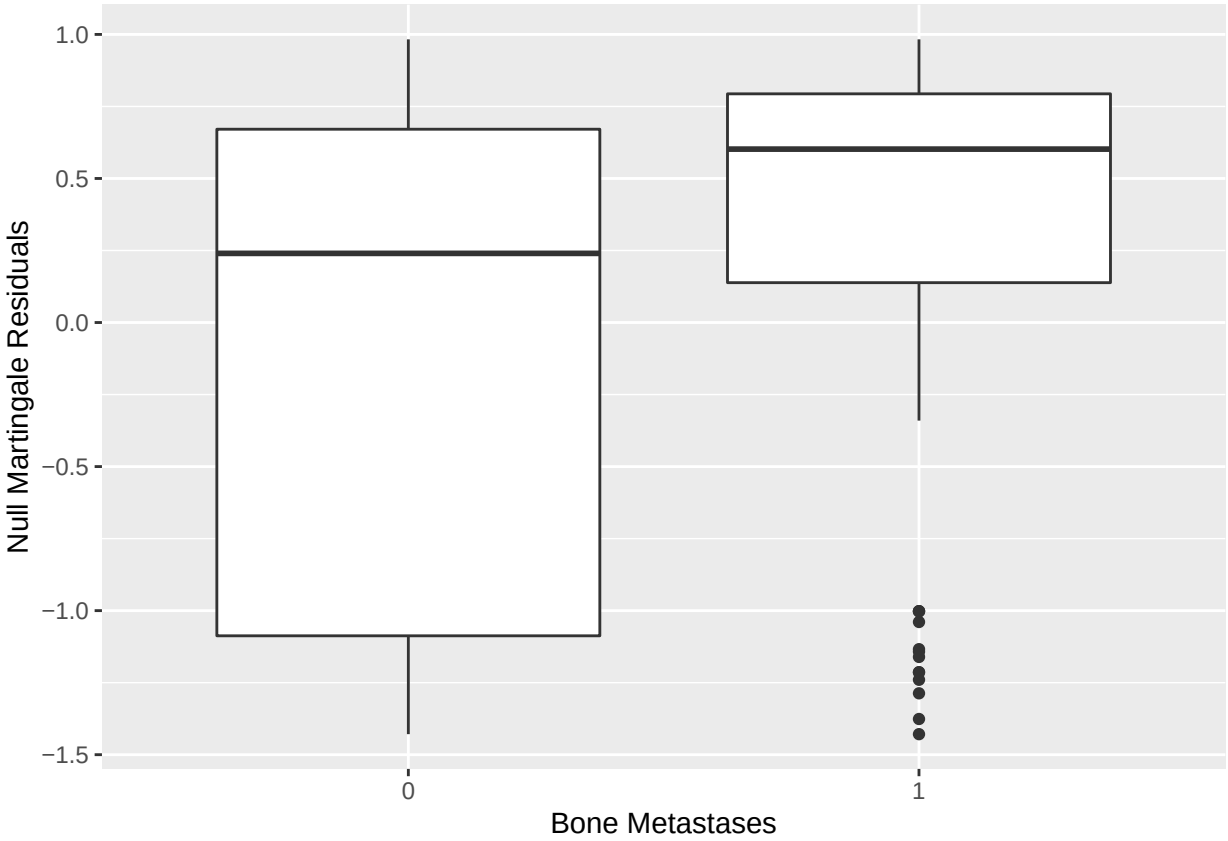
The below graph plots weight index against the null Martingale residuals. We can see from the blue line that `wt` has a clear effect on survival time and should be included in the model. Additionally it appears as though `wt` affects survival times in a non-linear fashion.



We can see the summary information for bone metastases below (0 indicates the cancer has not metastasized to the bones while a 1 indicates it has):

```
##   0   1
## 420  82
```

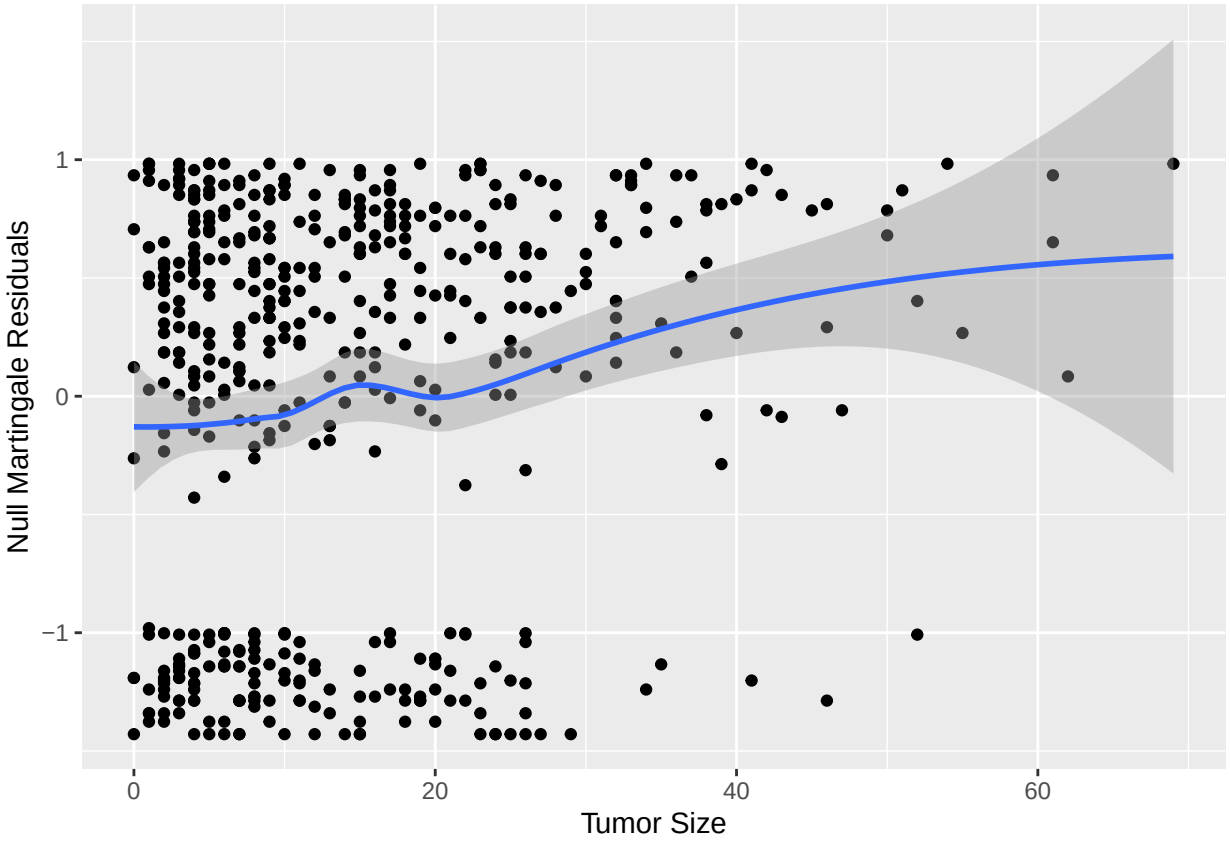
The below boxplot shows `bm` or bone metastases clearly has an effect on survival times and should be included in the model.



We can see the summary statistics for the variable `sz` or size of tumor after imputing values below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   11.00   14.61   21.00   69.00
```

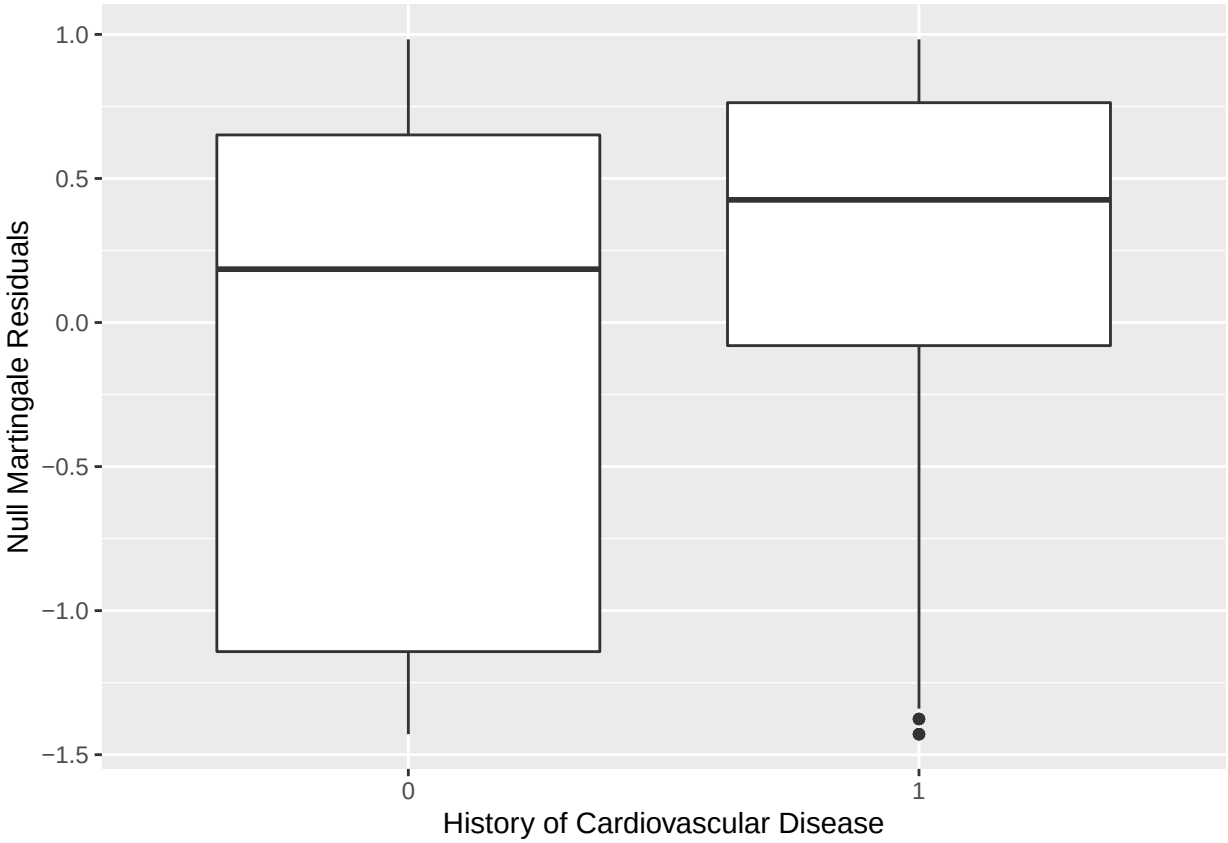
The below graph plots tumor against the null Martingale residuals. We can see from the blue line that `sz` has a clear effect on survival time and should be included in the model.



We can see the summary information for history of cardiovascular disease below (0 indicates no history of disease while a 1 indicates history of disease):

```
##   0   1
## 289 213
```

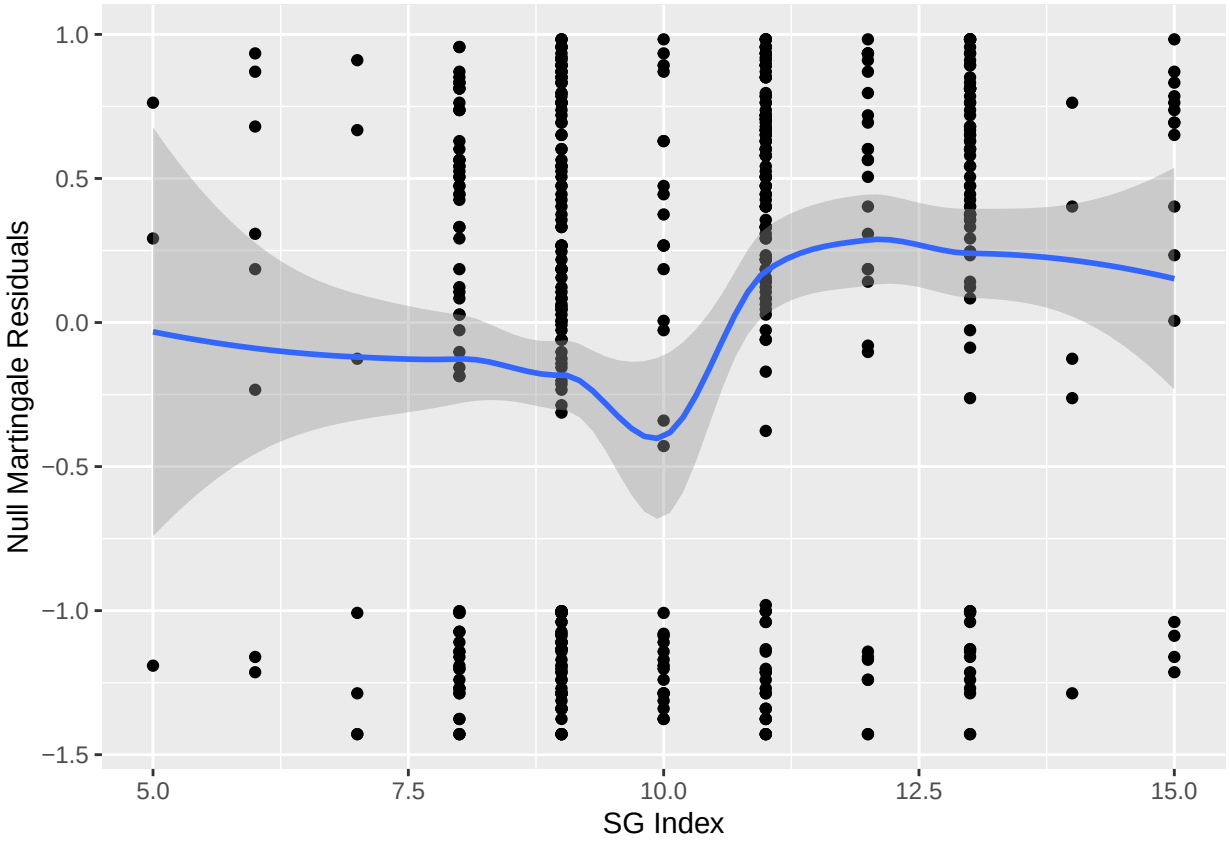
The below boxplot shows `hx` or history of cardiovascular disease clearly has an effect on survival times and should be included in the model.



We can see the summary statistics for the variable `sg` or stage and hist. index after imputing values below:

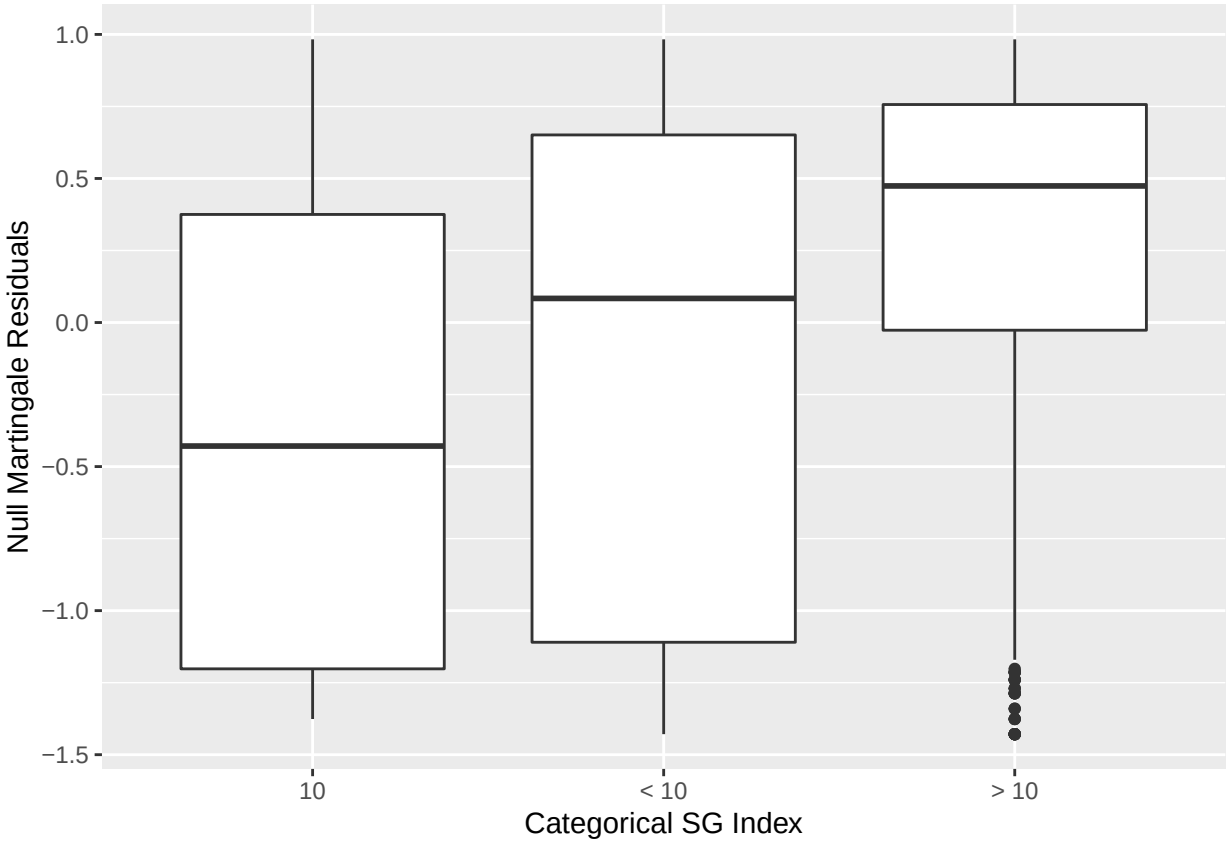
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00    9.00   10.00   10.32   11.75   15.00
```

The below plot clearly shows `sg` has an effect on survival times; however, there appears to be a significant grouping pattern:



Based on the above graph, it appears there are three distinct groupings. Values below 10, 10, and values above 10. Transforming this variable into a categorical variable may improve its predictive value.

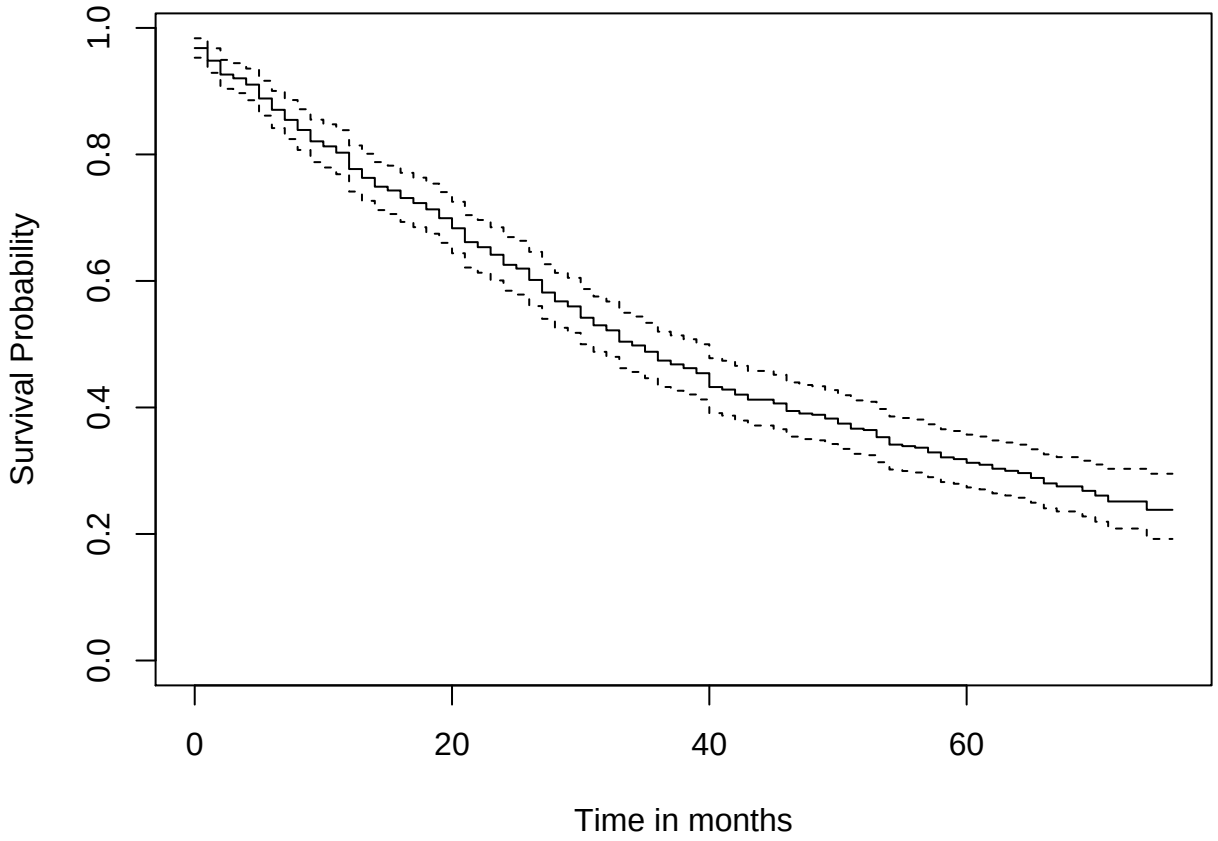
After creating a categorical version of this variable, we can see from the below boxplot it clearly has an effect on survival time and should be included in the model.



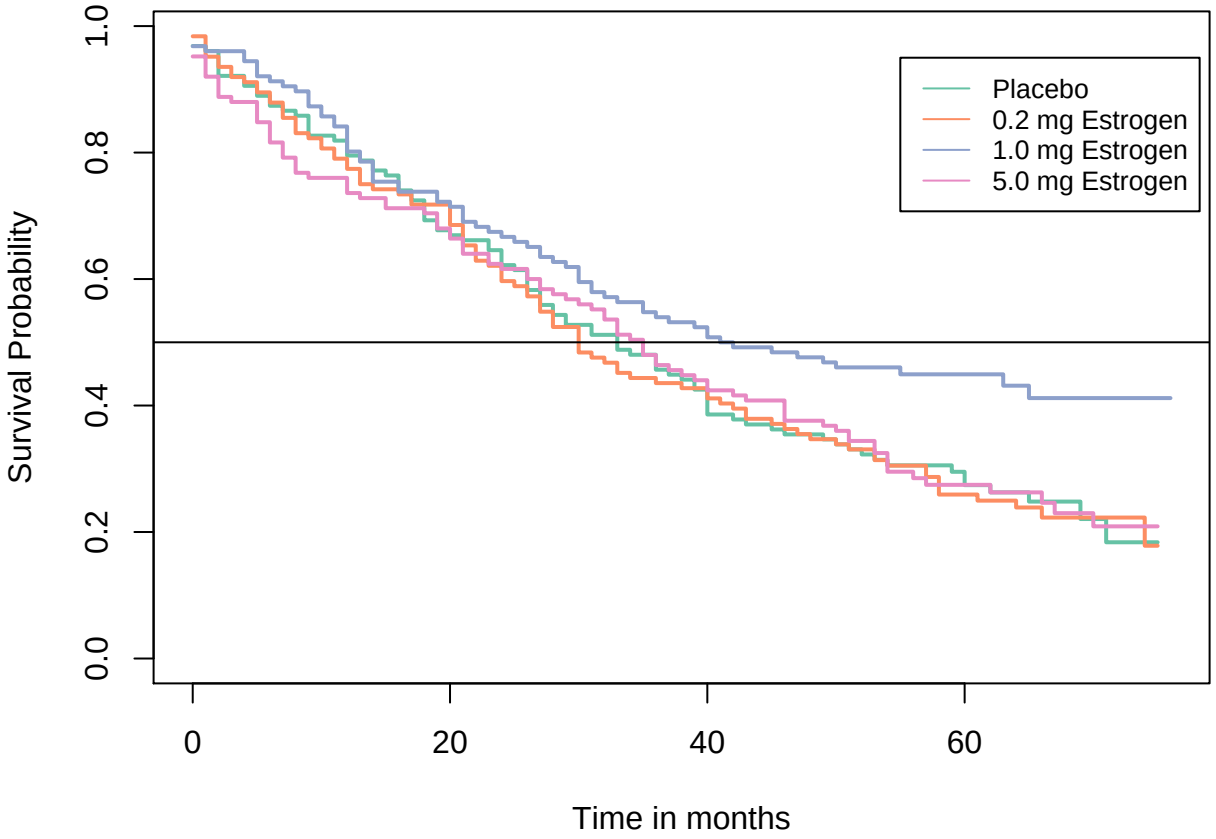
Kaplan Meier Non-Parametric Estimate

One method of estimating survival times is to use a Kaplan Meier Estimate. A disadvantage of this method is that it can only account for categorical variables and therefore is not capable of capturing some of valuable information that the Cox proportional hazard model is. An advantage is the visual nature of the estimates often make the information easier to communicate and more intuitive to understand.

The below plot shows the Kaplan Meier estimate of the entire study as well as the 95% confidence interval. Each tick downward indicates a patient death.



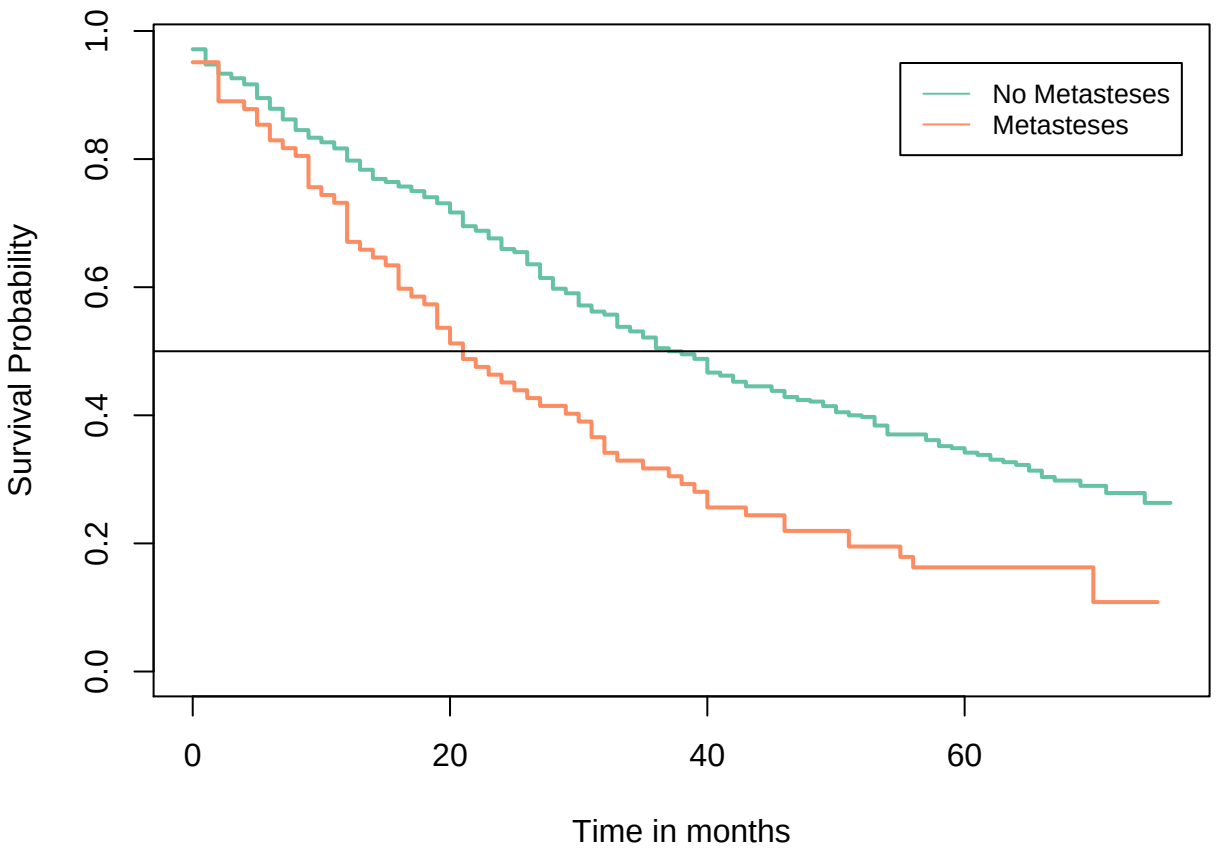
Taking this information we can then separate the survival functions by `rx` or treatment.



The above plot suggests the treatment 1.0 mg estrogen likely increases survival time. It is also worth noting the positive effects of the treatment do not appear to start working until after roughly 20 months.

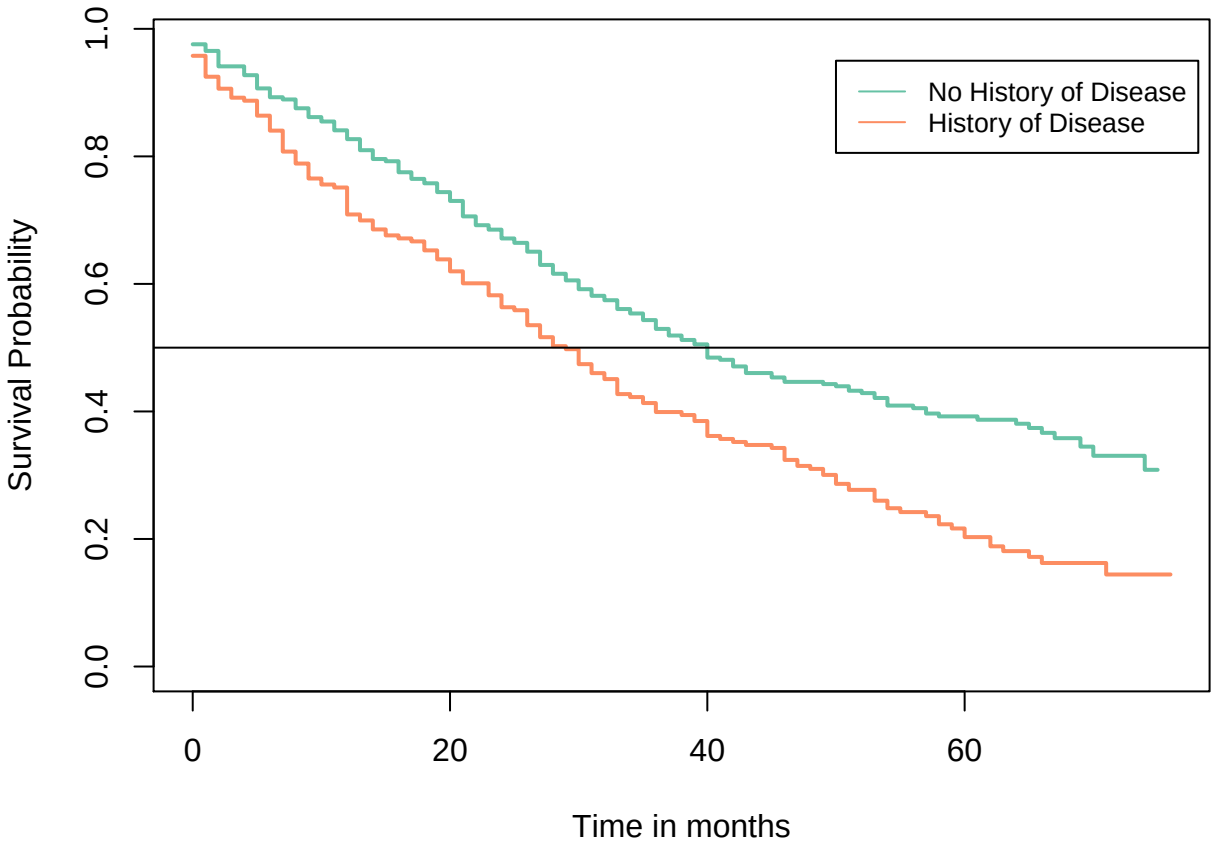
While there are slight variations in the survival times of the other treatments, it is unclear that the effects are significantly different from one another.

We can also examine the effect the variable `bm` or bone metastases has on survival times.



The above plot shows a clear difference in survival times between the two groups. Patients whose cancer has metastasized clearly have a much lower survival probability.

We can also examine the effect the variable `hx` or history of cardiovascular disease has on survival times.



The above plot shows a clear difference in survival times between the two groups. Patients who have a history of cardiovascular disease clearly have a much lower survival probability.

4. Cox Proportional Hazard Model Selection and Interpretation

Based on the above data characteristics section, it has been established there are clear correlations and patterns between the covariates and survival times.

Our final model includes the variables `rx`, `age`, `wt`, `bm`, `sz`, `hx`, and `sg`. Both the variables `age` and `wt` are included as non-linear terms using a penalized spline basis as the predictor. The variable `sg` was included as a categorical variable with the base level taken to be 10.

The model was built using the information on all 502 patients in the study. More detailed step by step information about how the model was built can be found in the Appendix section A. Plots of the Schoenfeld Residuals against each covariate can be found in Appendix section B. These plots are all centered at 0 indicating the proportional hazards assumption is satisfied.

The below information shows the coefficients, standard errors, and p-values of each predictor term.

```
## Call:
## coxph(formula = Surv(dtime, dead) ~ rx + pspline(i.age) + pspline(i.wt) +
##     bm + i.sz + hx + sgcat, data = pc)
##
##                            coef se(coef)      se2    Chisq   DF       p
## rx0.2 mg estrogen      -0.01378  0.14633  0.14614  0.00887 1.00  0.9250
## rx1.0 mg estrogen      -0.47012  0.16147  0.16127  8.47741 1.00  0.0036
```

```
## rx5.0 mg estrogen      -0.09851  0.15292  0.15216  0.41499 1.00  0.5194
## pspline(i.age), linear  0.02205  0.00772  0.00771  8.15141 1.00  0.0043
## pspline(i.age), nonlin                             7.65314 3.07  0.0568
## pspline(i.wt), linear  -0.01136  0.00406  0.00405  7.84386 1.00  0.0051
## pspline(i.wt), nonlin                              5.42255 3.08  0.1503
## bm1                     0.27491  0.15950  0.15897  2.97051 1.00  0.0848
## i.sz                    0.01538  0.00463  0.00461 11.01905 1.00  0.0009
## hx1                     0.49958  0.11231  0.11206 19.78789 1.00 8.7e-06
## sgcat< 10               0.27725  0.26113  0.26086  1.12721 1.00  0.2884
## sgcat> 10               0.59454  0.25897  0.25879  5.27078 1.00  0.0217
##
## Iterations: 5 outer, 14 Newton-Raphson
##      Theta= 0.879
##      Theta= 0.883
## Degrees of freedom for terms= 3.0 4.1 4.1 1.0 1.0 1.0 2.0
## Likelihood ratio test=102  on 16.1 df, p=2e-14
## n= 502, number of events= 354
```

When interpreting the model, one should note positive values of covariate coefficients indicate an increased likelihood of death while a negative coefficient indicates decreased likelihood of death. It is also important to note all coefficients are on a logarithmic scale and should be exponentiated for interpretation.

For example, the coefficient for tumor size is 0.0158. When we exponentiate this value, the result is 1.015499. This means holding all other variables constant, each increase of 1 cm sq in tumor size is associated with a 1.5499% increase in expected hazard.

With regards to treatment, the coefficient for 1.0 mg estrogen is -0.47012. When we exponentiate this value, the result is 0.6249. This means holding all other variables constant, receiving a treatment of 1.0 mg estrogen results in approximately a 37.5% decrease in hazard (increase in survival probability) relative to the placebo treatment.

Similarly to the Kaplan Meier estimate, it appears as if the 1.0 mg estrogen treatment is a significant predictor of survival probability. The high p-values of 0.2 mg estrogen and 0.5 mg estrogen indicate there is not statistical evidence to claim they affect survival probability.

In comparing the Kaplan Meier approach to the Cox Proportional Hazard approach, the Kaplan Meier estimate is easier to interpret visually and intuitively but cannot capture the effect of numerical variables as the Cox Proportional Hazard model can.

5. Summary and Concluding Remarks

In trying to predict survival times of prostate cancer patients, we find the variables `rx` (treatment), `age`, `wt` (weight index), `bm` (bone metastases), `sz` (tumor size), `hx`(history of cardiovascular disease), and `sg` (combined index of stage and hist. grade) to be significant predictors in a Cox proportional hazard model. Additionally with information from the Kaplan Meier estimate, we find a treatment of 1.0 mg estrogen to have a statistically significant effect in increasing survival probabilities while treatments of 0.2 and 5.0 mg do not have statistically significant evidence of being effective treatments. It is likely worth exploring if more recent data would show the same results. It would also be interesting to see the relative effectiveness of the estrogen treatments compared to chemotherapy.

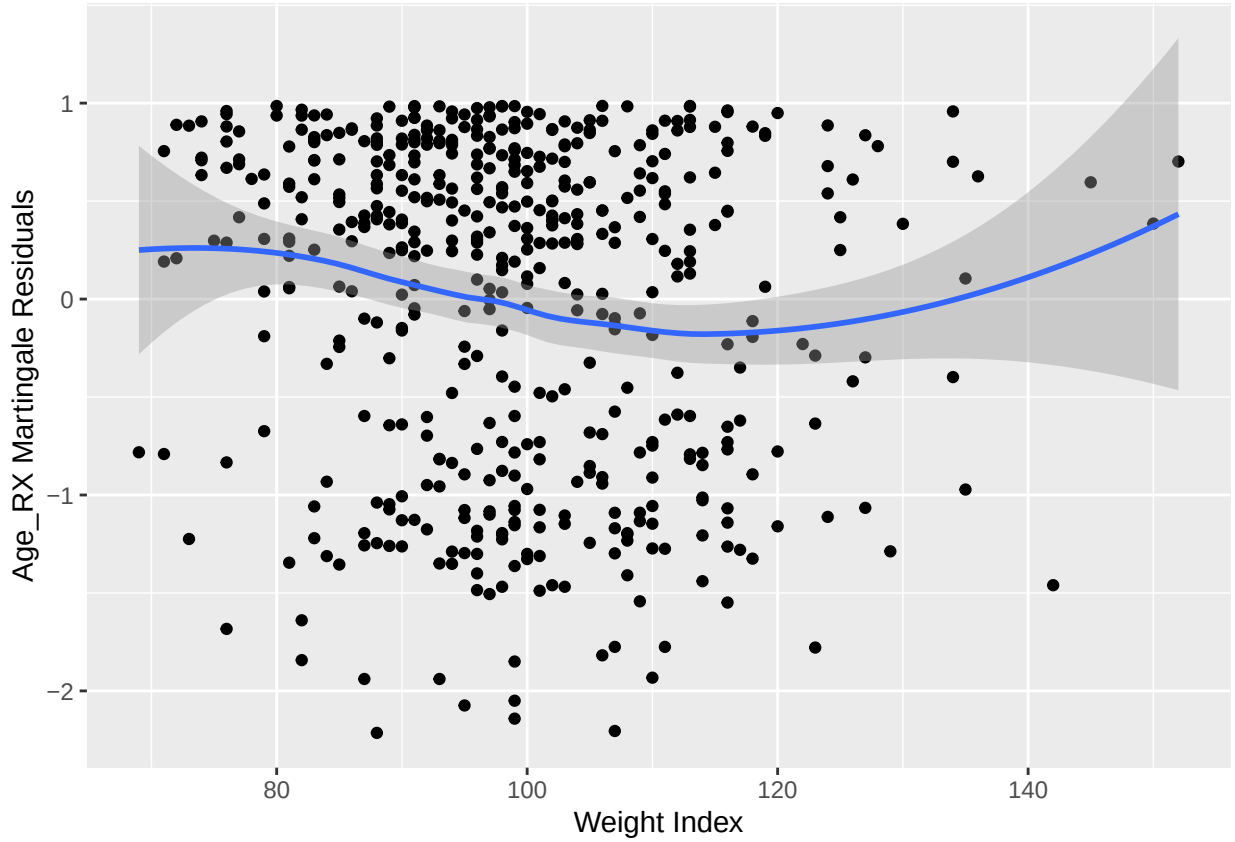
6. Appendix

Section A. Model Building Process

```
model0 <- coxph(Surv(dtime,dead)~rx + pspline(i.age), data = pc)

pc$mod0r <- residuals(model0, type = "martingale")
```

```
# WEIGHT VS agerx

ggplot(pc) +
  aes(x = i.wt, y = mod0r) +
  geom_point() +
  geom_smooth() +
  labs(x = "Weight Index", y = "Age_RX Martingale Residuals")
```

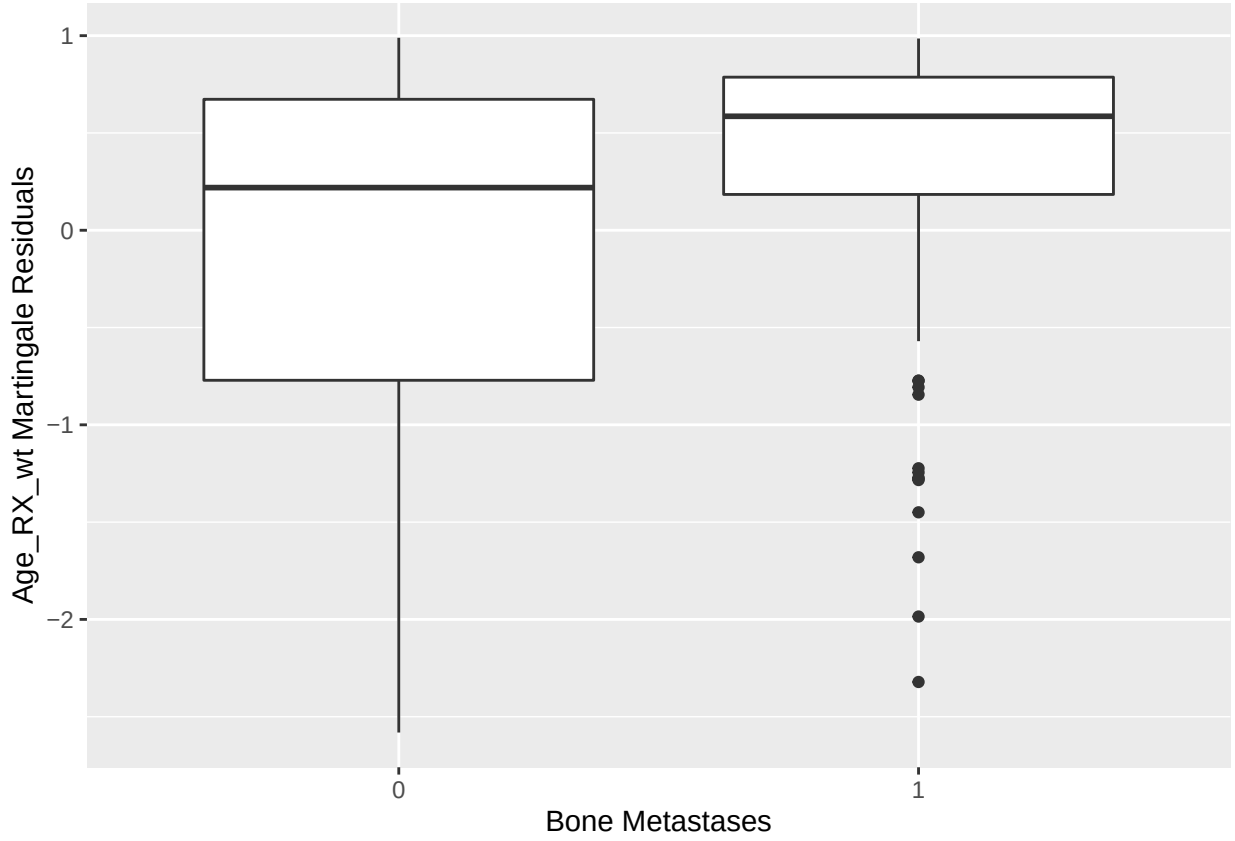
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
model1 <- coxph(Surv(dtime,dead)~rx + pspline(i.age) + pspline(i.wt), data = pc)
pc$mod1r <- residuals(model1, type = "martingale")

anova(model0,model1)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(dtime, dead)
##  Model 1: ~ rx + pspline(i.age)
##  Model 2: ~ rx + pspline(i.age) + pspline(i.wt)
##    loglik  Chisq     Df P(>|Chi|)
## 1 -1995.0
## 2 -1987.2 15.416 4.0702  0.004158 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# BM VS model with rx, age^, weight^

ggplot(pc) +
  aes(x = bm, y = mod1r) +
  geom_boxplot() +
  labs(x = "Bone Metastases", y = "Age_RX_wt Martingale Residuals")
```

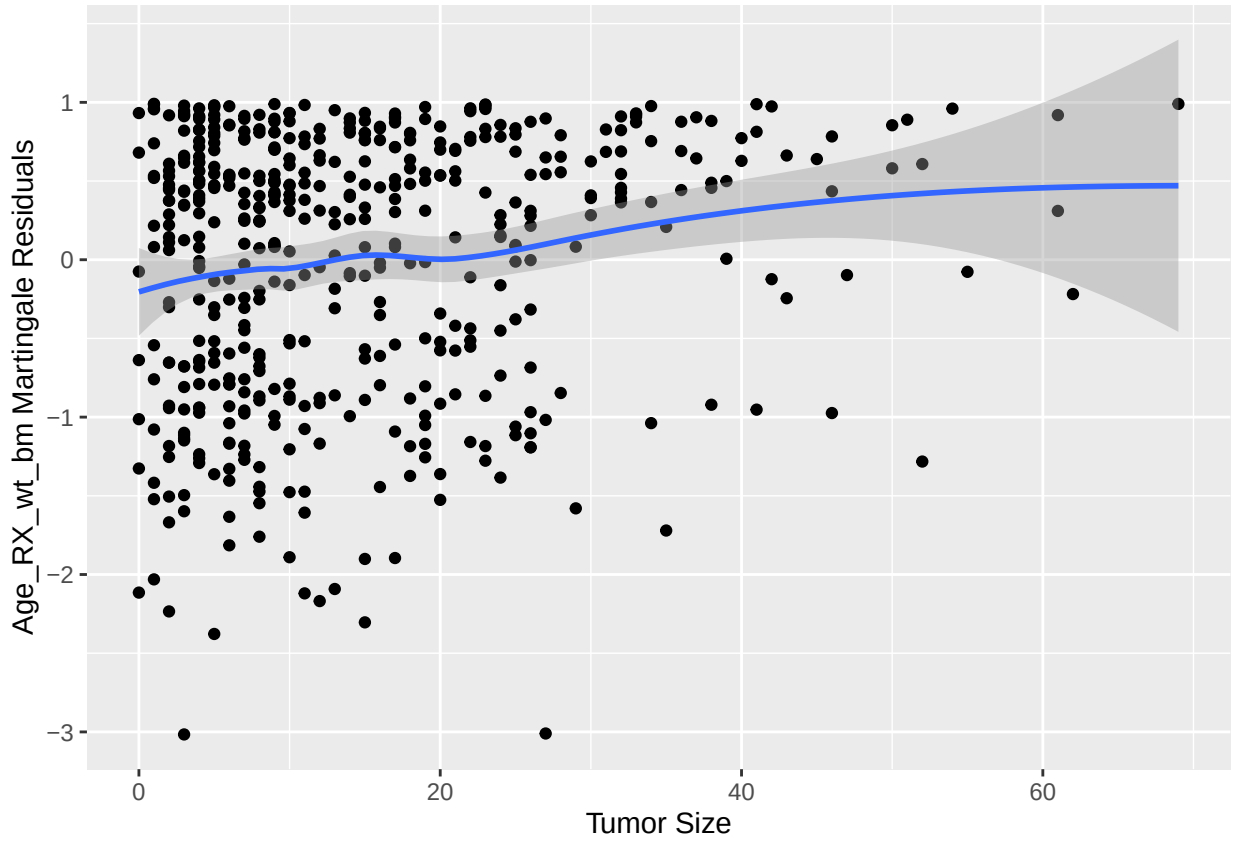



```
model2 <- coxph(Surv(dtime,dead)~rx + pspline(i.age) + pspline(i.wt) + bm, data = pc)
pc$mod2r <- residuals(model2, type = "martingale")

anova(model1,model2)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(dtime, dead)
##  Model 1: ~ rx + pspline(i.age) + pspline(i.wt)
```

```
##  Model 2: ~ rx + pspline(i.age) + pspline(i.wt) + bm
##    loglik  Chisq      Df P(>|Chi|)
## 1 -1987.2
## 2 -1981.2 12.055 0.99524 0.0005117 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Tumor size VS age_wt_bm_rx

ggplot(pc) +
  aes(x = i.sz, y = mod2r) +
  geom_point() +
  geom_smooth() +
  labs(x = "Tumor Size", y = "Age_RX_wt_bm Martingale Residuals")
```

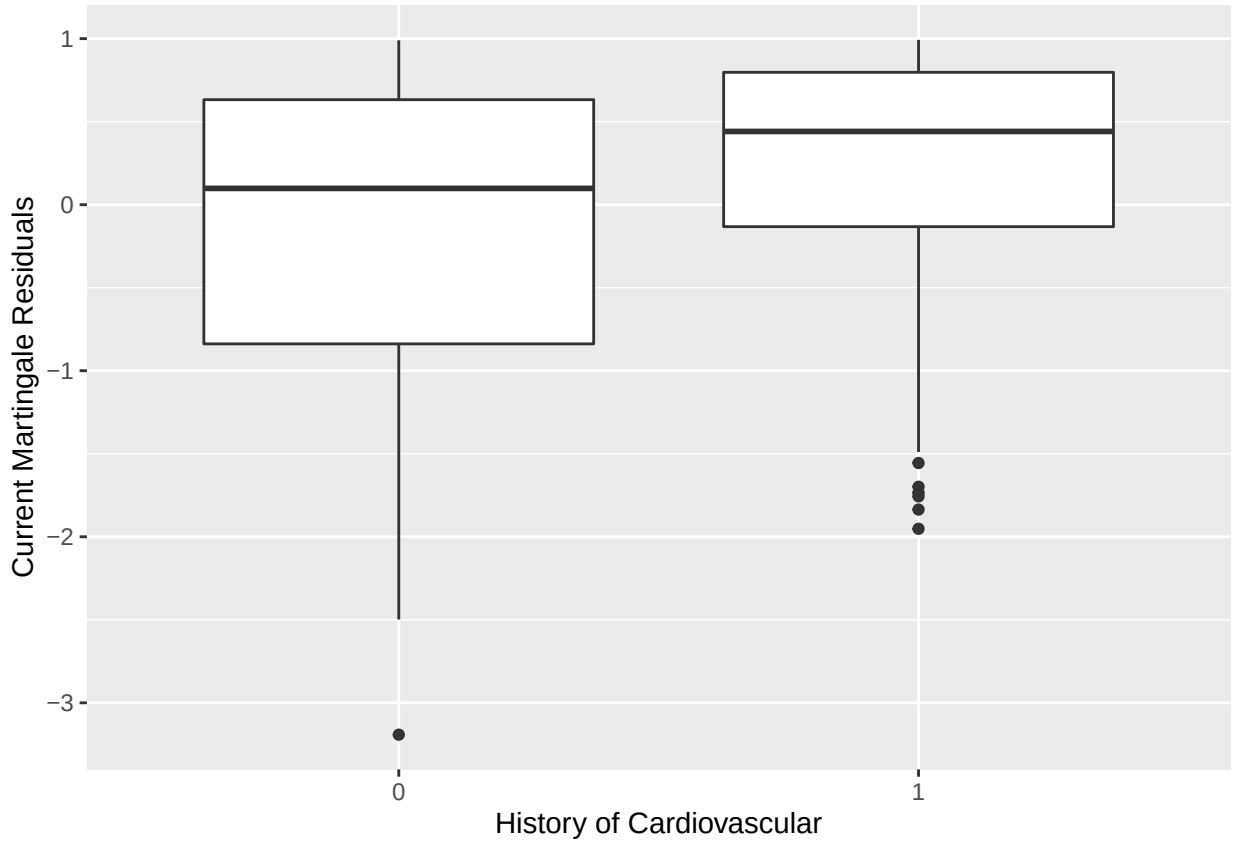



```
model3 <- coxph(Surv(dtime,dead)~rx + pspline(i.age) + pspline(i.wt) + bm + i.sz, data = pc)
pc$mod3r <- residuals(model3, type = "martingale")

anova(model2,model3)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(dtime, dead)
##  Model 1: ~ rx + pspline(i.age) + pspline(i.wt) + bm
##  Model 2: ~ rx + pspline(i.age) + pspline(i.wt) + bm + i.sz
##    loglik  Chisq      Df P(>|Chi|)
```

```
## 1 -1981.2
## 2 -1974.1 14.166 0.98569 0.0001626 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

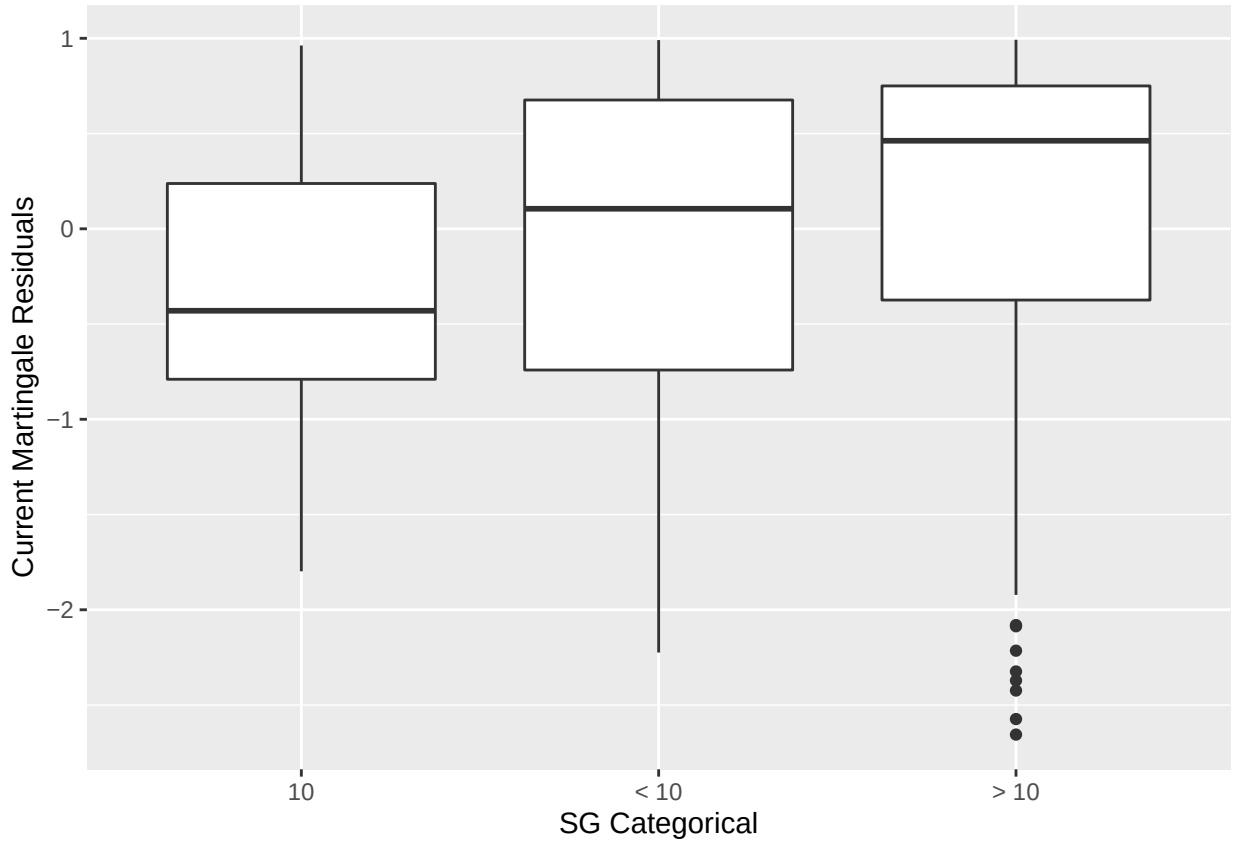
```
ggplot(pc) +
  aes(x = hx, y = mod3r) +
  geom_boxplot() +
  labs(x = "History of Cardiovascular", y = "Current Martingale Residuals")
```




```
model4 <- coxph(Surv(dtime,dead)~rx + pspline(i.age) + pspline(i.wt) + bm + i.sz + hx, data = pc)
pc$mod4r <- residuals(model4, type = "martingale")

anova(model3,model4)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(dtime, dead)
##  Model 1: ~ rx + pspline(i.age) + pspline(i.wt) + bm + i.sz
##  Model 2: ~ rx + pspline(i.age) + pspline(i.wt) + bm + i.sz + hx
##    loglik  Chisq      Df P(>|Chi|)
## 1 -1974.1
## 2 -1964.5 19.149 0.99891 1.206e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(pc) +
  aes(x = sgcat, y = mod4r) +
  geom_boxplot() +
  labs(x = "SG Categorical", y = "Current Martingale Residuals")
```

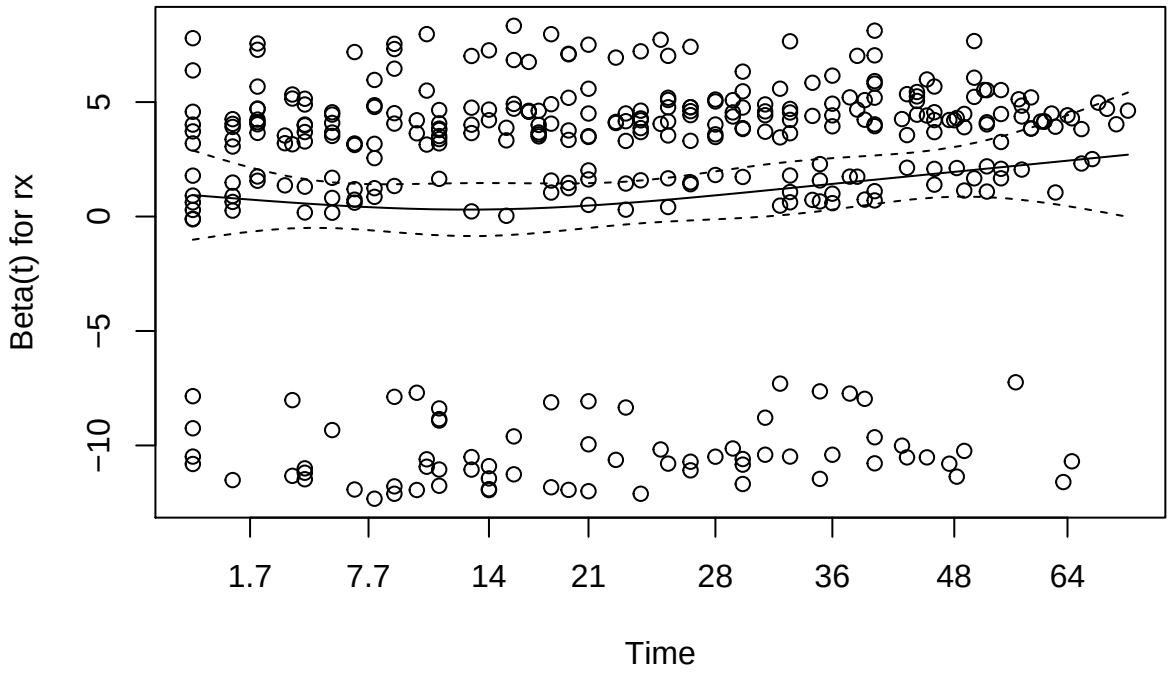


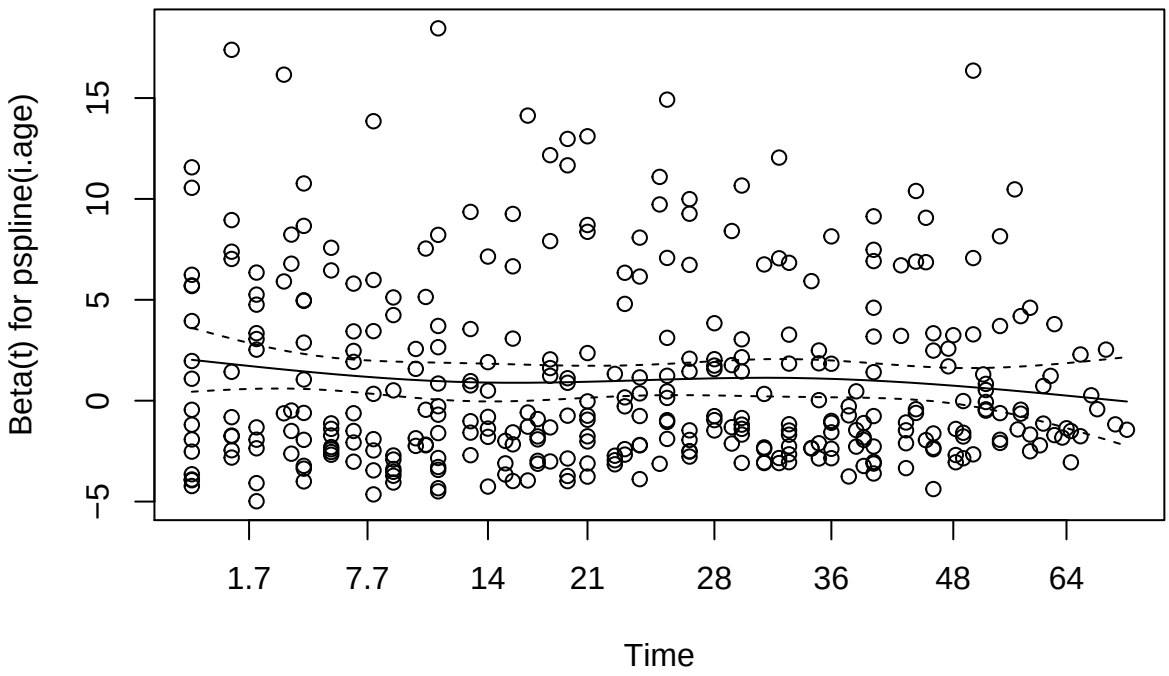

```
model5 <- coxph(Surv(dtime,dead)~rx + pspline(i.age) + pspline(i.wt) + bm + i.sz + hx + sgcat, data = pc
pc$mod5r <- residuals(model5, type = "martingale")

anova(model4,model5)
```

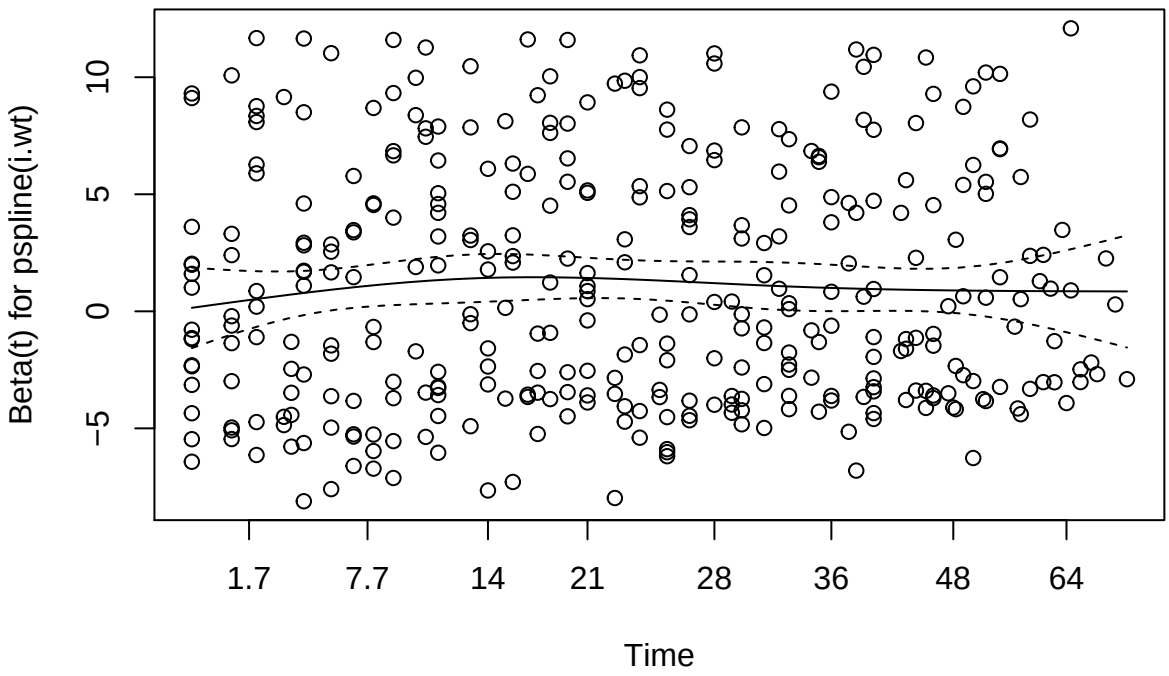
```
## Analysis of Deviance Table
##  Cox model: response is  Surv(dtime, dead)
##  Model 1: ~ rx + pspline(i.age) + pspline(i.wt) + bm + i.sz + hx
##  Model 2: ~ rx + pspline(i.age) + pspline(i.wt) + bm + i.sz + hx + sgcat
##    loglik  Chisq     Df P(>|Chi|)
## 1 -1964.5
## 2 -1959.7 9.6613 1.9857  0.007849 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

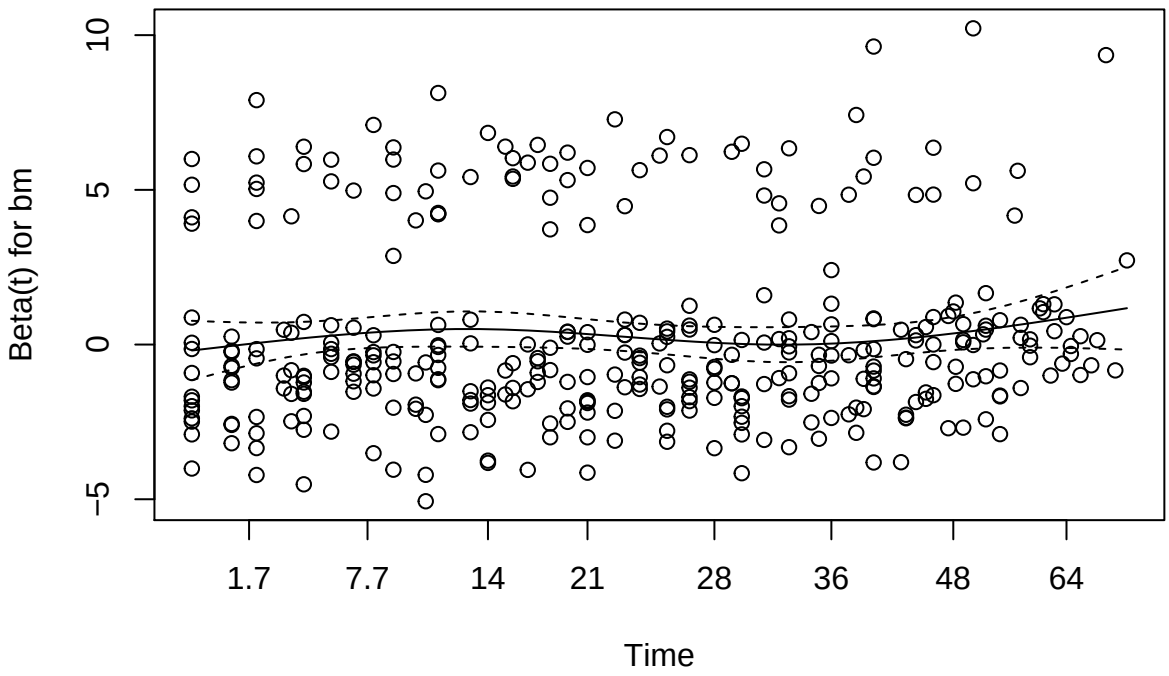
Section B. Schoenfeld Residual Plots

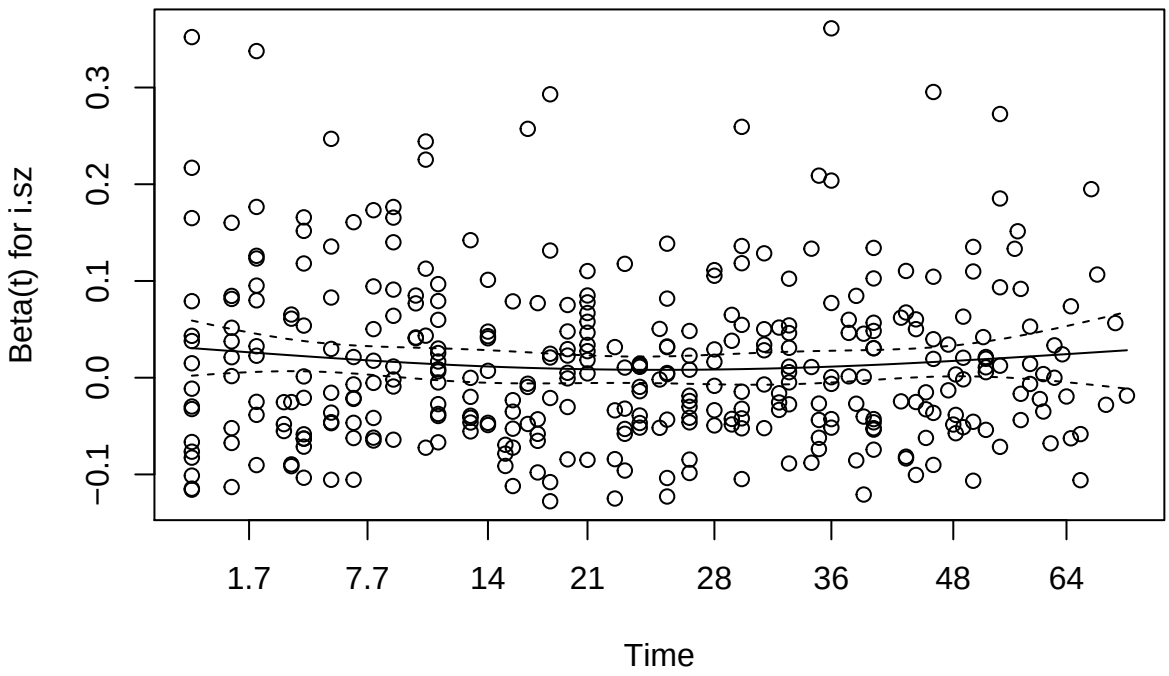
```r
p.h.asum <- cox.zph(model, transform = "km")

plot(p.h.asum)
```

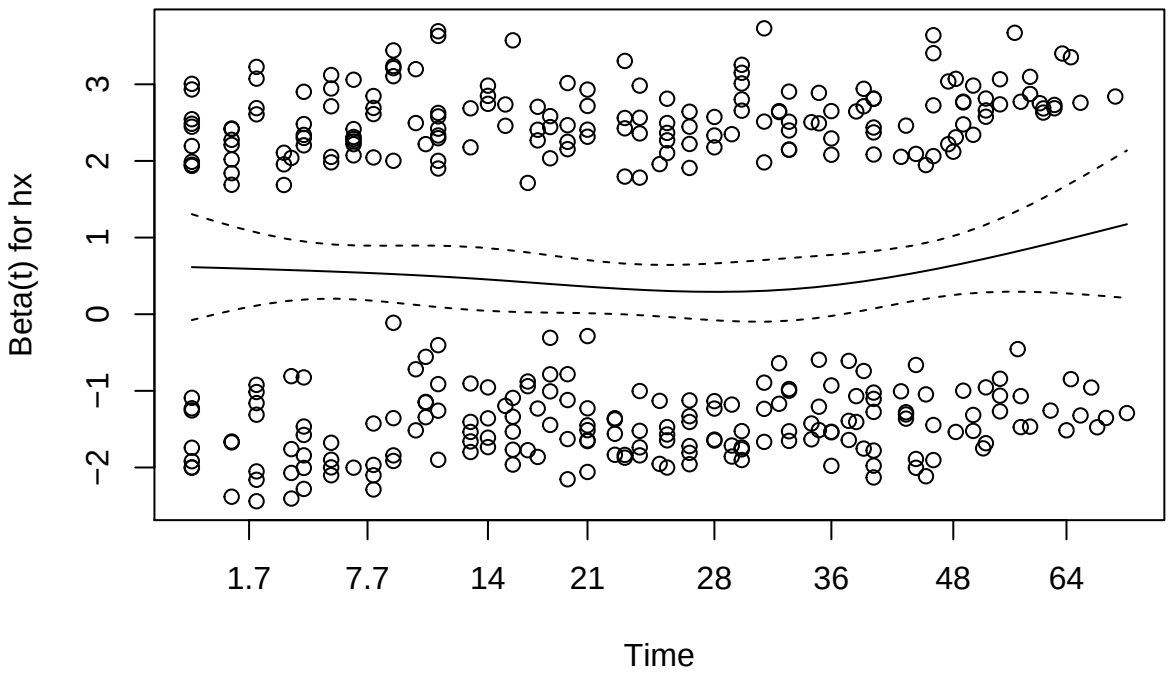


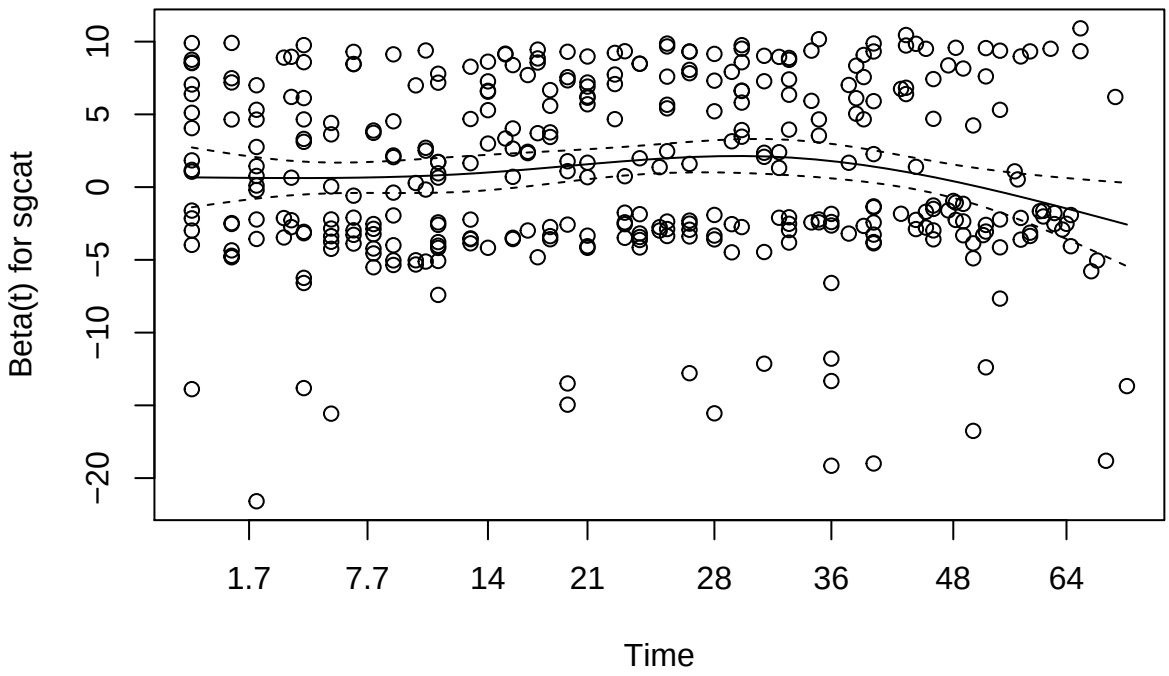

Beta(t) for pspline(i.age)
−5
0
5
10
15
1.7
7.7
14
21
28
36
48
64
Time


Beta(t) for pspline(i.wt)
10
5
0
−5
1.7
7.7
14
21
28
36
48
64
Time


Beta(t) for bm
10
5
0
−5
1.7
7.7
14
21
28
36
48
64
Time


Beta(t) for i.sz
−0.1
0.0
0.1
0.2
0.3
1.7
7.7
14
21
28
36
48
64
Time


Beta(t) for hx
3
2
1
0
−1
−2
1.7
7.7
14
21
28
36
48
64
Time


Beta(t) for sgcat
10
5
0
−5
−10
−20
1.7
7.7
14
21
28
36
48
64
Time