Arm_Guy_150Project_Assign1_due_Apr19

Arm and Guy

April 19, 2021

Assignment Instructions

- Include the names of the individuals in the group.
- Some sort of EDA (exploratory data analysis). Do not print the data, but do something thatindicates you've uploaded the data and know what some of the variables are. You might havesome summary statistics or a graph. This isnotan extensive assignment.
- Outline the "something new" part of the assignment. You should indicate who is doing what,what resources each of you will use to learn about your new topic, and a few sentences on whatthe topic is or how it relates to survival analysis / the analysis at hand. Additionally, for each "new" thing, provide 1-2 sentences describing what will be challenging about learning something new

1. Names of Group Members

a. Arm Wonghirundacha
b. Guy Thampakkul

2. Explnatory Data Analysis and Visuals

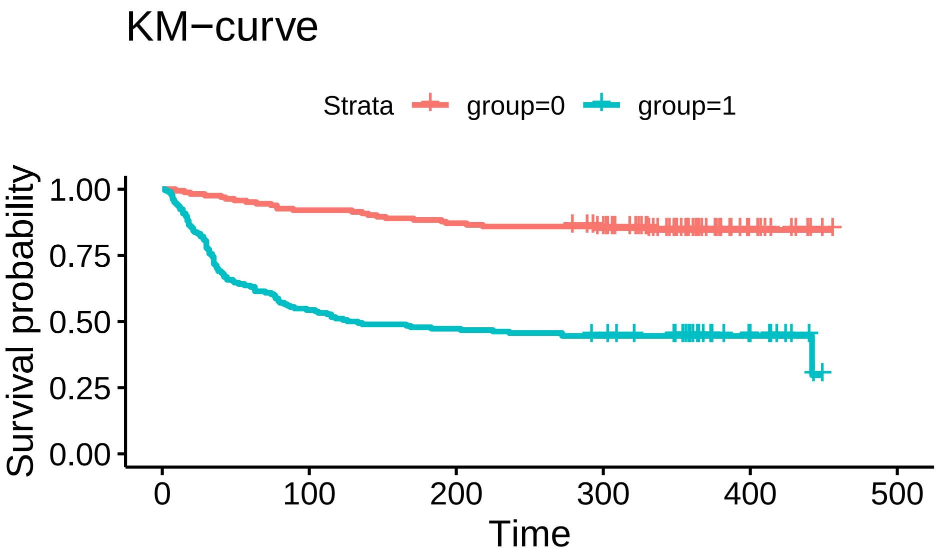
Reference Data Set: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1360-0443.2003.00328.x?sid=nlm%3Apubmed

```r
urlfile="https://raw.githubusercontent.com/gthampak/Arm_Guy_MATH150_Project/main/HELPdata.csv"

HELPdata <- read_csv(url(urlfile))
```

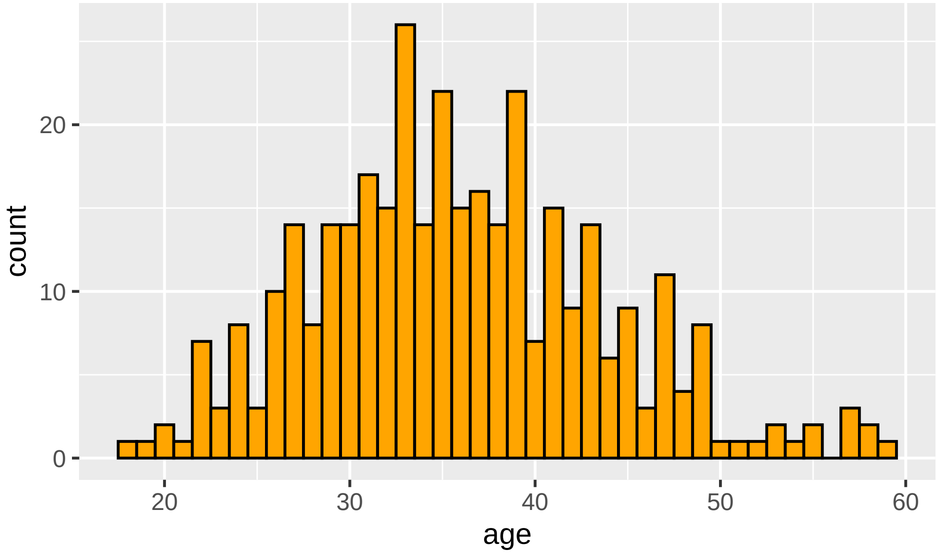
```r
HELPdata_survfit <- survfit(Surv(dayslink, linkstatus) ~ group, data=HELPdata)

ggsurvplot(HELPdata_survfit, conf.type = "TRUE") +
  ggtitle("KM-curve")
```

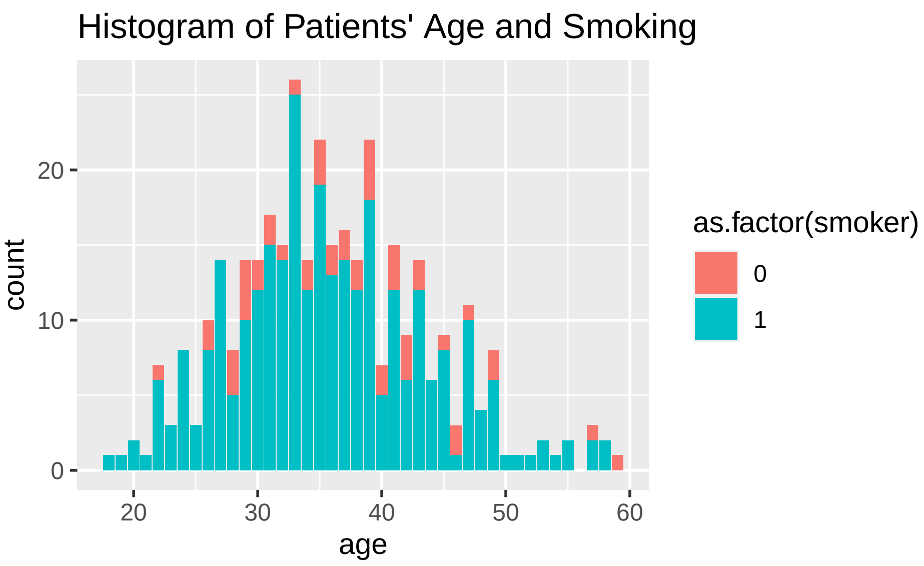


```
#histogram of age

ggplot(HELPdata, aes(x=age, fill = as.factor(smoker))) +
  geom_histogram(binwidth = 1, color="black", fill="orange")
```



```
ggplot(data=HELPdata) +
  geom_bar(mapping=aes(x=age, fill=as.factor(smoker))) +
  ggtitle("Histogram of Patients' Age and Smoking")
```



```
table(HELPdata$abuse)
```

```
##
##   0   1   2
## 101 180  64
```

3. Something New Outline

Topics (and possible sources):

- Investigation of the proportional hazards assumption (what does the R function *cox.zph* do?)

http://www.sthda.com/english/wiki/cox-model-assumptions

https://www.rdocumentation.org/packages/survival/versions/3.2-7/topics/cox.zph

https://bookdown.org/sestelo/sa_financial/how-to-evaluate-the-ph-assumption.html

- Exponential or Weibull PH regression (parametric survival model)

https://core.ac.uk/download/pdf/5172563.pdf

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5233524/

Weibull regression provides estimate of baseline hazard function in addition to coefficients for covariates. It is seldom used (as compared to semi-parametric proportional hazard model) because of its techinical difficulties.

- Deriving / detailing AIC & BIC for model selection on Cox PH

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2344147/

- Power analysis (a simulation?)
- Derivation of the sample size calculation for the log rank test (and application to the data)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2692036/

- An analysis of the Schoenfeld residuals (how are they calculated and why is that calculation relevant?)

https://towardsdatascience.com/schoenfeld-residuals-the-idea-that-turned-regression-modeling-on-its-head-b1f1fd293f87

- Bootstrapping the survival model (what are the assumptions? what do you conclude?)
- An analysis of possible time dependent covariates (do transformations help?)

https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf

- An analysis / understanding / simulation of the multiple comparisons issues for assessing many differentmodels (or other exploratory hypotheses).
- Andersen-Gill extension of the Cox PH model for time-varying covariates (available in rms R package).

https://rdrr.io/cran/rms/man/cph.html (R instructions)

http://finzi.psych.upenn.edu/library/rms/html/cph.html (R instructions)

https://projecteuclid.org/journals/annals-of-statistics/volume-10/issue-4/Coxs-Regression-Model-for-Counting-Processes--A-Large-Sample/10.1214/aos/1176345976.full

- Another topic related to survival analysis that you find interesting.
- Misc

Variable selection techniques for the Cox proportional hazards model: A comparative study: https://core.ac.uk/download/pdf/152600668.pdf

Arm -

Guy - New Topic(s): Investigation of the proportional hazards assumption

The function cox.zph() tests the proportional hazards assumption for each covariate included in a Cox regression model fit. I want to learn more about this because I think it is important in model building to know the assumptions really well in order to understand and/or predict any flaws the models we produce/report maay have. I think the challenge with learning this is to make connections between the theory (the assumptions themselves) and how it translates to model interpretation with real data and real models. I think that learning the assumptions will be fairly straightforward especially if there are concrete examples. However, the challenge will be to translate what we learn to the models we build because often, the examples used to explain/teach a new concept are 'perfect' whereas the models that we build from real data can have gray areas (which is what makes model building an art :)). For this topic, I plan to use the R documentation to learn about the *cox.zph* function and also look at real examples in statistical analysis papers to see how researchers use it to assess or interpret their models (links above).

(Will probably do an additional new something if time permits since learning about cox.zph might not take long. I suspect it is a topic that is on the simple side of things from the list of potential topics, but I also think understanding model assumptions and potential flaws in models that don't exactly fit assumptions is important.)