Survival Analysis on Linking Drug Dependent Adults to Primary Care

Introduction

Note: might want to think more about the TITLE, after finishing the doc. Is it informative, specific and precise?

BP1: What is survival analysis? What does it do?

Survival analysis estimates the expected duration of time until an event occurs. An event may be death, motorcycle breakdown, etc. In this survival analysis project, we use the dataset HELPfull, which contains data from the paper "Linking alcohol- and drug-dependent adults to primary medical care: a randomized controlled trial of a multi-disciplinary health intervention in a detoxification unit."

BP2: What is the dataset and key variables? What is the purpose of the research?

The HELP study measures the utilization of primary medical care of adult patients from a detox program. Some patients receive a combination of treatment that is intended to increase their primary care usage. Some of the key variables include group (treatment variable), dayslink (response variable), linkstatus (censoring variable), commonly used demographic variables (gender, age, education level), and medical variables (drug usage, mental health index, etc.).

BP3: How are the variables coded? What do you do with factor variables? Lots of missing, what should be done about that? Which variables are most interesting?

By default, the variables are encoded as strings. In order to make plotting easier, we encoded the categorical variables as factors. That way, ggplot can distinguish each group distinctly. We drop the missing data. This is a requirement to use the censboot function.

BP4: why do we want to research anything about the HELPfull data? What is its relevance?

The research is important, because it may provide insights that improve the treatment of detox programs. If more patients from detox programs can ask for medical help more often and sooner with nudging like the treatment, then it might be possible to address this societal issue more effectively. In addition, significant and positive results may give researchers more confidence to research applicable solutions to other social issues, such as homeslessness, bankruptcy, etc.

BP5: What are some of the initial goals we have in mind? What do we plan to do (exploration of variables, focus on a specific target, building a model, new thing)? Is the final paragraph a brief description of the hypothesis/goals and findings of the paper?

The primary goal is to build a Cox PH model that best predicts survival outcome (using primary healthcare). Our initial hypothesis is the treatment is effective at getting patients connected with primary healthcare. We want to first visualize and understand the variables in the dataset, then look for some interesting and potentially important variables to include in the final model. We will then refine the final model and apply additional techniques to examine the model, such as cox.zph function and bootstrapping techniques.

Methods

Note: The methods should be a source of detail about the approaches of the authors. Procedures that have been repeated by the authors should only be listed once. Variations to the procedure should be briefly summarized.

Our goal is to create a model that can predict survival outcome. Since the dataset and code are included and shared, we believe the work can be replicated. We will first visualize some data (including mutating some variables), figure out the variables we want to use, and then build the model.

Exploratory Data Analysis

For our exploratory data analysis, we will observe relationship the explanatory variables have with the response variables, linkstatus. Analyzing the relationship between the explanatory and response variable will help select the variables we want to include in the final model.

First, we will import the packages and dataset.

```
library(mosaic)
library(readr)
library(tidyverse)
library(broom)
library(survival)
library(survminer)
library(praise)
library(boot)
# import the dataset HELPFUL. Encode NAs as "*"
df <- read_csv("HELPdata.csv", na="*")
```

Now we will import the dataset and encode our variables into factors.

```
df <- df %>%
  mutate(yrs_education = as.numeric(a9),
         gender=a1,
         alcq_30 = as.numeric(alcq_30),
         marriage = as.factor(a10),
         employment = as.factor(a13),
         income = as.factor(case_when(a18 == 1 ~ "<5000",
                                      a18 == 2 ~ "5000-10000",
                                      a18 == 3 ~ "11000-19000",
                                      a18 == 4 ~ "20000-29000",
                                      a18 == 5 ~ "30000-39000",
                                      a18 == 6 ~ "40000-49000",
                                      a18 == 7 ~ "50000+")),
         income_1yr = as.factor(case_when(a18_rec1 == 0 ~ "$19,000",
                                      a18_rec1 == 1 ~ "$20,000-$49,000",
                                      a18_rec1 == 2 ~ "$50,000")),
         any_util = as.factor(case_when(any_util == 0 ~ "No",
                                        any_util == 1 ~ "Yes")),
         attempted_suicide = as.factor(case_when(g1c == 0 ~ "No",
                                                 g1c == 1 ~ "Yes")),
         employment = as.factor(
           case_when(a13 == 1 ~ "Full time",
```

```
                      a13 == 2 ~ "Part time",
                      a13 == 3 ~ "Student",
                      a13 == 4 ~ "Unemployed",
                      a13 == 5 ~ "Ctrl_envir")),
        homeless = as.factor(case_when(homeless == 0 ~ "No",
                                       homeless == 1 ~ "Yes")),
        hs_grad = as.factor(case_when(hs_grad == 0 ~ "No",
                                      hs_grad == 1 ~ "Yes")),
        group = as.factor(case_when(group == 0 ~ "Control",
                                    group == 1 ~ "Clinic")),
        # linkstatus = as.factor(case_when(linkstatus == 0 ~ "Did not link to primary care", linkstatu
        alcohol = as.factor(case_when(alcohol == 0 ~ "Not First Drug",
                                      alcohol == 1 ~ "First Drug Alcohol")),
        money_spent_on_alcohol = as.numeric(h16a),
        mh_index = as.numeric(mh),
        num_med_problems = as.numeric(d3),
        num_hospitilizations = as.numeric(d1),
        bothered_by_med = as.factor(case_when(d4 == 0 ~ "Not at all",
                                              d4 == 1 ~ "Slightly",
                                              d4 == 2 ~ "Moderately",
                                              d4 == 3 ~ "Considerably",
                                              d4 == 4 ~ "Extremely")),
        bothered = as.factor(case_when(d4_rec == 0 ~ "No",
                                       d4_rec == 1 ~ "Yes"))) %>%
  select(group, dayslink, linkstatus,
        yrs_education, gender, age,
        alcohol, alcq_30, marriage,
        employment, income, income_1yr,
        any_util, attempted_suicide, homeless,
        hs_grad, money_spent_on_alcohol,
        mh_index, num_med_problems,
        num_hospitilizations,bothered_by_med, bothered)
```

Note: based on our final results, we may end up removing some lines of unused variables (for the final report).

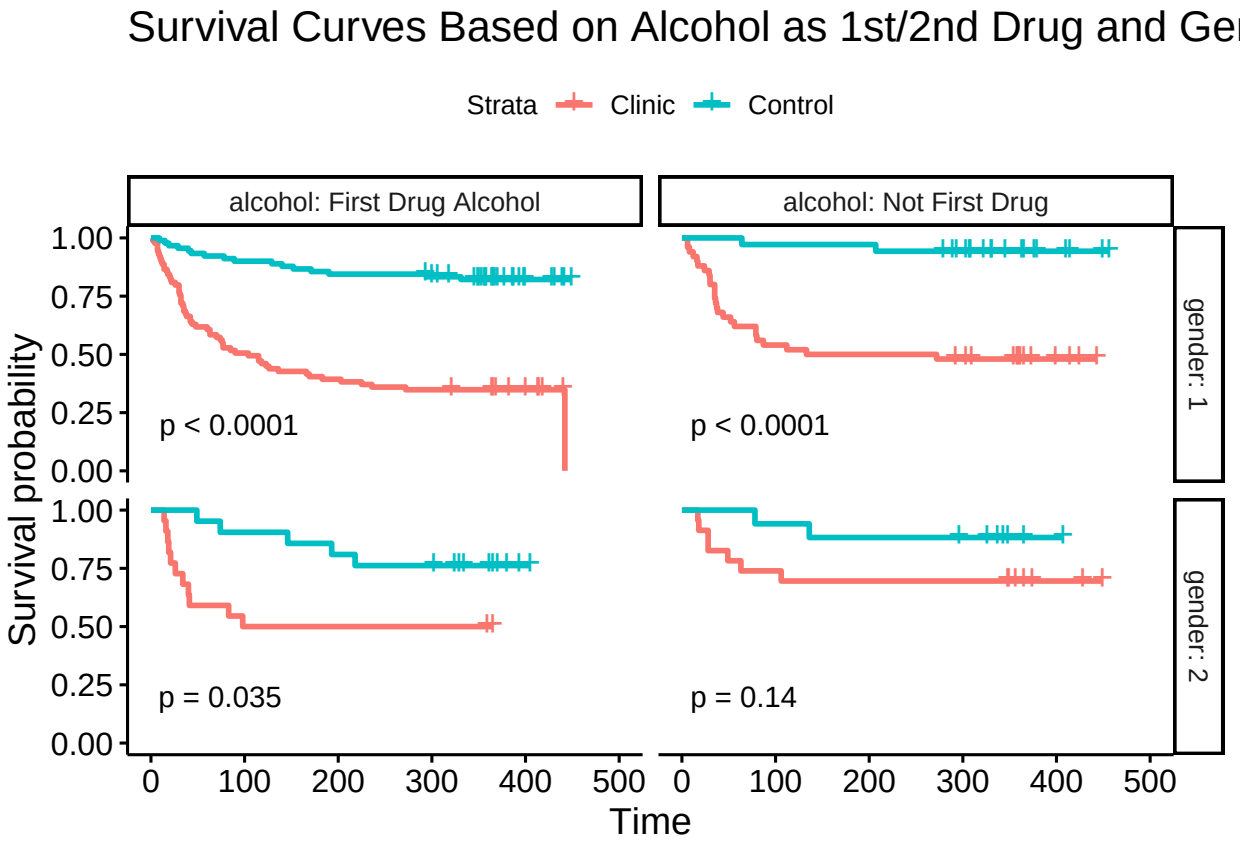
We begin by exploring some general variables in clinical research, such as age, gender, education level, and trial-specific variables, such as alcohol usage and medical conditions.

We want to create different data visualizations, in order to understand the relationship between variables and get more clues on the model building.

Since we are working with multiple categorical binary variables, we used the facet functionality to look at multiple survival probability plots simultaneously.

The plot below shows four survival probability plots separated by whether the individual's first drug was alcohol and gender. Gender is encoded as 1=Male and 2=Female. We mutated ALCOHOL to a string, alcohol "as first drug" or "not as first drug". The p-values represent significance for the log-rank test. A p-value less than 0.05 suggests evidence that the survival curves are not equal in favor of the alternative hypothesis, H_0 : the survival curves are equal.

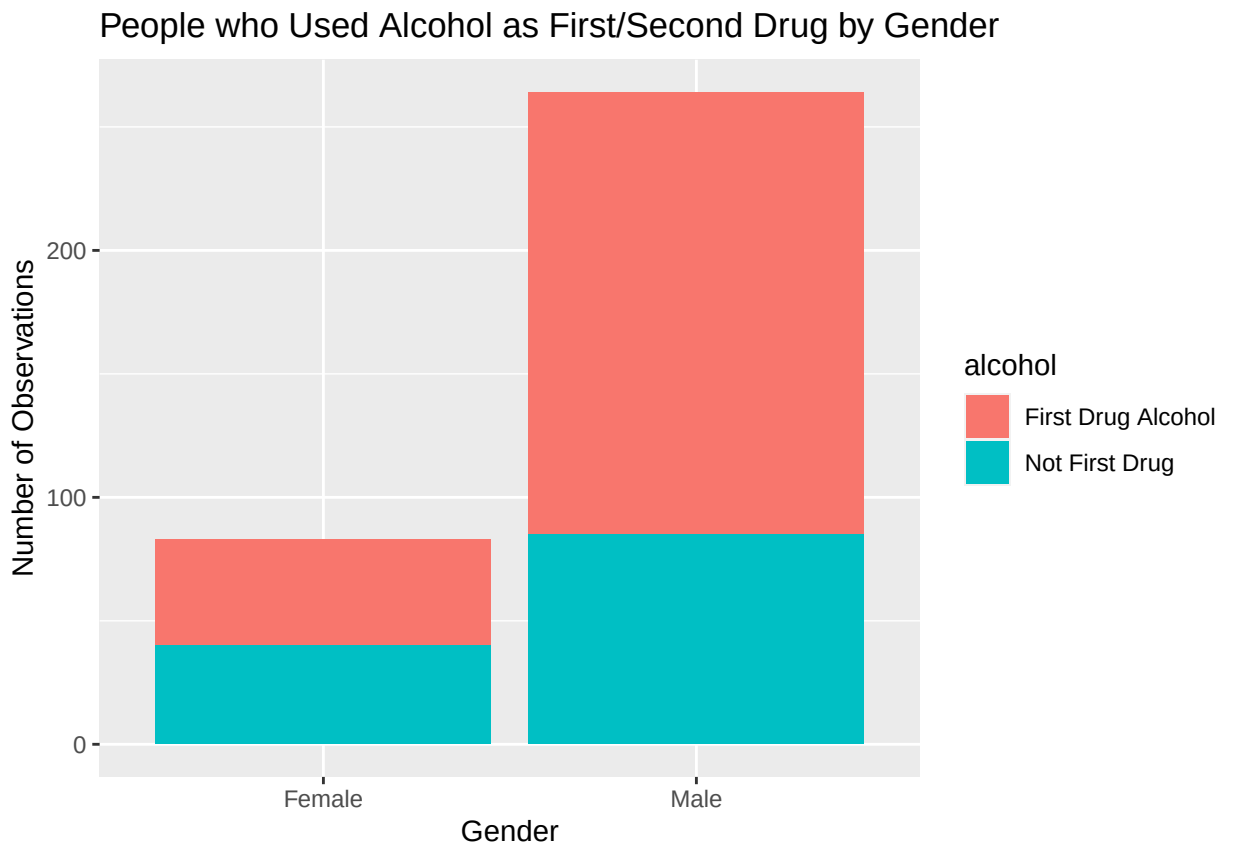
```
care_fit <- survfit(Surv(dayslink, linkstatus) ~ group, data=df)
ggsurvplot_facet(care_fit, df, facet.by = c("gender", "alcohol"), pval = TRUE) +
  ggtitle("Survival Curves Based on Alcohol as 1st/2nd Drug and Gender")
```



Looking at the plot, we see a p-value greater than 0.05 for observations whose first drug was not alcohol and gender is female. This means we fail to reject the null hypothesis that the survival curves are equal.

We want to see if there are some things to keep in mind, when it comes to the gender variable.

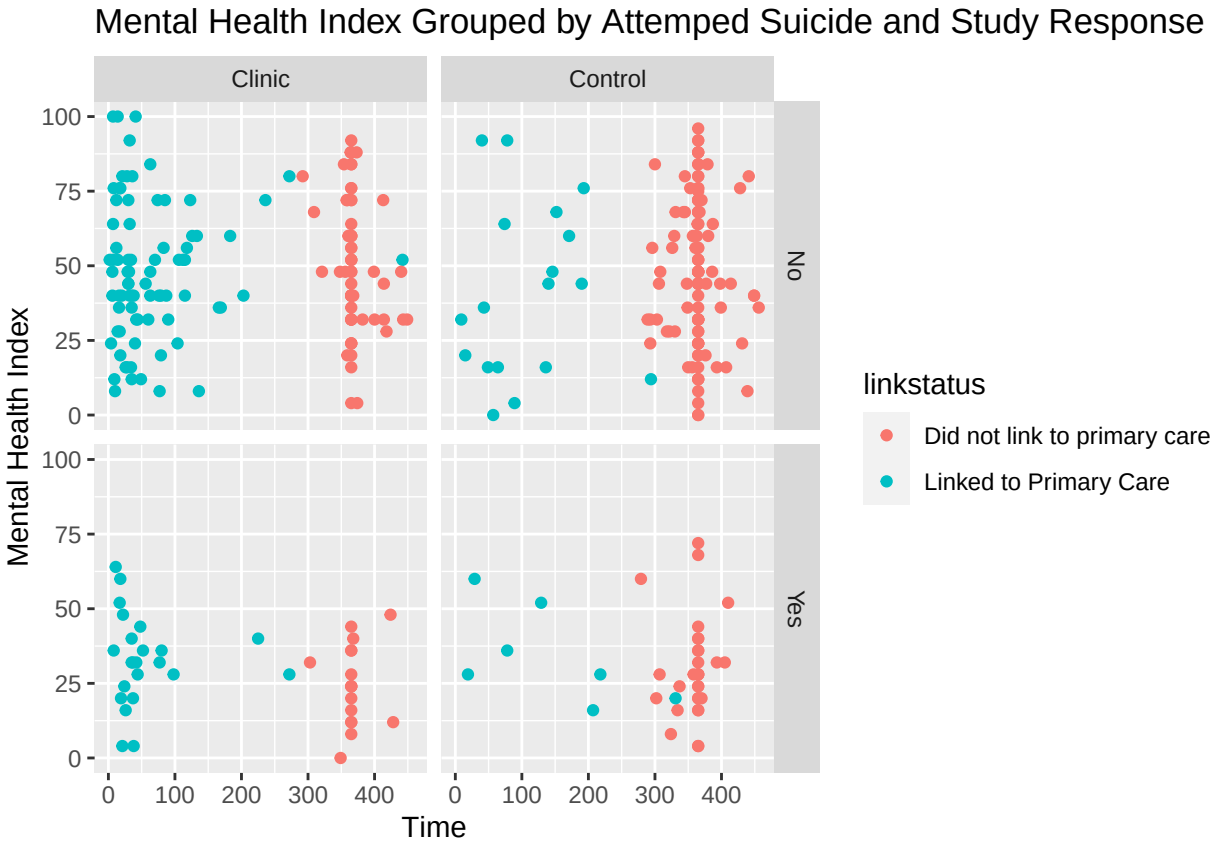
```
df %>%
  select(gender, alcohol) %>%
  mutate(gender_str = as.factor(case_when(gender == 1 ~ "Male",
                                          gender == 2 ~ "Female"))) %>%
  mutate(alcohol_str = as.factor(case_when(alcohol == 0 ~ "Not First Drug",
                                           alcohol == 1 ~ "First Drug Alcohol"))) %>%
  ggplot() + geom_bar(aes(x=gender_str, fill=alcohol)) +
  xlab("Gender") +
  ylab("Number of Observations") +
  ggtitle("People who Used Alcohol as First/Second Drug by Gender")
```



Note that the number of female participants are much less than the number of male participants. We will take this into consideration, when building the final model.

Going in another direction, here we observe the relationships mental health index and attempted suicide have on the linkstatus.

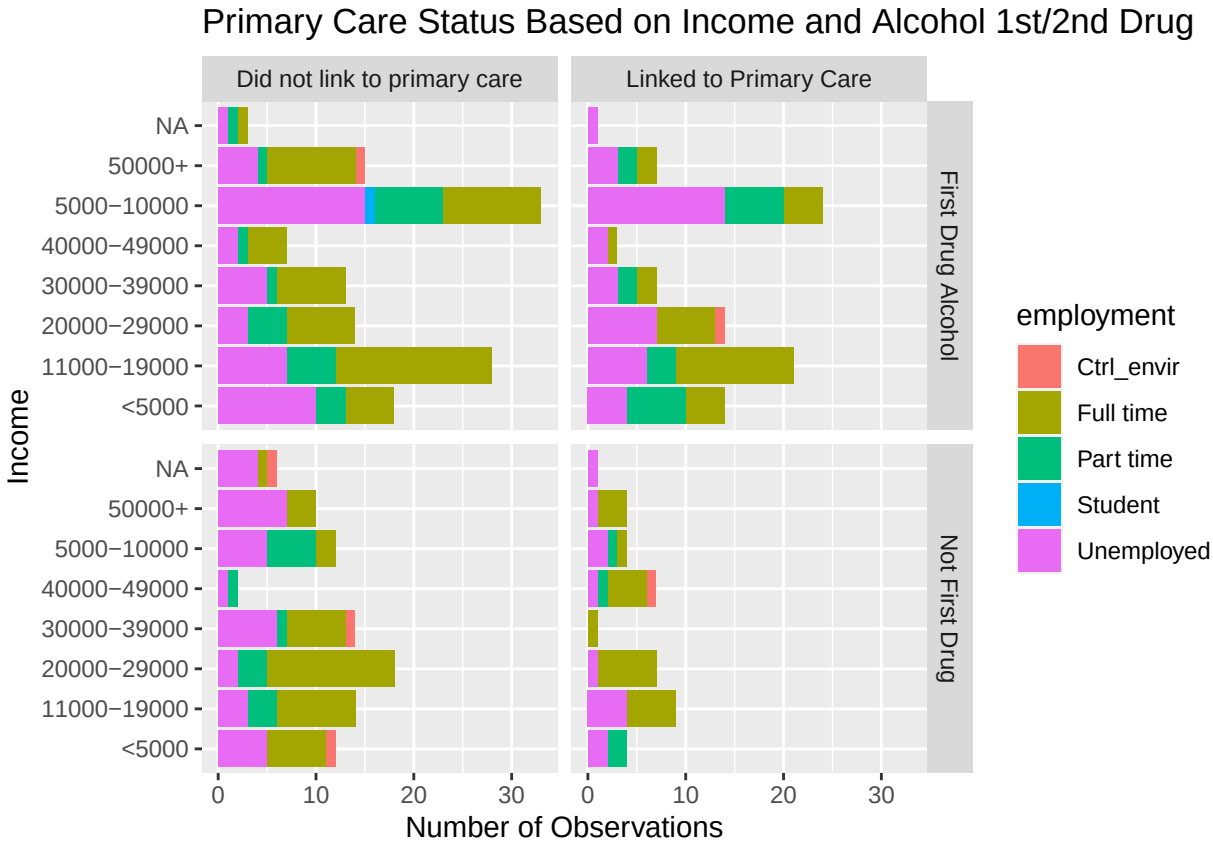
```
df %>%
  mutate(linkstatus = as.factor(case_when(linkstatus == 0 ~ "Did not link to primary care",
                                          linkstatus == 1 ~ "Linked to Primary Care"))) %>%
  select(group, linkstatus, dayslink, income, mh_index, attempted_suicide) %>%
  ggplot() +
  geom_point(aes(x=dayslink, y=mh_index, color=linkstatus)) +
  facet_grid(vars(attempted_suicide), vars(group)) +
  ylab("Mental Health Index") +
  xlab("Time") +
  ggtitle("Mental Health Index Grouped by Attemped Suicide and Study Response")
```



We see a distinct separation in response variable divided by the clinic and control group. In general, more individuals in the clinic group linked to primary care sooner than the control group. On the y-axis, a lower mental health index seems to be associated with more suicide attempts.

Next, we look at income and first drug alcohol and relation to the linkstatus.

```
df %>%
  select(income, employment, alcohol, group, linkstatus) %>%
  mutate(linkstatus = as.factor(case_when(linkstatus == 0 ~ "Did not link to primary care",
                                          linkstatus == 1 ~ "Linked to Primary Care")),) %>%
  ggplot() + geom_bar(aes(x=income, fill=employment)) +
  facet_grid(vars(alcohol), vars(linkstatus)) +
  coord_flip() +
  ggtitle("Primary Care Status Based on Income and Alcohol 1st/2nd Drug") +
  xlab("Income") +
  ylab("Number of Observations")
```



Observe that the number of people in the $\$40,000 - \$49,000$ income bracket is much smaller than the other brackets. This could pose statistical confusion in our model building progress, as the small number in that group could skew results.

Based on the exploratory data analysis, non-explanatory variables can influence the primary care linkstatus. For our model building, we will separate the medical and socioeconomic variables. We hypothesize that patients with more medical related problems would be inclined to connect to primary care. (anything else?)

Stepwise Regression

We proceed with stepwise regression. For each potential variable, we build a coxph model with the variable and one without. Then, we take the loglik deviation from the models and run a drop-in-deviance test. The drop-in-deviance test will help us determine if this additional variable, x_i, should be included in the model. The G statistic equals $2 * (logLik_{biggermodel} - logLik_{smallermodel})$. Calculating degrees of freedom is taking the difference in the number of parameters of the full model minus the restricted model. Finding the p-value is the "percentage of the X^2 distribution that exceeds G". We calculate the p-value by finding the converse of $phcisq(...)$, or $1 - pchisq(...)$. The null and alternative hypothesis for this test is $H_0 : \beta_i = 0$ and $H_a : \beta_i \neq 0$.

Since there were many variables to consider, we wrote a function that takes in the full and small model and returns a p-value from the drop-in-deviance test.

(very specific, very good, wonder if we need to add or trim any info)

```
# input: (small model's glance output, big model's glance output, degrees of freedom)
drop_in_dev <- function(smallmodel, bigmodel, df){
  small_loglik <- smallmodel$logLik
  big_loglik <- bigmodel$logLik
```

```
  G = 2*(big_loglik - small_loglik)
  return(1-pchisq(G, df))
}
```

Medical Explanatory Variables

We first consider the variable "attempted suicide".

```
full_model <- coxph(Surv(dayslink, linkstatus) ~ group + attempted_suicide, data=df) %>% glance()
restricted_model <- coxph(Surv(dayslink, linkstatus) ~ group, data=df) %>% glance()
drop_in_dev(restricted_model, full_model, 1)
```

```
## [1] 0.543385
```

The change in deviance is 0.3692, $(H_0 : \gamma = 0)$, so with one degree of freedom the p-value is 0.543385, which is greater than 0.05. We fail to reject the null hypothesis and do not need this variable in the model.

Next, we consider the mental health index.

```
full_model <- coxph(Surv(dayslink, linkstatus) ~ group + mh_index, data=df) %>% glance()
restricted_model <- coxph(Surv(dayslink, linkstatus) ~ group, data=df) %>% glance()
drop_in_dev(restricted_model, full_model, 1)
```

```
## [1] 0.6621623
```

The change in deviance is 0.1908, $(H_0 : \gamma = 0)$, so with one degree of freedom the p-value is 0.6622516, which is greater than 0.05. We fail to reject the null hypothesis and do not need this variable in the model.

Now, we consider gender.

```
full_model <- coxph(Surv(dayslink, linkstatus) ~ group + gender, data=df) %>% glance()
restricted_model <- coxph(Surv(dayslink, linkstatus) ~ group, data=df) %>% glance()
drop_in_dev(restricted_model, full_model, 1)
```

```
## [1] 0.07660948
```

The change in deviance is 3.1354, $(H_0 : \gamma = 0)$, so with one degree of freedom the p-value is 0.07660959, which is greater than 0.05. Note that the p-value is close to 0.05, which suggests that there could be little evidence. We fail to reject the null hypothesis and do not need this variable in the model.

We consider the variable "alcohol", or whether alcohol is the user's first or second drug.

```
full_model <- coxph(Surv(dayslink, linkstatus) ~ group + alcohol, data=df) %>% glance()
restricted_model <- coxph(Surv(dayslink, linkstatus) ~ group, data=df) %>% glance()
drop_in_dev(restricted_model, full_model, 1)
```

```
## [1] 0.004042625
```

The change in deviance is [1] 8.2646. ($H_0 : \gamma = 0$), so with one degree of freedom the p-value is 0.004042557, which is less than 0.05. We reject the null hypothesis that $\gamma = 0$ in favor of $H_a : \gamma \neq 0$ and should include first drink alcohol in the model.

Since gender has a p-value close to 0.05, we are interested in seeing if the additive model with alcohol will pass the drop-in-deviance test.

```
full_model <- coxph(Surv(dayslink, linkstatus) ~ group + gender + alcohol, data=df) %>% glance()
restricted_model <- coxph(Surv(dayslink, linkstatus) ~ group + alcohol, data=df) %>% glance()
drop_in_dev(restricted_model, full_model, 1)
```

```
## [1] 0.1809207
```

The p-value is 0.1809207. This does not suggest evidence to reject the null hypothesis that $\beta_i = 0$. (additional analysis sentence)

Socioeconomic Explanatory Variables

The p-values from the drop-in-deviance tests do not suggest adding all of the medical based variables, except for first drink alcohol. We will look to socioeconomic variables like age, income, employment, homeless, and high school graduation.

First, we consider age.

```
full_model <- coxph(Surv(dayslink, linkstatus) ~ group + alcohol + age, data=df) %>% glance()
restricted_model <- coxph(Surv(dayslink, linkstatus) ~ group + alcohol, data=df) %>% glance()
drop_in_dev(restricted_model, full_model, 1)
```

```
## [1] 0.04584322
```

$0.04584322 < 0.05$. We reject the null hypothesis and include age.

Now we consider income.

```
full_model <- coxph(Surv(dayslink, linkstatus) ~ group + alcohol + age + income, data=df) %>% glance()
restricted_model <- coxph(Surv(dayslink, linkstatus) ~ group + alcohol + age, data=df) %>% glance()
drop_in_dev(restricted_model, full_model, 1)
```

```
## [1] 1.733154e-08
```

A p-value of 1.733154e-08 means we reject the null hypothesis that $\beta_i = 0$. We include income in our model.

Now, the years of education.

```
full_model <- coxph(Surv(dayslink, linkstatus) ~ group + alcohol + age + yrs_education, data=df) %>% gla
restricted_model <- coxph(Surv(dayslink, linkstatus) ~ group + alcohol + age, data=df) %>% glance()
drop_in_dev(restricted_model, full_model, 1)
```

```
## [1] 1.126592e-05
```

A p-value of 1.126592e-05 means we reject the null hypothesis that $\beta_i = 0$. We include the "years of education" variable in our model.

(biggest question: how much of the above should be in Results instead, and which ones should be shortened?)
FINISH LATER

Results

The Cox PH analysis should include: an interpretation of your final survival model including a discussion of the sign of the coefficients (note: feel free to use interactions)

Note: Is the content appropriate for a results section? Simple introduction to the scientific question. Clear description of the results for each experiment. Analysis of those results.

Note: Are the results/data analyzed well? Given the data in each figure, is the interpretation accurate and logical? Is the analysis of the data thorough or are some aspects of the data ignored? Does the author make connections between ideas (graphs, models, etc.) within the text? Are the data interpreted in a larger context?

Note: Figures. Are the figures appropriate for the data being discussed?. Are the figure legends and titles clear and concise?

The tentative final model is $exp\{\beta_1 group + \beta_2 age + \beta_3 alcohol + \beta_4 income + \beta_5 YearsEducation\}$.

```
coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + income + yrs_education, data=df)
```

```
## Call:
## coxph(formula = Surv(dayslink, linkstatus) ~ group + age + alcohol +
##     income + yrs_education, data = df)
##
##                            coef exp(coef) se(coef)      z        p
## groupControl           -1.84690   0.15772  0.23327 -7.917 2.43e-15
## age                     0.02591   1.02625  0.01209  2.143  0.03208
## alcoholNot First Drug -0.41864   0.65794  0.21025 -1.991  0.04647
## income11000-19000       0.32875   1.38923  0.30125  1.091  0.27515
## income20000-29000       0.24058   1.27199  0.32938  0.730  0.46514
## income30000-39000       0.10428   1.10991  0.43231  0.241  0.80939
## income40000-49000       0.70615   2.02617  0.40202  1.756  0.07900
## income5000-10000        0.15755   1.17064  0.30889  0.510  0.61003
## income50000+           -0.18031   0.83501  0.38408 -0.469  0.63875
## yrs_education          -0.13184   0.87648  0.04967 -2.654  0.00794
##
## Likelihood ratio test=95.37  on 10 df, p=4.586e-16
## n= 333, number of events= 125
##    (14 observations deleted due to missingness)
```

Above are the coefficients for the model. Note that the the p-values of income levels are all greater than 0.05. In fact, the p-value for income40000-49000 is smaller than the other income coefficients by a factor of 100. One reason behind the smaller p-value is that the number of individuals with income between 40000-49000 are less than the other income groups.

Looking the plot, "Primary Care Status Based on Income and Alcohol 1st/2nd Drug", we see a drop in the number of individuals for the 40000-49000 income group. So even though the drop-in-deviance test is significant, the outliers could be dragging the p-value down. Thus, we will not be proceeding with income.

Our final model is as follows:

```
df %>% drop_na()
```

```
## # A tibble: 328 x 22
##    group dayslink linkstatus yrs_education gender   age alcohol alcq_30 marriage
```

```
##    <fct>    <dbl>      <dbl>         <dbl>  <dbl> <dbl> <fct>     <dbl> <fct>
##  1 Cont~      377          0            12      1    36 First ~     338 6
##  2 Clin~       49          1            12      2    27 Not Fi~     180 6
##  3 Clin~      365          0            12      1    39 First ~    1020 6
##  4 Clin~      365          0            12      1    24 Not Fi~       0 6
##  5 Clin~       42          1            11      1    29 First ~     340 6
##  6 Cont~      370          0            11      1    37 First ~      32 6
##  7 Cont~      387          0            10      1    35 First ~     750 6
##  8 Cont~      329          0            12      2    34 First ~     120 6
##  9 Clin~       21          1             7      2    33 First ~      52 1
## 10 Clin~       28          1            11      2    35 Not Fi~      12 6
## # ... with 318 more rows, and 13 more variables: employment <fct>,
## #   income <fct>, income_1yr <fct>, any_util <fct>, attempted_suicide <fct>,
## #   homeless <fct>, hs_grad <fct>, money_spent_on_alcohol <dbl>,
## #   mh_index <dbl>, num_med_problems <dbl>, num_hospitilizations <dbl>,
## #   bothered_by_med <fct>, bothered <fct>
```

```
model1 <- coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education, data=df)
model1 %>% tidy(conf.int = TRUE)
```

```
## # A tibble: 4 x 7
##   term                   estimate std.error statistic  p.value conf.low conf.high
##   <chr>                     <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 groupControl           -1.77      0.227      -7.81 5.78e-15 -2.21      -1.33
## 2 age                     0.0261    0.0118      2.21 2.68e- 2  0.00300    0.0492
## 3 alcoholNot First Drug  -0.429     0.201      -2.13 3.30e- 2 -0.824     -0.0347
## 4 yrs_education          -0.118     0.0472     -2.50 1.24e- 2 -0.211     -0.0256
```

```
# cox.zph(model1)
```

More formally,

$$h_i(t) = h_0 exp\{-1.76952 \cdot \text{Group} + 0.02609 \cdot \text{Age} - 0.42942 \cdot \text{Alcohol} - 0.11815 \cdot \text{YearsOfEducation}\}.$$

The most drastic change in risk for $\hat{HR}$ comes from a unit increase in group. That is, a change from the treatment group to the control group, decreases risk of linking to primary care by a factor of $exp(-1.76952) = 0.1704148$ with $CI(0.1092994, 0.2657042)$. $\hat{HR}$ for a unit increase in age increases the risk of linking to primary care by a factor of $exp(0.02608945) = 1.026433$, $CI(1.003001, 1.050412)$. The change in risk for a transition from alcohol being a first drug to alcohol not as a first drug decreases the risk of linking to primary care by $exp(-0.42942) = 0.6508865$, $CI(0.4385955, 0.965922)$. Last, a unit increase in years in education decreases the risk of linking to a primary care by $exp(-0.11815) = 0.8885628$, $CI(0.8100146, 0.9747364)$.

The final model suggests the patients in the treatment group showed the greatest decrease in risk. Alcohol first drug, showed the second greatest reduction in risk. Since all the p-values are significant, this model provides evidence to suggests that the patients who exhibit unit increases in these variables are subject to a decrease in risk of not linking to primary care. This partially supports our initial hypothesis that medical related problems can prompt individuals to connect to primary care. Age and alcohol first drug are significant coefficients that impact the risk of linking to primary care. The only socioeconomic variable that is significant is years of education, which decrease risk but not as greatly as group or alcohol.

We do not want to imply causation. While the study was conducted in the greater Boston area over 12 months, the patients were recruited from detox units. The variables we are using are widespread so everyone can be categorized. Having fewer, general variables means we can get more data points from people. However, one draw back to the model is that it is only additive and the patients do not represent the general population.

So, there could be interaction among the variables that we are not accounting more. In other words, we chose a simplified model given 777 explanatory variables. We conclude that the model can be generalized to rehab patients.

New Ideas

We choose to learn two new ideas, related to survival analysis. Michael will investigate assumptions about proportional hazard. Oliver will investigate the bootstrapping methods for the survival model.

Cox.zph

Cox PH models have an underlying assumption that hazards are proportional. The ratio between two results are constant over time, or presents a linear relationship. But hazard ratios are sometimes non-proportional, such as when KM curves cross, or one tapers off and another drops to zero. In fact, if we do not assume a baseline hazard rate and time invariant (meaning constant) coefficients and variables, the Cox PH model will not be accurate. To test the proportionality assumption of the Cox PH model, we use the cox.zph function.

The cox.zph function tests the proportionality of every (prediction) variable in the model. It does so by creating interactions with time in various time transformation. (https://stats.idre.ucla.edu/r/examples/asa/r-applied-survival-analysis-ch-6/) If the p-value is less than 0.05 for a coefficient, then it means the coefficient does not contribute linearly to the PH model and violates the proportionality asumption. If the p-values are big, then we do not reject the null hypothesis. This indicates the model hazard is porportional.

We want to make sure that the hazards are proportional to make the Cox PH model work. If they are not, then we want to see if transforming the time variable can bring back proportionality. If such attempt is successful, we can transform the original time variable in the dataset and revise the model, without starting from scratch. (http://st47s.com/Math150/Notes/survival-analysis.html) The code is given below, note that the basic cox.zph model and the completed default cox.zph model produce the same results.

```
cox.zph(model1)
```

```
##                 chisq df      p
## group          6.9850  1 0.0082
## age            0.0818  1 0.7749
## alcohol        1.1684  1 0.2797
## yrs_education 0.5564  1 0.4557
## GLOBAL         8.0528  4 0.0897
```

```
# without tranform, it assumes transform = "km"
# includes every condition in the vignette, the results are still the same by default settings
# cox.zph(fit, transform="km", terms=TRUE, singledf=FALSE, global=TRUE)
cox.zph(coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education, data=df), terms = TRU
```

```
##                 chisq df      p
## group          6.9850  1 0.0082
## age            0.0818  1 0.7749
## alcohol        1.1684  1 0.2797
## yrs_education 0.5564  1 0.4557
## GLOBAL         8.0528  4 0.0897
```

```r
# model interacts with log(time)
cox.zph((model1), transform="log")
```

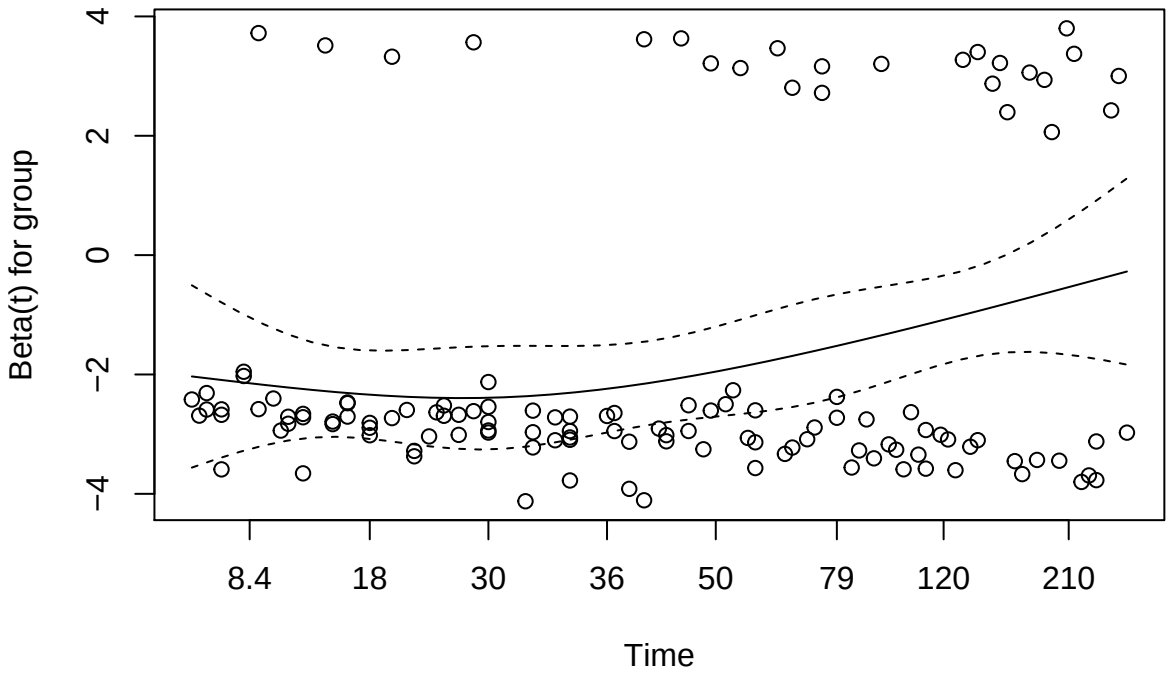
```
##                  chisq df     p
## group         6.5151  1 0.011
## age           0.0822  1 0.774
## alcohol       1.0440  1 0.307
## yrs_education 1.4434  1 0.230
## GLOBAL        8.1760  4 0.085
```

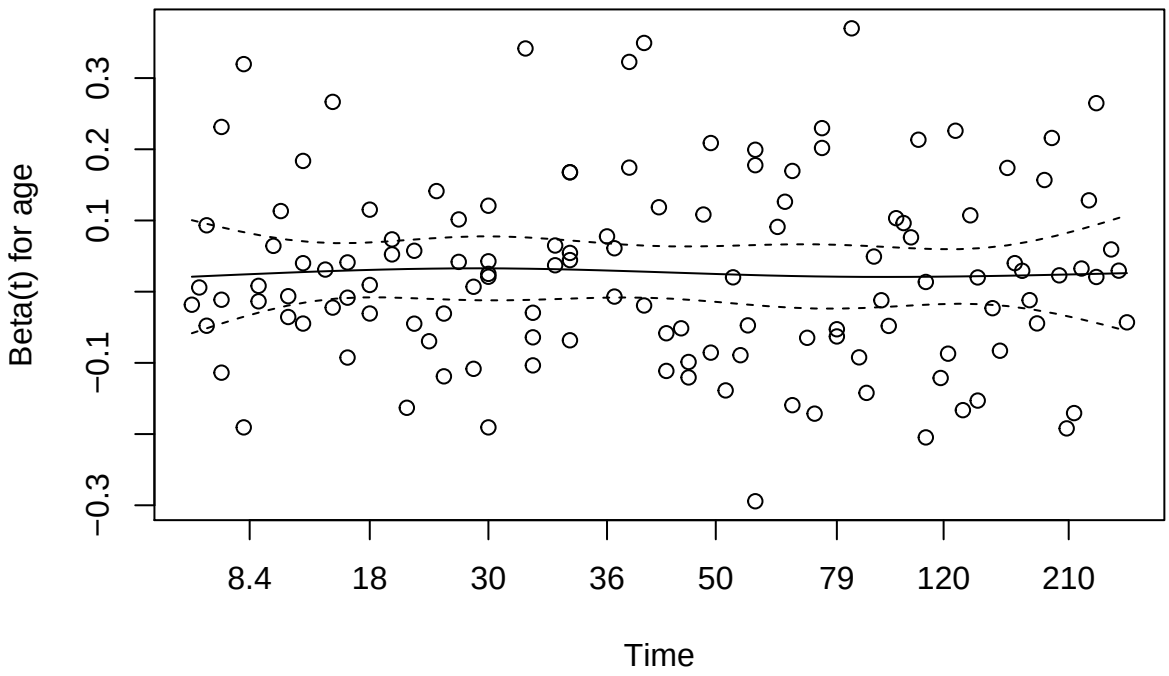
Here we see that for both methods of transformation, only group does not contribute linearly to the proportionality assumption. However, this may not be unacceptable, since group is a binary explanatory variable. We can leave it as as is. The other predictors pass the proportionality test.

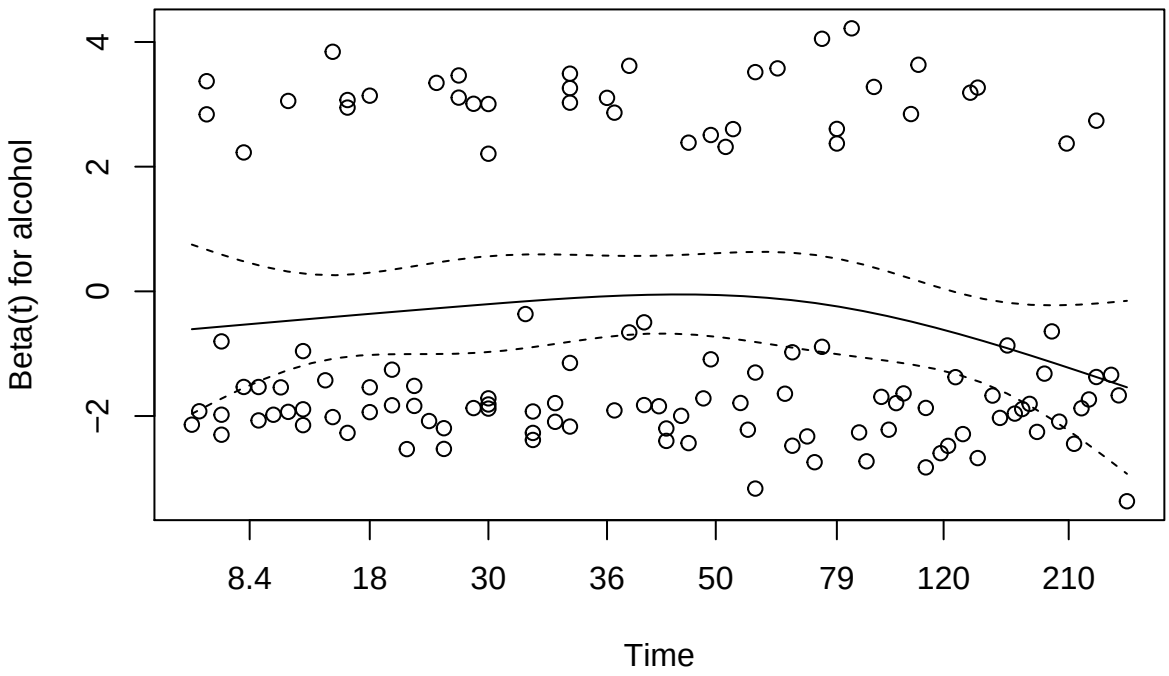
We also want to look at the residual plots of the model, since non-proportional hazards indicate variables that do not contribute linearly to the model. These residuals are called the Scoenfeld residuals, each predictor variable has its own plot. The plot function and the ggcoxzph function both plot the same residuals, the only difference is the former plots them separately, the latter plots them together similar to a facet function. The plots will also produce three regression lines.

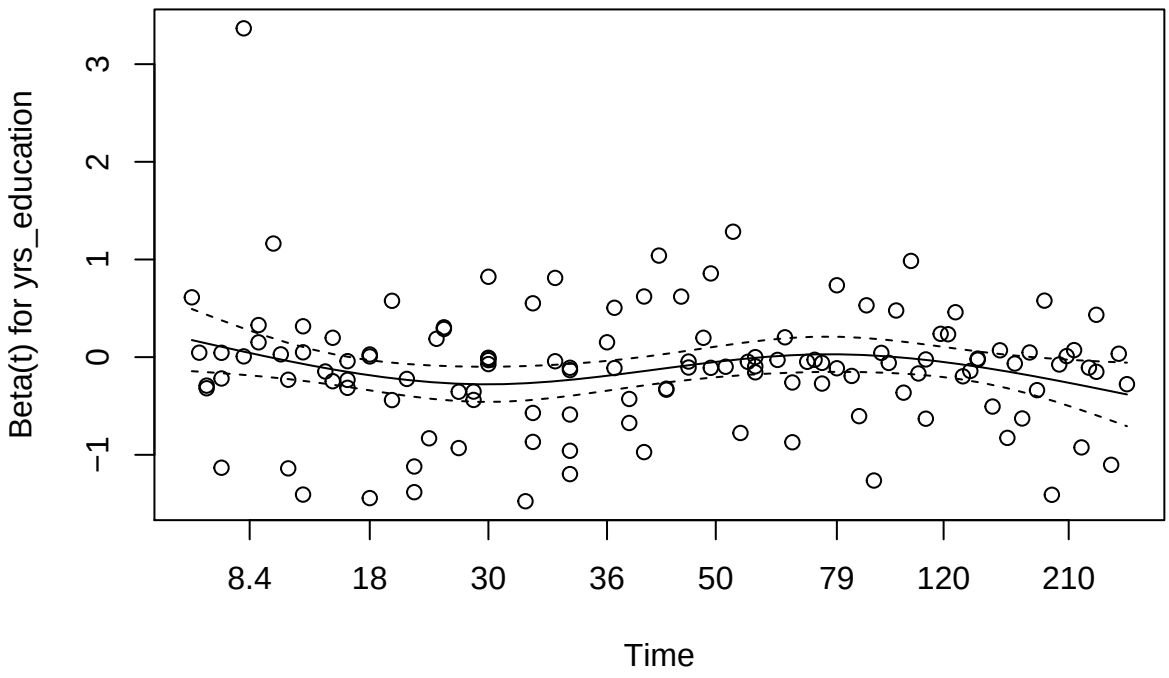
Schoenfeld residuals essentially revert the regression process. Let's call the explanatory variable X and the response variable Y. Previously we want to predict Y given X and the residual is the difference between the predicted Y and the observed Y. With Shoenfeld residuals, we want to predict X given Y and find the difference between the predicted X and the observed X. The following article is a comprehensive source on Schoenfeld residuals with Python. (https://towardsdatascience.com/schoenfeld-residuals-the-idea-that-turned-regression-modeling-on-its-head-b1f1fd293f87)

```r
plot(cox.zph(model1))
```

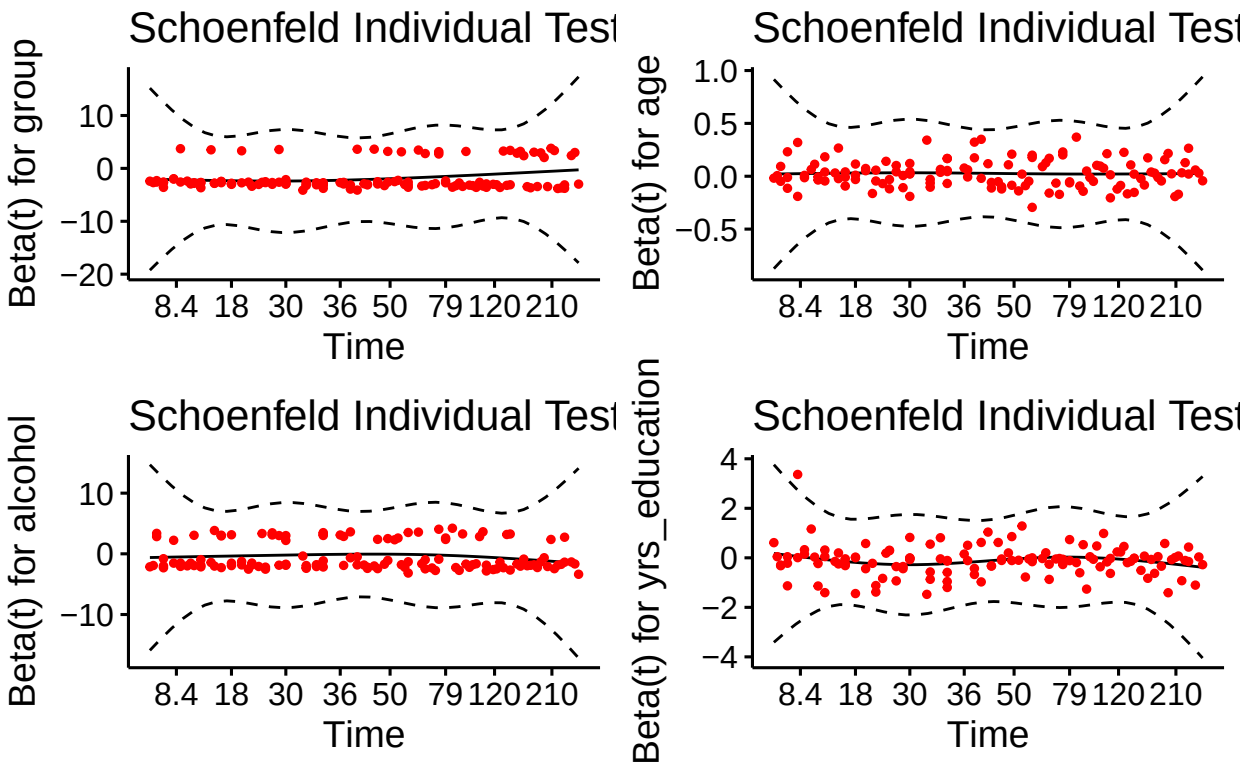

Beta(t) for group
4
2
0
−2
−4
8.4
18
30
36
50
79
120
210
Time


Beta(t) for age
0.3
0.2
0.1
−0.1
−0.3
8.4
18
30
36
50
79
120
210
Time


Beta(t) for alcohol
4
2
0
−2
8.4
18
30
36
50
79
120
210
Time



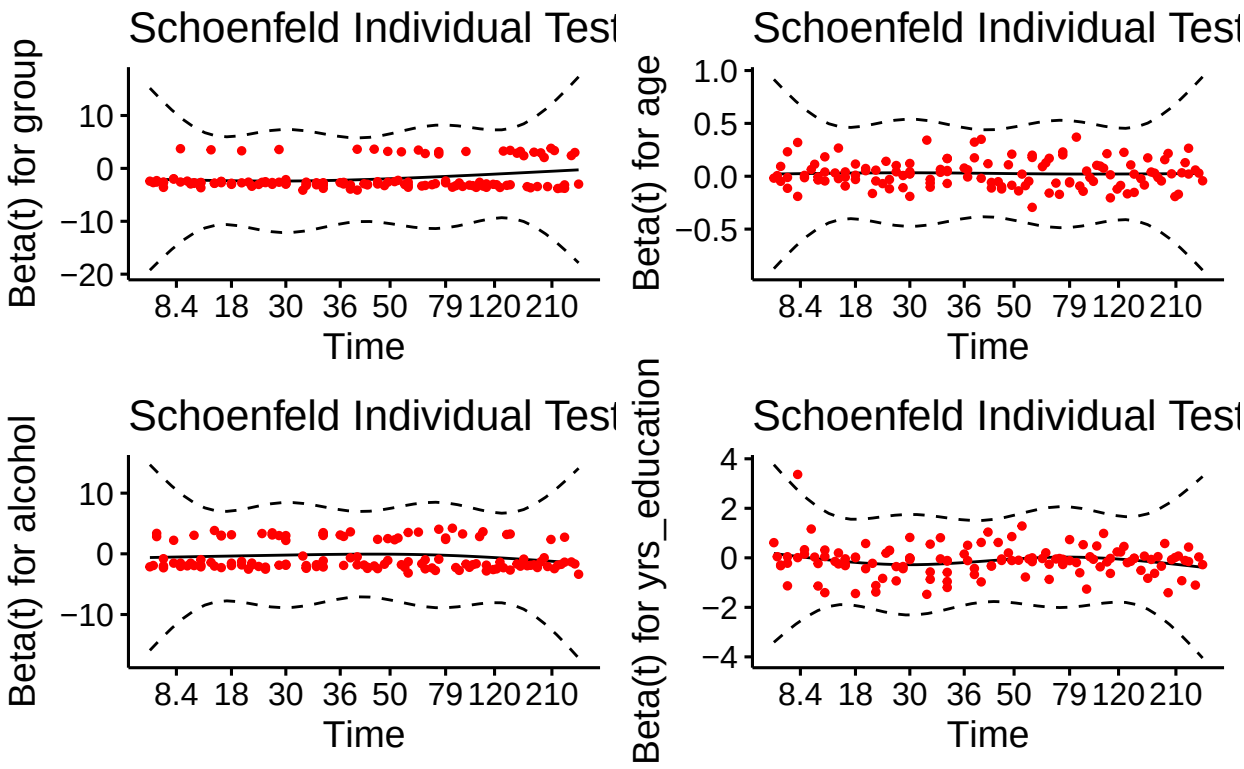

```
# the plot shows the Schoenfeld residuals in 4 plots
ggcoxzph(cox.zph(coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education, data=df)))
```




```
# the plot shows the Schoenfeld residuals in 1 big plot
```

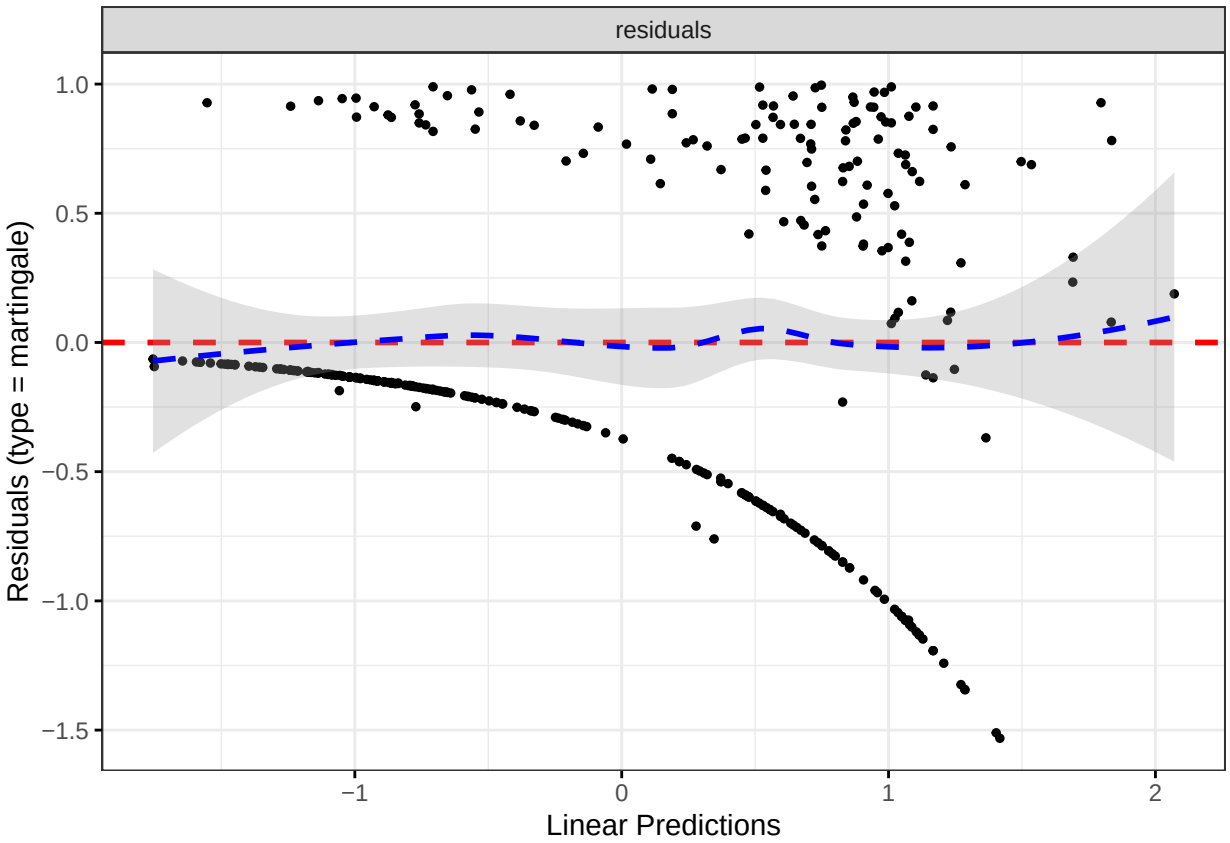
The mean of Schoenfeld residuals is zero, if the regression coefficients of the Cox PH model are not dependent on time. In fact, if the Cox PH model is indeed proportional, then the Schoenfeld residuals should be randomly distributed given a large dataset. We observe a more randomly distributed Schoenfeld residual plots, except the group variable. Therefore we are okay with using this Cox PH model, since group is our explanatory variable and the other predictor variables pass the proportionality test. (https://towardsdatascience.com/schoenfeld-residuals-the-idea-that-turned-regression-modeling-on-its-head-b1f1fd293f87)

```
ggcoxzph(cox.zph(coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education, data=df)))
```




```
ggcoxdiagnostics(coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education, data=df))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Bootstrapping the Survival Model

Bootstrapping the survival model (what are the assumptions? what do you conclude?)

Bootstrapping the Survival Model

Bootstrapping is a nonparametric technique that resamples the sampled data and samples from the resampled data. It is most useful for building confidence intervals without the assumption of the central limit theory. However, one important assumption is that the observations were randomly sampled. This holds true for this study as the paper mentions 235 observations were randomized into the help clinic. From bootstrapping we hope to estimate the true parameter of the population, which are the coefficients to our proportional hazard model.

We are interested in bootstrapping the survival model because the dataset is right censored. The variable, Linkstatus, indicates whether an observation connected to primary care within 12 months. We right censor when an individual did not connect to primary care by 12 months. Given that we have censored data offers multiple resampling options. In bootstrapping, we care about the resampling method because it will directly affect our statistical inference. We will be building a survival model out of resampled data. The library "censboot" will help facilitate the various resampling algorithms. Censboot references "Bootstrap Methods and their Applications" by Davison and Hinkley (1997) for their resampling options. There are four different simulation options. We will explore the ordinary and conditional sampling methods. The censboot function requires some statistic to be returned. Since we are comparing multiple sampling methods, we will estimate the survival model coefficients and the loglik deviations from the model. We can calculate the t-staistic from the loglik deviations and then find the p-value for the Likelihood Ratio test. The null hypothesis is $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. The alternative hypothesis is $H_A :$ at least one coefficient is not zero.

Ordinary Resampling

As the name suggests, Ordinary resampling is sampling under a random distribution. Ordinary resapmling is appropriate when the dataset is "subject to random censorship" (A.C. Davision 1997) According to the paper we are analysis, the patients were enrolled in a randomized control trial. So if we were to use ordinary resampling, the sampling distribution would be equivalent to the original data set. Below, we bootstrap the coxph coefficients and loglik deviance using the ordinary simulation.

```
cox.fun <- function(data) {
     data <- data %>% drop_na()
     model <- coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education, data = data)
     out <- c(model$coefficients, model$loglik)
     out
}
model1_data <- df %>% select(group, age, alcohol, yrs_education, dayslink, linkstatus) %>% drop_na()
censboot(data = model1_data, cox.fun , R = 499, sim = "ordinary")
```

```
##
## CASE RESAMPLING BOOTSTRAP FOR CENSORED DATA
##
##
## Call:
## censboot(data = model1_data, statistic = cox.fun, R = 499, sim = "ordinary")
##
##
## Bootstrap Statistics :
##          original        bias    std. error
## t1*   -1.76951814 -0.0289273248  0.22404354
## t2*    0.02608945  0.0004876738  0.01157350
## t3*   -0.42942499  0.0095624360  0.20942285
## t4*   -0.11814557 -0.0006739915  0.05391067
## t5* -711.25670498  1.0957085658 47.68553752
## t6* -667.45187238  3.1763259904 47.21814062
```

The coefficients from the bootstrapped model are very similar to the original dataset. In fact, they are only a factor of 1000 off from the original model. And the loglike deviance give a p-value of 0, which means we reject the null hypothesis that the coefficients are not zero. We pass the Likelihood Ratio test. It is important to point out that the low bias usually means there is higher variance. A high variance could indicate that the model is capturing noise instead of signal.

Conditional Bootstrap

Conditional resampling works under the assumption that the censoring variable is unrelated to the response variable. Since we do not have some distribution to sample from, "simulations should be conditional" on the censorship pattern (A.C. Davision 1997) The pattern of the censorship can be estimated by providing the original censored data and the reversed censored data. To conditionally sample from the censorship pattern, censboot estimates failure times, which is the time an observation dies off. Each observation is censored or not censored depending on its failure time and the censor distribution.

Here is the estimated coefficients and loglik using the conditional coefficients.

```r
cond.fun <- function(data) {
  cox <- coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education, data = data)
  c(cox$coefficients, cox$loglik)
}
df <- df %>% select(dayslink, linkstatus, group, age, alcohol, yrs_education) %>% mutate(group = as.num
df <- df[order(df$group),]
df <- as.data.frame(df)
s1 <- survfit(Surv(dayslink, linkstatus) ~ group, data = df)
s2 <- survfit(Surv(dayslink-0.001*linkstatus, 1-linkstatus) ~ 1, data = df)
censboot(df, cond.fun, R = 499, strata = df$group,
     F.surv = s1, G.surv = s2, sim = "cond")
```

```
##
## STRATIFIED CONDITIONAL BOOTSTRAP FOR CENSORED DATA
##
##
## Call:
## censboot(data = df, statistic = cond.fun, R = 499, F.surv = s1,
##     G.surv = s2, strata = df$group, sim = "cond")
##
##
## Bootstrap Statistics :
##          original      bias    std. error
## t1*   -1.76951814  0.11683728  0.22706448
## t2*    0.02608945 -0.02598209  0.01346166
## t3*   -0.42942499  0.41685865  0.19300883
## t4*   -0.11814557  0.11846716  0.04510366
## t5* -711.25670498 -1.75969340 43.01336043
## t6* -667.45187238 -9.17835828 42.23973609
```

The conditional sampling method returned the same coefficients and loglik deviance as the ordinary resampled and original model. One noteworthy difference is that the bias is much bigger in the bootstrapped model using conditional simulation. A larger bias suggests a lower variance, which could miss data signal.

Discussion

We hypothesized that medical related variables like age, gender, and alcohol would prompt more patients to connect to primary care. Our model building process revealed that only age and alcohol first drug passed the drop-in-deviance test. Since we had to drop multiple variables we looked towards adding socioeconomic explanatory variables. The second stepwise regression concluded with years in education passing the drop-in-deviance test. So the final additive model included the treatment group, age, alcohol first drug, and years in education. An additive model with four variables is simple and general enough to capture most data points. The significant coefficients and significant Likelihood Ratio test, lead us to conclude that there is a strong association with treatment group, age, alcohol, and years in education influencing the risk of linking to primary care. Hence, the model supports our hypothesis that medical and socioeconomic explanatory variables affect an individual's outcome to linking to primary care. As such, this research can be generalized to rehab patients in the United States with the caveat that the myriad variables warrants further research.

======= For your analysis, you should give details of what is going on, how it is relevant, what are the technical conditions, what are the conclusions, etc.

Discussion

Does the author clearly state whether the results answer the question? (i.e. support or disprove the hypothesis?) Were specific data cited from the results to support each interpretation? Does the author clearly articulate the basis for supporting or rejecting the hypothesis ? Does the author adequately relate the results of the current work to previous research? Does the author appropriately discuss to whom the results can be generalized?

References (for new ideas)

Are the references appropriate and of an adequate quantity? Are the references cited properly (both within the text and at the end of the paper)?

Writing Quality (final check in the evening, read it out loud)

Is the paper well organized? (Paragraphs are organized in a logical manner) Is each paragraph well written? (Clear topic sentence, single major point) Is the paper generally well written? (Good use of language, sentence structure)

References

A.C. Davision, D.V Hickey. 1997. *Bootstrap Methods and Their Applications.* United Kingdom: Cambridge Series on Statistical; Probabilistic Mathematics.

Brian Ripley. 2020. *Censboot: Bootstrap for Censored Data.* https://www.rdocumentation.org/packages/boot/versions/1.3-27/topics/censboot.

Jeffrey H Samet, Nicholas J Horton, Mary Jo Larson. 2003. *Linking Alcohol- and Drug-Dependent Adults to Primary Medical Care: A Randomized Controlled Trial of a Multi-Disciplinary Health Intervention in a Detoxification Unit.* https://pubmed.ncbi.nlm.nih.gov/12653820/.

Johanna Hardin. 2020. *Methods in Biostatistics Class Notes.* Claremont, California: Pomona College. http://st47s.com/Math150/Notes/index.html.