EDA and Something New

Ethan Ashby

4/19/2021

Group members

Just me!

```
knitr::opts_chunk$set(message=FALSE, warning=FALSE, fig.height=5, fig.width=8,
                      fig.align = "center")
```

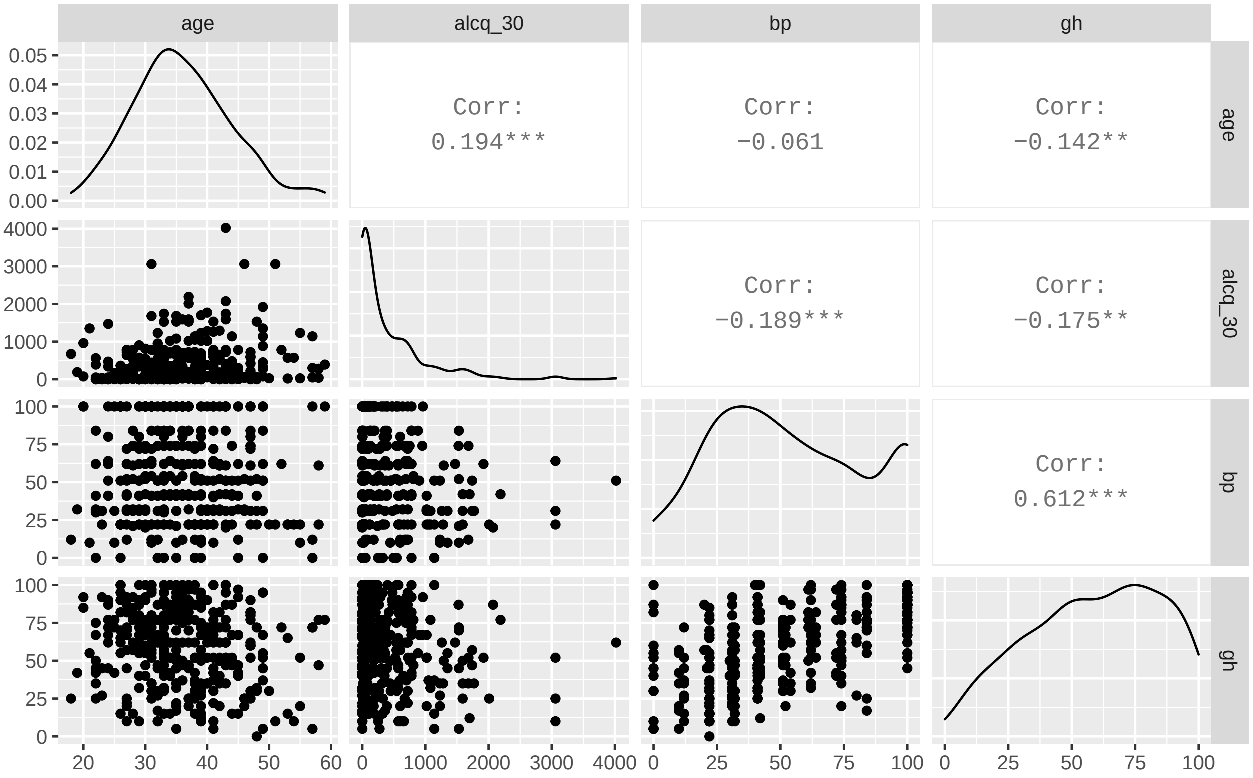
```
library(tidyverse)
library(GGally)
library(glmnet)
library(survival)
library(survminer)
```

Reading in the data and some EDA

```
data<-read.csv("HELPdata.csv")
#only permit columns with less than 10% missing values
num.permissible.na.values<-round(dim(data)[1]/10)

data_filter<-data[,colSums(is.na(data))<num.permissible.na.values]

data_filter %>% select(age, alcq_30, bp, gh) %>% ggpairs()
```

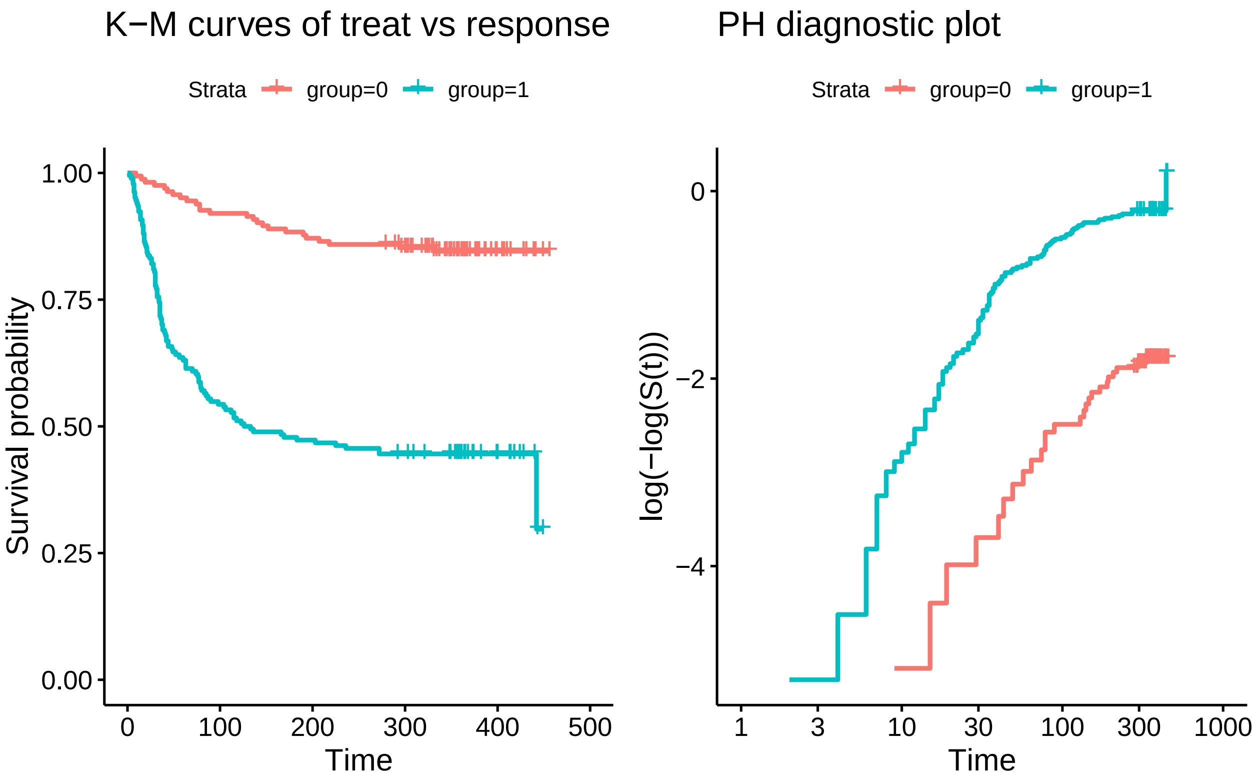


As a first step, I am going to read in the HELP data and remove columns that contain > 10% missing values. The resulting dataframe `data_filter`, contains 500 features summarizing 347 data points.

As a first pass at the data, I selected 4 continous variables `age`, `alcq_30` (total num drinks in last 30 days), `bp` (SF-36 pain index), and `gh` (SF-36 general health perceptions) and generated pairwise scatterplots. Based on these scatterplot, it looks to me like `bp` and `gh` are positively correlated variables, suggesting that they may be picking up on the same signals.

```
simple_surv<-survfit(Surv(dayslink, linkstatus)~group, data=data_filter)

p1<-survminer::ggsurvplot(simple_surv, censor=TRUE)+ggtitle("K-M curves of treat vs response")
p2<-survminer::ggsurvplot(simple_surv, censor=TRUE, fun="cloglog")+ggtitle("PH diagnostic plot")
arrange_ggsurvplots(list(p1, p2), ncol=2, nrow=1)
```



Next, I fit survival curves to the different groups and assessed the proportional hazards assumption using a $\log(-\log(S(t)))$ plot. The survival curves show drastically different survival probabilities (i.e. differences in scheduling an appointment with a physician) between the different treatment groups (multidisciplinary assessment and motivational intervention OR usual care). The relatively constant difference between the grouped $\log(-\log(S(t)))$ curves on the diagnostic plot supports the PH assumption.

In conclusion, this first pass at the data suggests to me that the `group` (treatment) variable conforms to the PH assumption and is likely very significant in estimating probability of visiting a physician at time t. Thus, I anticipate seeing the `group` variable included in my final model!

Something new

A major challenge when working with the HELP dataset is that the number of predictors outnumber the number of observations. In search of good generalization performance (i.e. on a held-out dataset), how does one elect which variables to include in the model and which to exclude?

The elastic net is a regularized regression approach that fits a model with constraints on the weighted average of the L_1 and L_2 norms of the β coefficients. Elastic net boasts the benefits of inducing model sparsity through the L_1 penalty, by forcing small βs to 0, and capturing sufficient model complexity by allowing the maximum number of features to exceed n (the number of observations) and by identifying groups of highly correlated features . In essence, the elastic net automatically does variable selection and selects entire groups of highly-related features to model. This can lead to a simpler model with improved generalization performance.

For this project, I will fit regularized Cox models with an elastic net penalty on the negative log of the partial likelihood via the R package `glmnet`. I will use the Coxnet vignette published by Tay, Simon, Friedman, Hastie, Tibshirani, and Narasimhan (2021) to guide my learning regarding how to implement the model. I believe my biggest challenge in this project will be understanding and explaining how elastic net is implemented in the context of Cox regression.

Data dictionary

Data dictionary