Survival Analysis on Linking Drug Dependent Adults to Primary Care

Introduction

Note: might want to think more about the TITLE, after finishing the doc.

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

Survival analysis estimates an outcome on an observation or experimental study. The dataset, HELPFUL, 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? why do we want to research anything about the HELPfull data? any applications or insights? Why is the work important? What is its relevance?

(explain some of the main variables already)

BP3: 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?

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?

Methods

first paragraph method: what is our goal? can we replicate the results & process? (yes, because we will use the same dataset and the same code)

second paragraph (and more): how we approached the data exploration & cox.ph model building

Note: the order is visualize some data (code given, including mutations), figure out which ones we want to focus on, and proceed to build the Model

include dataset and packages used (lines of code)

model building process: ?

Keep filling in, after writing about the EDA and model building

Exploratory Data Analysis

For our exploratory data analysis, we will observe relationship the explanatory variables have with the response vairables, linkstatus. Analyzing the relationship between the explantory and response variable will help indicate which 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)

# import the dataset HELPFUL. Encode NAs as "*"
df <- read_csv("HELPdata.csv", na="*")
```

Now that the dataset is imported, we can select our variables of interest. Note that the original dataset has some variables encoded as characters types. We converted these variables to a factor type so categorical variables are easily distinguishable in our plots.

```
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).

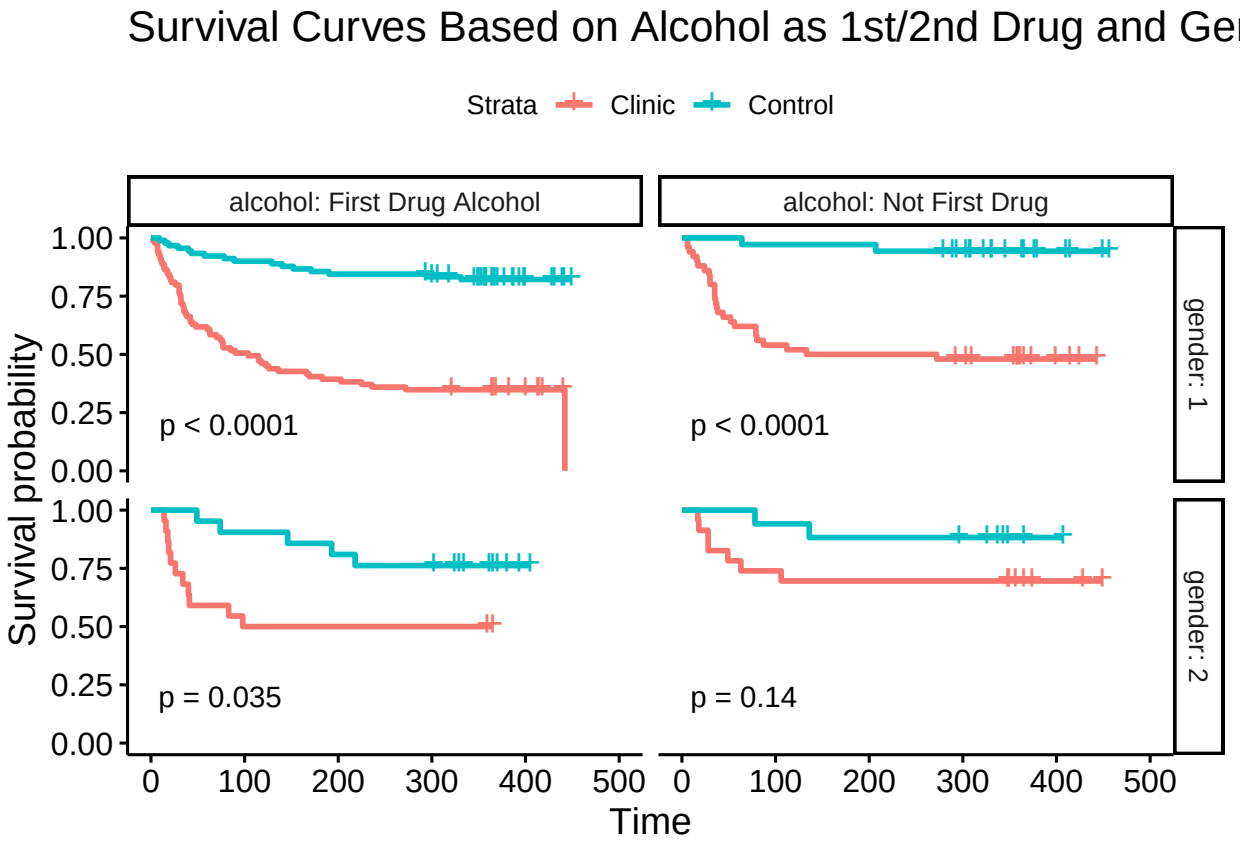
We begin by exploring some general variables in clinical research, such as age, gender, education level, and trial-specific variables, such as alcohol usage and medical conditions, etc.

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

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.

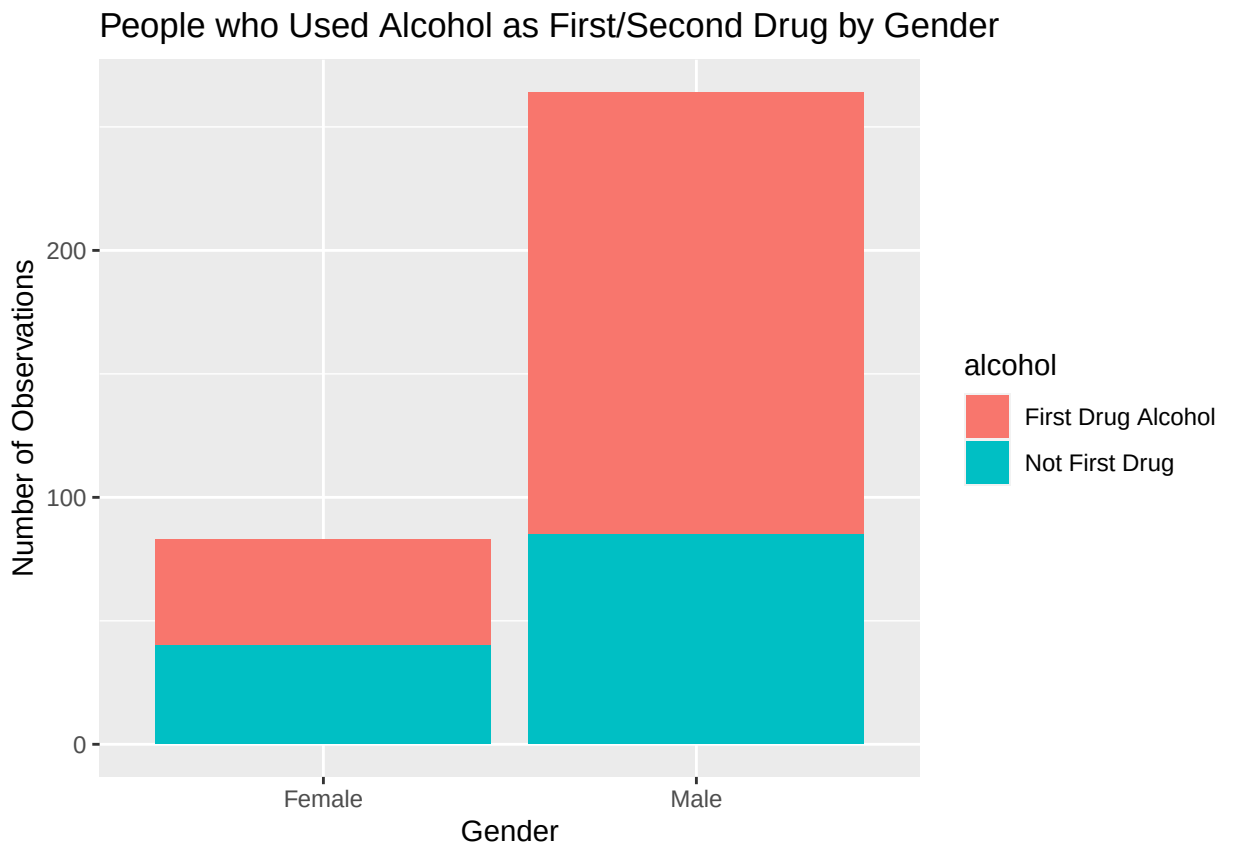
```
care_fit <- survfit(Surv(dayslink, linkstatus) ~ group, data=df)
ggsurvplot_facet(care_fit, df, facet.by = c("gender", "alcohol"), pval = TRUE) +
  ggtitle("Survival Curves Based on Alcohol as 1st/2nd Drug and Gender")
```



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.

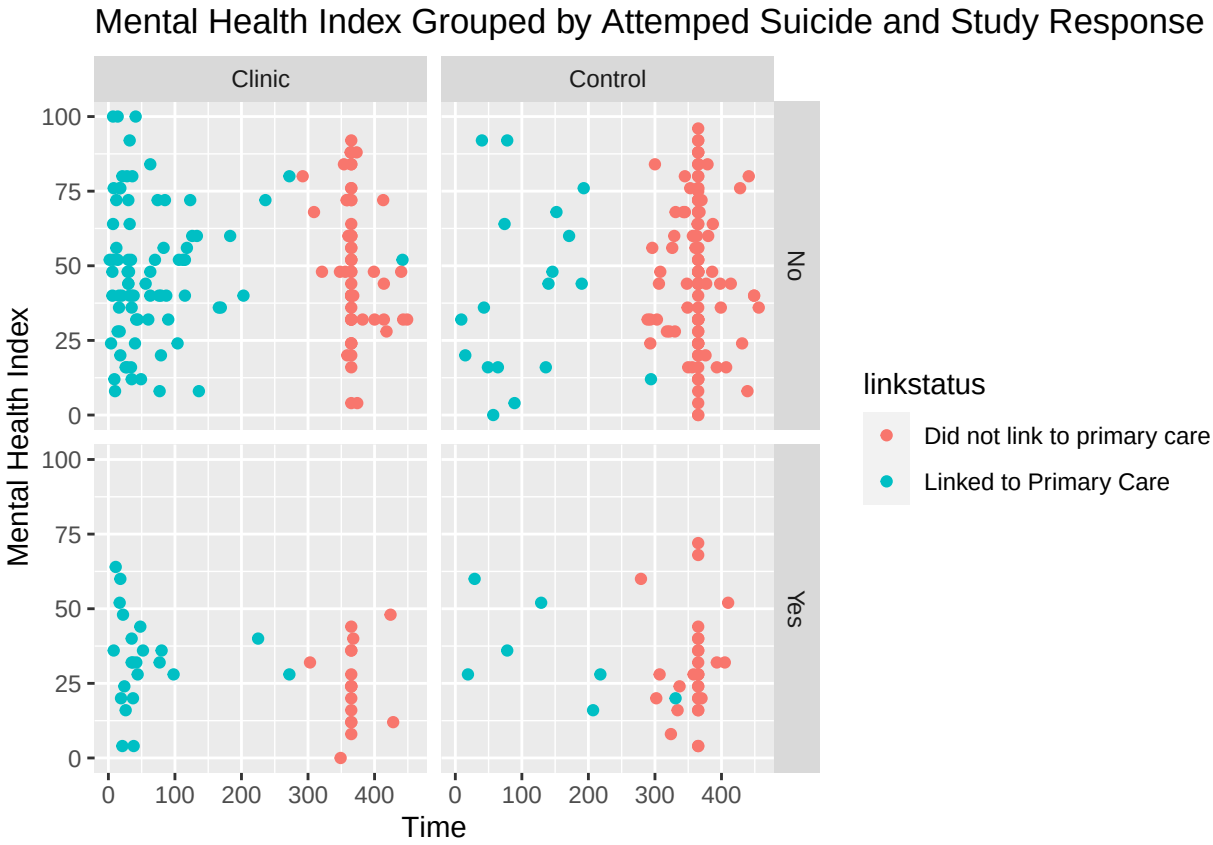
```
df %>%
  select(gender, alcohol) %>%
  mutate(gender_str = as.factor(case_when(gender == 1 ~ "Male",
                                          gender == 2 ~ "Female"))) %>%
  mutate(alcohol_str = as.factor(case_when(alcohol == 0 ~ "Not First Drug",
                                           alcohol == 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 way less than the number of male participants. (talk about how this factors into not using Gender in final model)

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

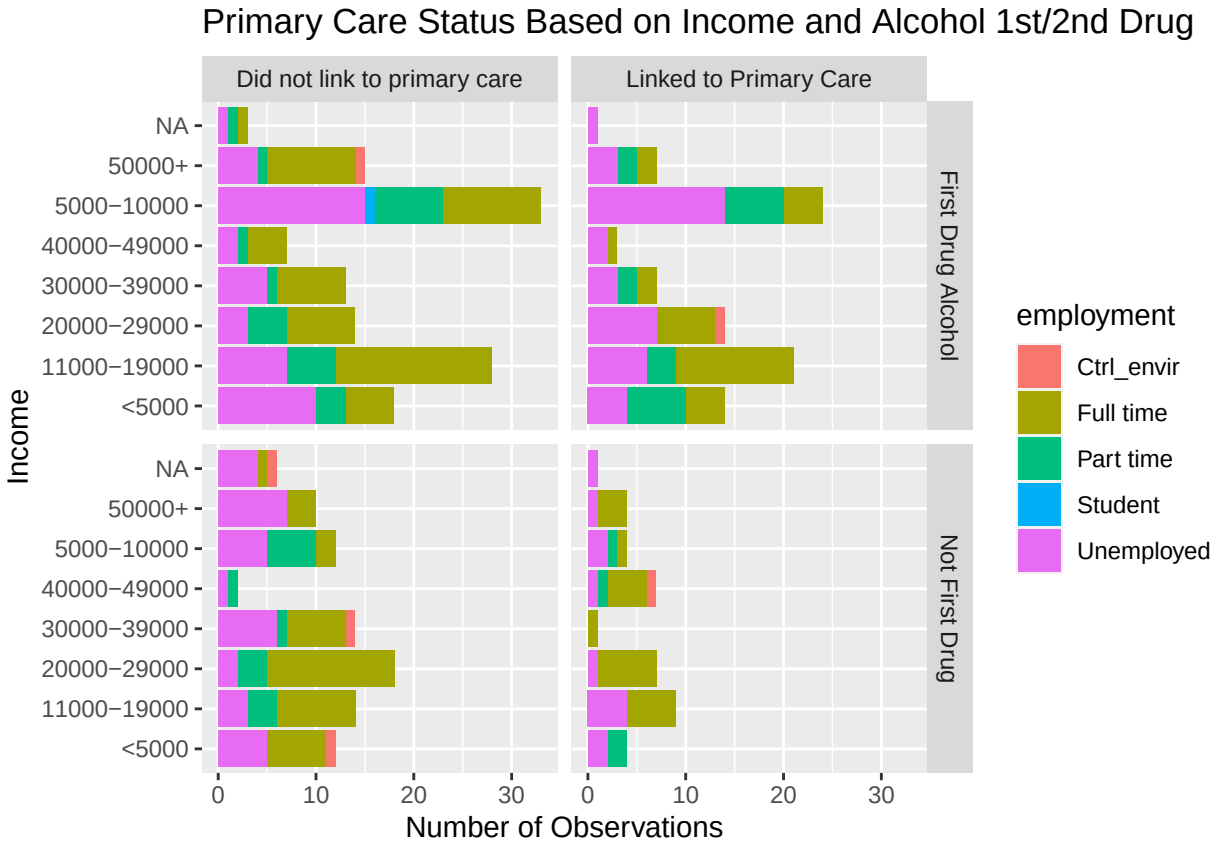
```
df %>%
  mutate(linkstatus = as.factor(case_when(linkstatus == 0 ~ "Did not link to primary care",
                                          linkstatus == 1 ~ "Linked to Primary Care"))) %>%
  select(group, linkstatus, dayslink, income, mh_index, attempted_suicide) %>%
  ggplot() +
  geom_point(aes(x=dayslink, 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 the results.

Based on the exploratory data analysis, medical and variables can influence the primary care link status. For our model building process, 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.

Stepwise Regression

We proceed with stepwise regression. For each potential variable, we build a coxph model including it and a model without the additional variable. 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 an 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 $phcisq(...)$, so $1 - pchisq(...)$. 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.

```
# 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 type in the model.

Next, 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 type in the model.

```
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 type in the model.

```
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.

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

```
## [1] 0.07209521
```

We see that the p-value for the additive model for treatment groups and the number of medical problems is insignificant (0.07209521). Note that the p-value is relatively close to 0.05, but do not include number of medical problems because we aim to produce the simplest model.

Since gender has a p-value close to zero, 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 $\beta_i = 0$.

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 graduate.

First, 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 should include age.

Next, we consider employment.

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

```
## [1] 0.06941881
```

A p-value of 0.06941881 means we fail to reject the null hypothesis that $\beta_i = 0$.

Consider the binary variable, homeless.

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

```
## [1] 0.7095458
```

A p-value of 0.7095458 means we fail to reject the null hypothesis that $\beta_i = 0$.

```
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.

```
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 income in our model.

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

```
## [1] 0.08662291
```

A p-value of 0.08662291 means we fail to reject the null hypothesis that $\beta_i = 0$. We do not include income in our model.

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 are income 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 amount of individuals with income between 40000-49000 were 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.

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) Which variable(s) are in? Which are out? What do you conclude about linking to primary care? Is there anything worth mentioning about how you got to your final model? What can you say about causation? What can you say about generalizing to a larger population?

Our final model:

```
model1 <- coxph(Surv(dayslink, linkstatus) ~ group + age + alcohol + yrs_education, data=df)
model1
```

```
## Call:
## coxph(formula = Surv(dayslink, linkstatus) ~ group + age + alcohol +
##     yrs_education, data = df)
##
##                            coef exp(coef) se(coef)      z        p
## groupControl           -1.76952   0.17042  0.22661 -7.809 5.78e-15
## age                     0.02609   1.02643  0.01178  2.214   0.0268
## alcoholNot First Drug -0.42942   0.65088  0.20141 -2.132   0.0330
## yrs_education          -0.11815   0.88857  0.04722 -2.502   0.0124
##
## Likelihood ratio test=87.61  on 4 df, p=< 2.2e-16
## n= 344, number of events= 127
##    (3 observations deleted due to missingness)
```

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 unit increase in the binary encoding, a change in the treatment group to the control group, has a decrease in risk by a factor of $exp\{-1.76952\} = 0.1704148$. The change in risk for a transition from alcohol being a first drug to non-alcoholic first drug decreases by $exp\{-0.42942\} = 0.6508865$. Last, a unit increase in years in education decreases by $exp\{-0.11815\} = 0.8885628$.

New Ideas

Cox.zph

Bootstrapping the Survival Model