

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 HELPFUL 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 variables, linkstatus. Analyzing the relationship between the explanatory 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 character 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", linkstatus == 1 ~ "Linked to primary care")),
         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_hospitalizations = 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)

```

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.

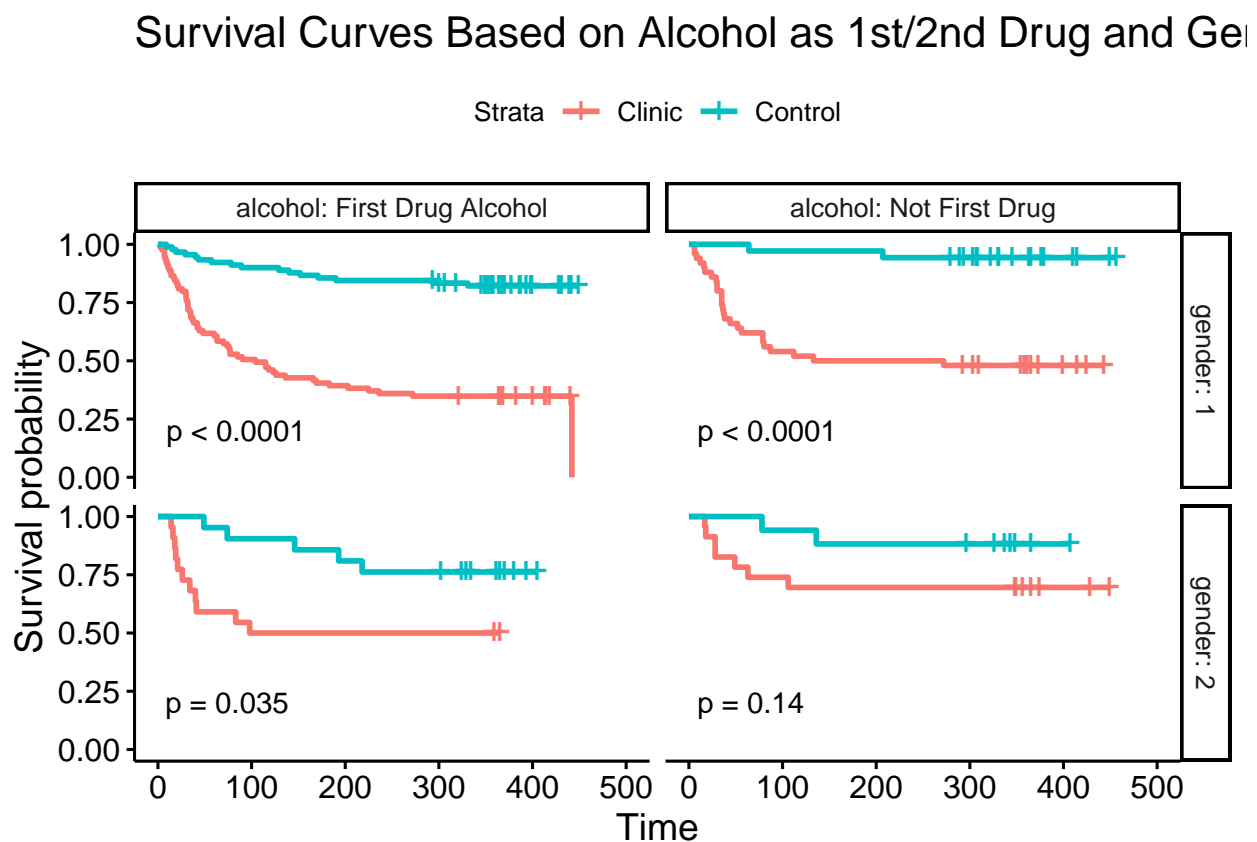
Since we are working with multiple categorical binary variables, we used the facet functionality to look at multiple survival probability plots simultaneously.

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")

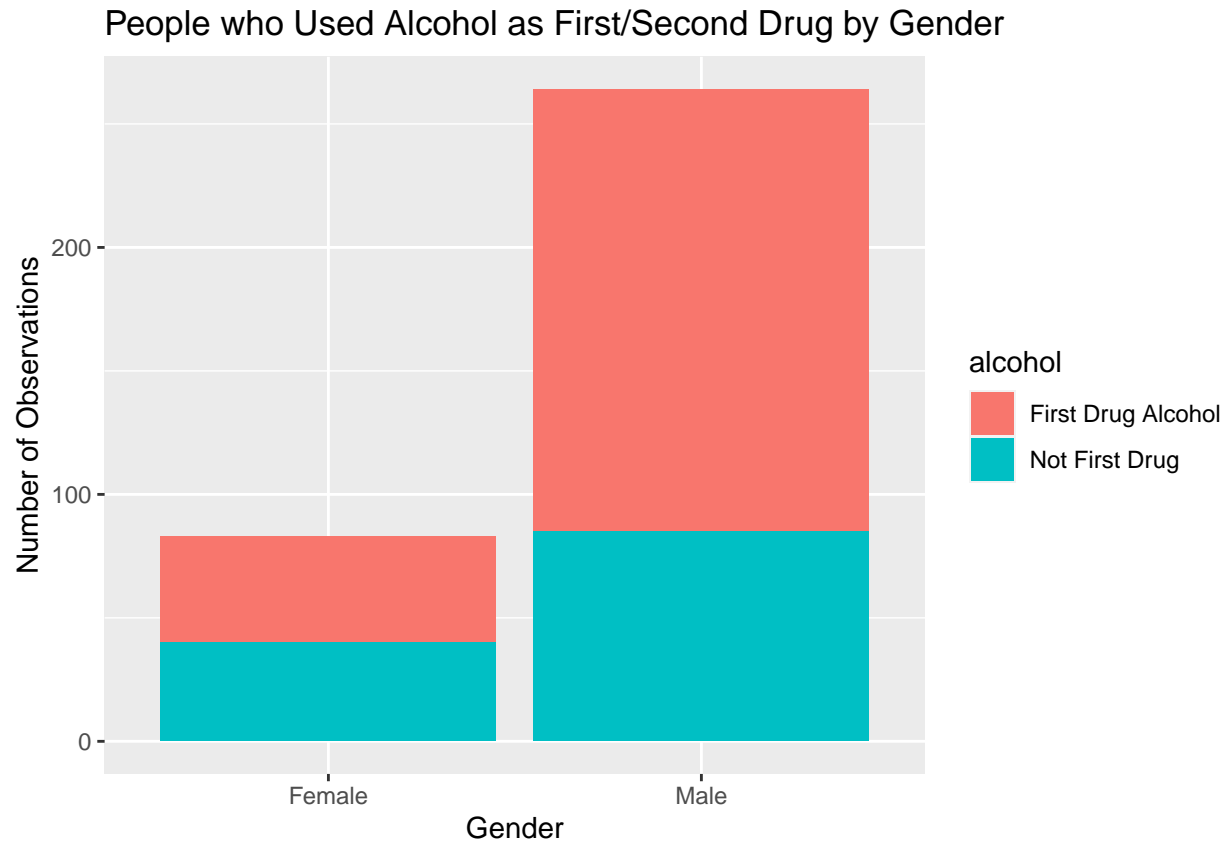
```



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")
```

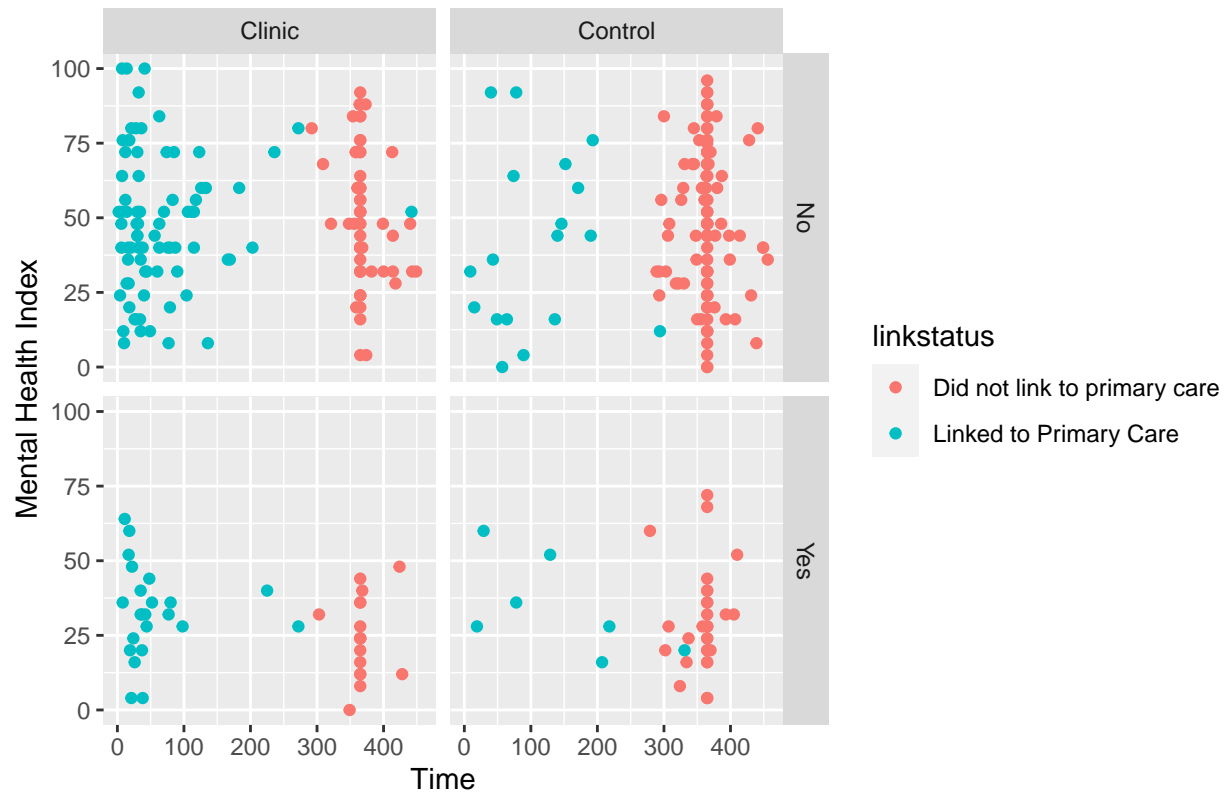


Note that the number of female participants are way less than the number of male participants. (talk about how this factors into not using Gender in final model)

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))
```

Mental Health Index Grouped by Attempted Suicide and Study Response

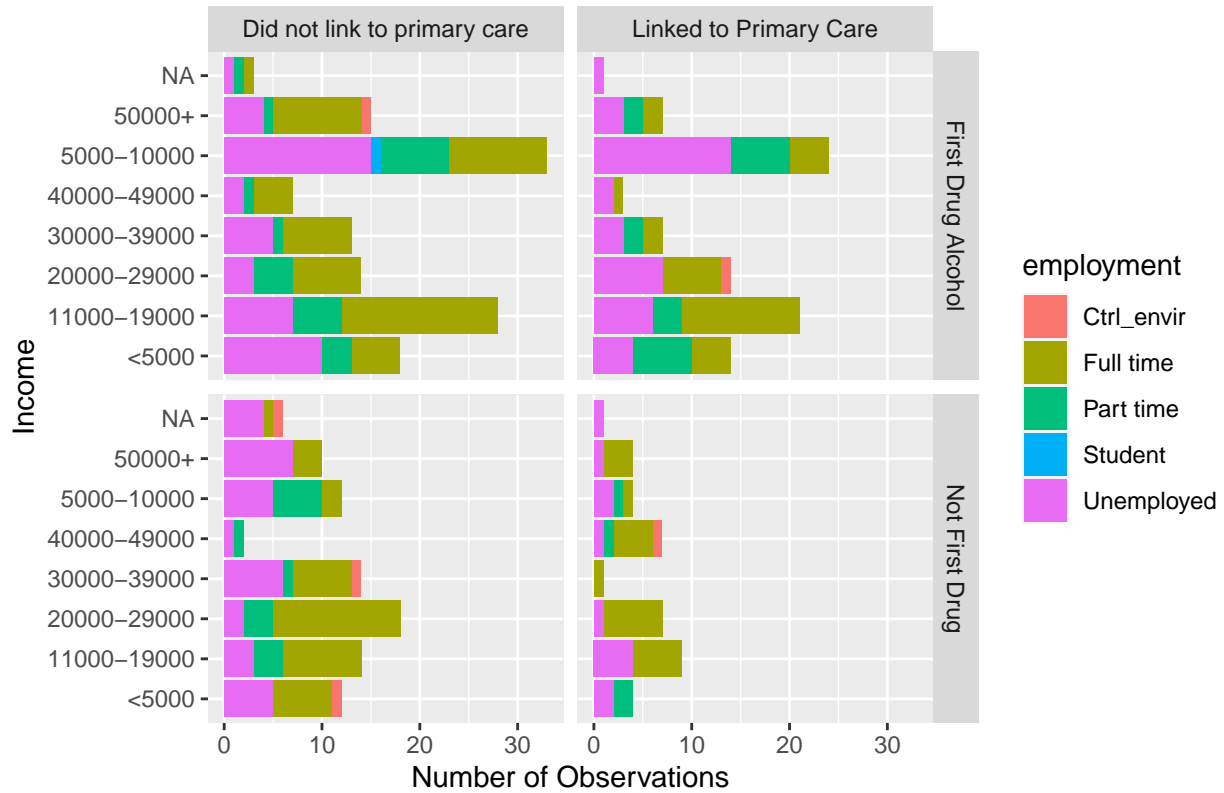


We see a distinct separation in response variable divided by the clinic and control group. In general, more individuals in the clinic group linked to primary care sooner than the control group. On the y-axis, a lower mental health index seems to be associated with more suicide attempts.

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()
```

Primary Care Status Based on Income and Alcohol 1st/2nd Drug



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 $pchisq(...)$, 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

```
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
```

Socioeconomic Explanatory Variables

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?

New Ideas

Cox.zph

Bootstrapping the Survival Model