Nathanael Roy

Stat 255

Final Project: Mayo Clinic Data for PBC Treatment

Summary of Baseline Information

The data set for this project consists of data for a randomized controlled trial comparing the drug DPCA with a placebo in treating patients with Primary biliary cirrhosis of the liver. This disease is rare but fatal. While the randomized controlled trial is the gold standard and randomization should account for balance among treatment group versus control group we are still interested in analyzing the baseline covariates to see if just by chance some imbalance might have resulted in giving us results that may not reflect a normal population or if we can perhaps identify some heterogeneity in survival time among the various patient groups.

For each patient the outcome recorded in the data, our Y value essentially, consists of the time, or number of days between registration in the study and the time of death, censorship from transplantation, or censorship from the end of the study. The three possible statuses of what happened to the patient were also recorded in the data, although we only care about death or censorship and treat censorship as independent. Obviously, the patients were recorded for whether they were in the treatment group or the placebo group.

The first baseline covariate recorded was age. The patients have a range of ages from 26 years old to 78 years old. The mean age overall is about 50 while in the drug group the age is around 51 and in the placebo group the age is around 49. The distributions for the two split age groups also look roughly equivalent. There are a lot more females than males in the study with about 87% of the group given the drug being female and about 90% of those given the drug male.

There are a few discrete baseline covariates recorded for each patient. Ascites are abnormal buildup of fluid in the abdomen and are present in just about 8 percent of patients with about the same in each treatment group. Hepatomegaly is the symptom of enlarged liver and is present in just about half of patients, although around 56% of patients given the placebo had this symptom and only 46 percent of those in the treatment group. Spiders are a symptom of swollen blood vessels found in around 29% of patients in both treatment and control. Finally, edema is divided into three groups, one without edema, one with edema without diuretic therapy or resolved by diuretic therapy, and one group with edema despite diuretic therapy. This is in about equal numbers in treatment and non-treatment groups with just about 84% presenting with no edema.

Covariate	Presence in Treatment	Presence in Control	Overall
Acites	8.86%	6.49%	7.69%
Hepatomegaly	46.2%	56.5%	51.3%
Spiders	28.4%	29.2%	28.8%
Edema*	11.4%	10.7%	11.1%

*Treated as 50% true if coded as .5

Finally, there are 6 more baseline covariates collected for each patient that involve biomarkers. Cholesterol is a familiar biomarker though the dictionary does not say whether it is LDL or HDL cholesterol. The normal range for LDL (bad) cholesterol is less than 190 mg/dl while the normal for HDL is greater than 40-60 mg/dl. The range for of cholesterol in patients is 120 to 1775 which doesn't quite make much sense. According to the CDC from 1999-2000 18.3% of the population had LDL cholesterol above 240 mg/dl, but for our patient population we

have 80% having that high cholesterol. It is difficult to justify the values in the table. Though the number of very high values does drop off a bit so perhaps there was an issue with data collection.

Bilirubin levels range from .3 to 28 with the normal range going from .3 to 1.9, although a majority did fall under 1.9 so perhaps this value can be used reliably at least as an indicator of whether an individual had normal or lower bilirubin levels versus higher or abnormal bilirubin levels. Albumin on the other hand looks much more reliable as the normal range is from 3.4-5.4 g/dl while the data ranges from around 2 to 4.6.

Alkaline levels are normal from 53-128 U/L for 20 to 50 year old men and normal for 20 to 50 year old women for values from 42 to 98 U/L. The claim in the excel dictionary is that the units are correct but the range is from 289 to 13862.4. If we think there might be a transcription error and all the values are a decimal off we try dividing these values by 10 gives us more reasonable though still high values with the first quartile already at 87.15 and the median value being high for women at 125 U/L. But perhaps this is a feature of the population as a whole with a lot of patients having high alkaline levels.

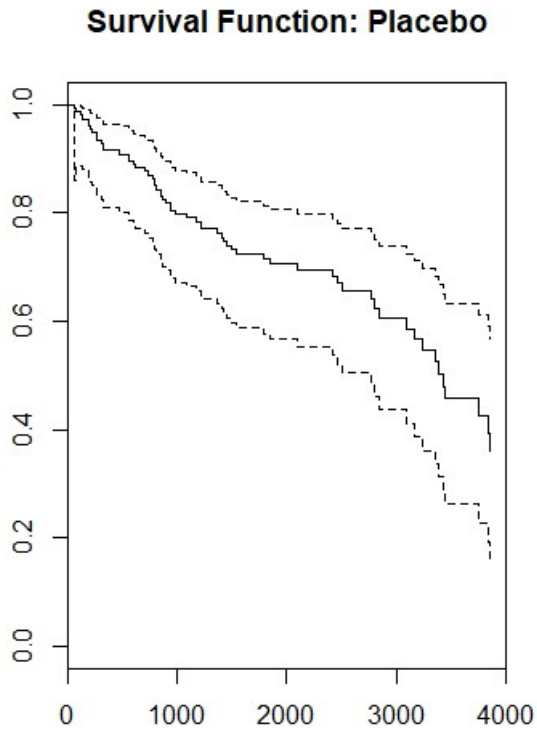
Platelets are measured in count per cubic ml/1000. While normal is measured in microliter and ranges from 150000 to 450000. With a range from 62 to 563 it seems platelets are actually measured in microliter divided by 1000 with the middle quartile then laying within the normal range.

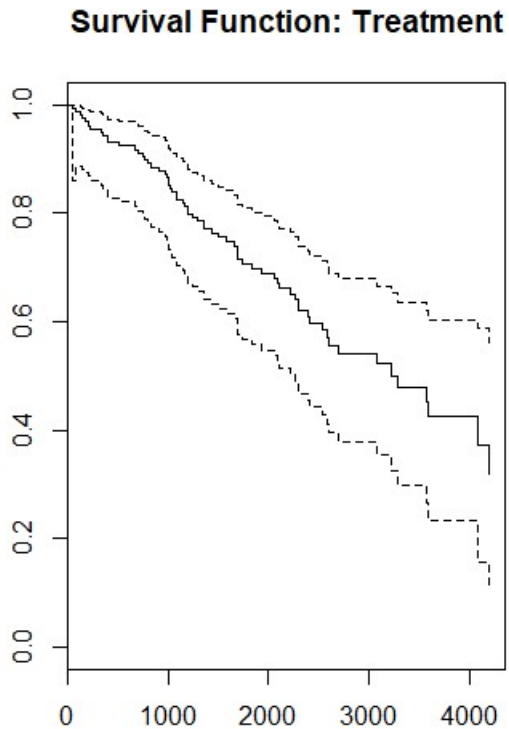
Finally, prothrombin is measured in time in seconds from 9 to 17 seconds, although half of the treated group are within 10-11 range. A typical result falls between 10 to 14 seconds so the interesting population might be the slower clotting times of those who took the test.

Summarizing this data:

Covariate	Min	Max	Median	Normal Range
Cholesterol	120 mg/dl	1775 mg/dl	309.5 mg/dl	Less than 190mg/dl
Bilirubin	.3 mg/dl	28 mg/dl	1.35 mg/dl	.3 to 1.9 mg/dl
Albumin	2 g/dl	4.6 g/dl	3.6 g/dl	3.4-5.4 g/dl
Alkaline	289 U/L	13862.4 U/L	1259 U/L	42 to 98 U/L (women)
Platelets	62000 /nanoL	563000 / nanoL	257000/ microL	150000 to 450000
Prothrombin	9 seconds	17 seconds	10.6 seconds	10 to 14 seconds

Overall (unadjusted) Comparison of Treatment Groups





The above graphs are the survival functions in time of days for the placebo and control groups with the kaplan meier confidence interval estimates. These survival functions do not look

different from each other and indeed when we make a comparison between the two we find the p value is .7 and so we conclude that the treatment has no meaningful effect on the survival of the patient.

Adjusted (for baseline covariates) Comparisons

We might consider tests adjusted for baseline covariates as well to see if we can tease out any significant difference between those treated with the drug and those treated with a placebo.

Stratified Comparisons

The survival functions do not look significantly different, but we might see if there is survival difference in the end or beginning of the trial by doing a weighted comparison. We might also try stratifying by age and sex, since clearly age is a major factor in how much longer you are going to survive and sex can have an impact on the particular biology of a person. In addition, we might consider stratifying on whether each variable is within the normal range. We run these tests of stratification on each of the discrete variables as well as each of the numeric variables and get the following results.

Test	Chi-Squared Value	P-Value
No adjustment	.1	.7
Weighted: Later Values	0	.9
Weighted: Earlier Values	.2	.7
Stratified: Age	0	.9
Stratified: Sex	.1	.8
Stratified: High Bilirubin	0	.8
Stratified: Low Albumin	0	.9

Stratified: Abnormal Platelet	0	.9
Stratified: High Prothrombin	.1	.7
Stratified: Ascites	.2	.7
Stratified: Hepatomegaly	.5	.5
Stratified: Spiders	.2	.7
Stratified: Edema	.2	.7

As we can see, none of these test are particularly enlightening, with the highest p value only reaching up to .5. It appears that the treatment is simply not significant

Developing a Cox Regression Model

The Cox regression model may be useful in assessing prognosis for an individual even without treatment effects and since we have a number of baseline covariates we are interested in fitting a Cox proportional hazard regression model to be able to give individuals a good idea of what we might be able to predict their survival probability over the next few years might be.

Before fitting the Cox Proportional hazard model, we need to deal with missing data for two of our covariates, cholesterol and platelet counts. How I addressed this was a four step process. First, I removed the missing data. Second, I fit a linear model to the data with the missing data as an independent variable. Third, using this new model I interpolated the missing data.

The problem with this method is that it might lead to lower estimate of variance within our data because our new data will follow a linear model exactly. To account for this, we should add some variance to our new predicted data. However, when we fit the model using a stepwise method we find that neither platelets nor cholesterol are included in the model. Our model includes the following: treatment (we add in by default in the end because we are interested),

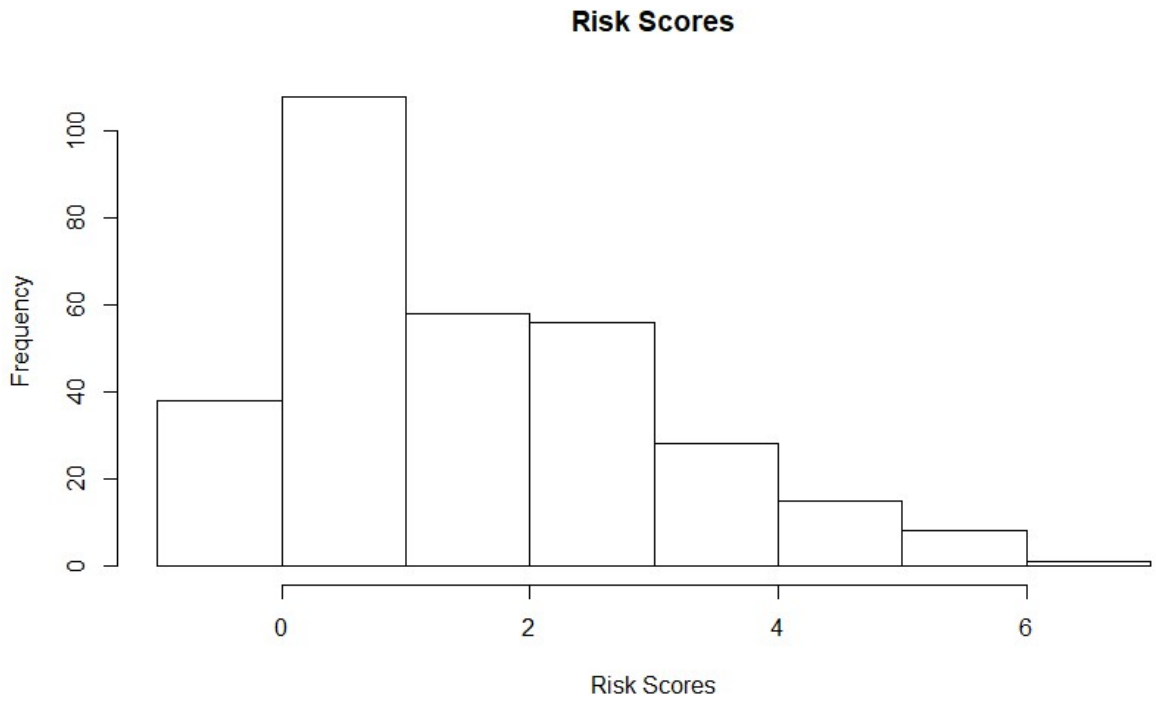
albumin, prothrombin, bilirubin, age, edema, sex, and an indicator variable of whether bilirubin is high.

Treatment Comparison in the Cox Model

In our Cox model the coefficient for treatment is very small and crosses 0. The p-value is .6 so we conclude, yet again, that treatment is not significantly different from placebo.

Other Analyses

One thing we could consider is breaking out patients into high and low risk groups. We do this by assigning a risk score based on the coefficients fit in the proportional hazard model. For the moment we remove treatment so that we can get a baseline prognosis without treatment. We then divide patients to those who, once the coefficients are added together and multiplied with the respective variable, have a risk score of 3 or above. We chose this so that we had a reasonable cutoff based on the histogram of risk scores:



From a subjective standpoint, it looks like the bulk of the patients are at a risk score of below 3. After running a baseline difference on just those with a risk score of higher than 3 we find that we get a p value of .1. This is still not significant but does give us some indication that perhaps a well targeted population of high risk patients may be good candidates for the drug.

Conclusion and Discussion

Our drug does no better than placebo in helping with the survival of patients in the trial. Several of the baseline covariates seem to be useful to us in giving us a decent idea of prognosis for patients over a time frame of several years. This information may be useful in deciding who might be a good candidate for transplantation. However, treatment with DPCA will likely not be much help for patients as it performs no better than placebo and will only result in costs of money and extra side-effects to the patients.

Appendix: Some Code Used to Analyze the Data

```
library(readxl)

Mayo_Clinic_PBC_trial_data_restricted <- read_excel("C:/Users/Acer/Desktop/Mayo Clinic PBC trial data restricted.xlsx")

delta = as.numeric(Mayo_Clinic_PBC_trial_data_restricted$status==2)

Mayo_Clinic_PBC_trial_data_restricted$delta = delta

#This function gives us the percentage of patients who have each symptom in overall population,

#control group, and treatment

findpct = function(a){

 overall = mean(a) #overall

 drug = mean(a[which(treatment==1)]) #drug

 control = mean(a[which(treatment==2)]) #control

 val = c(overall,drug,control)

 names(val) = c("overall","drug","control")

 return(val)

}

#analyzing each baseline symptom (along with sex)

findpct(sex)

findpct(ascites)

findpct(hepatomegaly)

findpct(spiders)

findpct(edema)

hist(ascites)

hist(hepatomegaly)

hist(spiders)

hist(edema)

library(km.ci)

library(KMsurv)

library(survival)
```

```
library(My.stepwise)

attach(Mayo_Clinic_PBC_trial_data_restricted)

#The following gives us which within the data are not censored. The rest are considered indpendently censored at the time given.

#Took a look at this to get a quick sense of the baseline data

summary(Mayo_Clinic_PBC_trial_data_restricted)

#The following were used to create the survival functions for drug and placebo along with K-M intervals

surv.trt1 = Surv(time[treatment==1],delta[treatment==1])

surv.trt2 = Surv(time[treatment==2],delta[treatment==2])

fit1 = survfit(surv.trt1~1)

fit2 = survfit(surv.trt2~1)

band.1 = km.ci(fit1,method="logep")

plot(band.1,main="Survival Function: Treatment")

band.2 = km.ci(fit2,method="logep")

plot(band.2,main="Survival Function: Placebo")

sv = Surv(time,delta)

#Before fitting the Cox model we have to deal with some missing data. We might try fitting cholesterol and platelets for the missing data

#based on some of the other values such as age, sex, and bilirubin. For this missing data we will simply

my.variable.list =
c("sex","age","ascites","hepatomegaly","spiders","edema","bilirubin","albumin","alkaline","prothrombi
n")

trunc.cholest =
Mayo_Clinic_PBC_trial_data_restricted[which(Mayo_Clinic_PBC_trial_data_restricted$cholesterol!="."),
]

trunc.cholest$cholesterol = as.numeric(trunc.cholest$cholesterol)

trunc.platelets =
Mayo_Clinic_PBC_trial_data_restricted[which(Mayo_Clinic_PBC_trial_data_restricted$platelets!="."),]

trunc.platelets$platelets = as.numeric(trunc.platelets$platelets)

#My.stepwise.lm(Y="cholesterol",variable.list = my.variable.list,data=trunc.cholest)

#Gives us a model that includes bilirubin, intercept, edema, age, alkaline, and prothrombin
```

```
model.cholest = lm(cholesterol~1+bilirubin+edema+age+alkaline+prothrombin,data=trunc.cholest)

#replaces missing data for cholesterol using the linear model.

Mayo_Clinic_PBC_trial_data_restricted$cholesterol[which(Mayo_Clinic_PBC_trial_data_restricted$cholesterol==".")] =
predict(model.cholest,Mayo_Clinic_PBC_trial_data_restricted[which(Mayo_Clinic_PBC_trial_data_restricted$cholesterol=="."),])

Mayo_Clinic_PBC_trial_data_restricted$cholesterol =
as.numeric(Mayo_Clinic_PBC_trial_data_restricted$cholesterol)

My.stepwise.lm(Y="platelets",my.variable.list,data=trunc.platelets)

model.platelets =
lm(platelets~1+edema+hepatomegaly+alkaline+prothrombin+albumin+sex,trunc.platelets)

Mayo_Clinic_PBC_trial_data_restricted$platelets[which(Mayo_Clinic_PBC_trial_data_restricted$platelets==".")] =
predict(model.platelets,Mayo_Clinic_PBC_trial_data_restricted[which(Mayo_Clinic_PBC_trial_data_restricted$platelets=="."),])

#replaces missing data for platelets

Mayo_Clinic_PBC_trial_data_restricted$platelets =
as.numeric(Mayo_Clinic_PBC_trial_data_restricted$platelets)

Mayo_Clinic_PBC_trial_data_restricted$older= as.numeric(age > median(age))

Mayo_Clinic_PBC_trial_data_restricted$high.bili = as.numeric(bilirubin > 1.9)

Mayo_Clinic_PBC_trial_data_restricted$low.alb = as.numeric(albumin < 3.4)

Mayo_Clinic_PBC_trial_data_restricted$plate.abnormal = as.numeric(platelets < 150000 | platelets > 450000)

Mayo_Clinic_PBC_trial_data_restricted$prothrombin.over = as.numeric(prothrombin>14)

Mayo_Clinic_PBC_trial_data_restricted$prothrombin.under = as.numeric(prothrombin < 10)

test1 = survdiff(sv~treatment, rho = 0)

#Tests for difference weighted to earlier and later as well as stratifying for a variety of factors

test2 = survdiff(sv~treatment, rho = 1)

test3 = survdiff(sv~treatment, rho = -1)

test4 = survdiff(sv~treatment+strata(older))

testsex = survdiff(sv~treatment+strata(sex))

test5 = survdiff(sv~treatment+strata(high.bili))
```

```
test6 = survdiff(sv~treatment+strata(low.alb))

test7 = survdiff(sv~treatment+strata(plate.abnormal))

test8 = survdiff(sv~treatment+strata(prothrombin.over))

test9 = survdiff(sv~treatment+strata(ascites))

test10 = survdiff(sv~treatment+strata(hepatomegaly))

test11 = survdiff(sv~treatment+strata(spiders))

test12  = survdiff(sv~treatment+strata(edema))

#Creating a Cox model

my.variable.list =
c("low.alb","plate.abnormal","prothrombin.over","high.bili","treatment","sex","age","ascites","hepato
megaly","spiders","edema","platelets","cholesterol","bilirubin","albumin","alkaline","prothrombin")

My.stepwise.coxph(Time = "time", Status = "delta", variable.list = my.variable.list,

            data = Mayo_Clinic_PBC_trial_data_restricted)

My.stepwise.coxph(Time = "time", Status = "delta", variable.list = my.variable.list, in.variable =
c("treatment"),

             data = Mayo_Clinic_PBC_trial_data_restricted)

#The above was run to get the final model:

library(MASS)

cox.model = coxph(formula = Surv(time, delta) ~ + high.bili + albumin + prothrombin + bilirubin + age +
edema, data = Mayo_Clinic_PBC_trial_data_restricted, method = "efron")

cox.model.trt = coxph(formula = Surv(time, delta) ~ treatment + high.bili + albumin + prothrombin +
bilirubin + age + edema, data = Mayo_Clinic_PBC_trial_data_restricted, method = "efron")

summary(cox.model)

#Just to check to compare the model with and without treatment

anova(cox.model,cox.model.trt)

#Other analysis: Created “risk factor” data

risk.factors <- cbind(high.bili,albumin,prothrombin,bilirubin,age,edema)

risk.score = rep(0,312)

#calculating the risk score for each person

for(i in 1:312){
```

```
  risk.score[i] = sum(cox.model$coefficients*risk.factors[i,])
}
risk.score
sv.highrisk = Surv(time[risk.score>3],delta[risk.score>3])
sv.lowrisk = Surv(time[risk.score<=3],delta[risk.score<=3])
fithigh = survfit(surv.highrisk~1)
fitlow = survfit(surv.lowrisk~1)
band.1 = km.ci(fithigh,method="logep")
plot(band.1,main="Survival Function: High Risk")
band.2 = km.ci(fitlow,method="logep")
plot(band.2,main="Survival Function: Low Risk")
testhigh = survdiff(sv.highrisk~treatment[risk.score>3])
testlow = survdiff(sv.lowrisk~treatment[risk.score<=3])
```