Time Until Relapse Among Smokers

Sam Kinghorn

2023-01-26

1 Introduction

This project looks at the data produced by Steinberg et al. (2009) in their randomized clinical trial from 2005 to 2007. They collected data on 127 medically ill smokers for the purposes of studying two treatments for smoking cessation: nicotine patch alone or a triple-combination intervention consisting of nicotine patch, nicotine oral inhaler, and bupropion. The nicotine patch was administered over a 10 week tapering program, as described for the patch producers, while the triple-combination therapy was given as needed, with a focus on tapering. The examination period began with a target quit date and the administration of the therapies. There is a relatively high amount of censorship, with more than 25% of subjects lost to follow-up.

The goal of my analysis is to determine the efficacy of the two treatments, patch only vs. triple-combination therapy, in the time until relapse, or cessation time, of smokers. Since my focus is on time until relapse, this analysis will be constrained to a survival analysis where relapse is the event of interest.

2 Data

The data contains demographic and lifestyle information on each subject, like age, gender, race, employment, number of years smoking, level of smoking, number of attempts to quit, and longest period of time without smoking. The primary variable of interest is the group in which the subject was randomly assigned to treatment. Additionally, there are survival variables describing time until relapse and whether the observation was censored or not.

3 EDA

In this section I'll describe and visualize the data. To start, let's look at the data itself.

```
# packages
library(asaur)
library(tidyverse)
library(survival)
library(ggfortify)
```

```
# load data
data("pharmacoSmoking")
smoking = pharmacoSmoking
head(smoking)
```

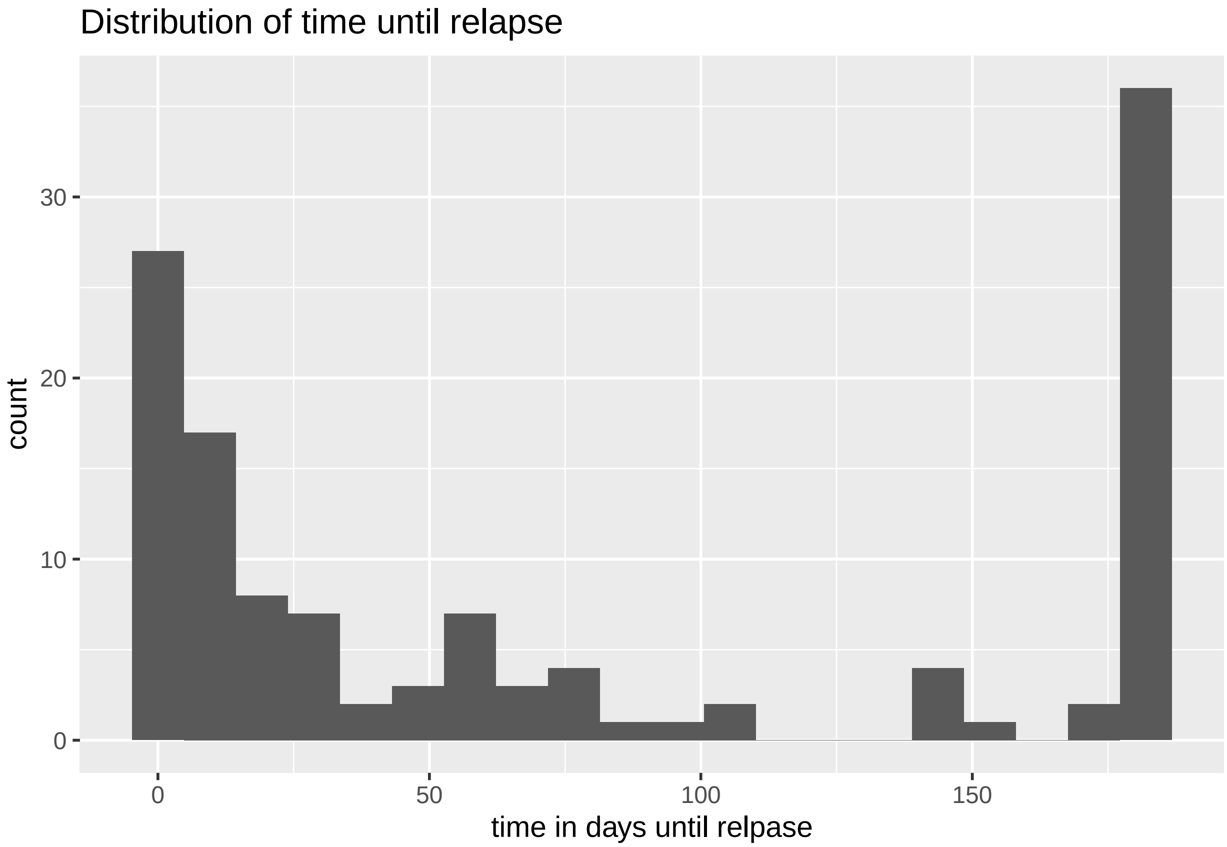
```
##    id ttr relapse         grp age gender     race employment yearsSmoking
## 1  21 182       0   patchOnly  36   Male    white         ft           26
## 2 113  14       1   patchOnly  41   Male    white      other           27
## 3  39   5       1 combination  25 Female    white      other           12
## 4  80  16       1 combination  54   Male    white         ft           39
## 5  87   0       1 combination  45   Male    white      other           30
## 6  29 182       0 combination  43   Male hispanic         ft           30
##   levelSmoking ageGroup2 ageGroup4 priorAttempts longestNoSmoke
## 1        heavy     21-49     35-49             0              0
## 2        heavy     21-49     35-49             3             90
## 3        heavy     21-49     21-34             3             21
## 4        heavy       50+     50-64             0              0
## 5        heavy     21-49     35-49             0              0
## 6        heavy     21-49     35-49             2           1825
```

```
dim(smoking)
```

```
## [1] 125  14
```

We can see the first six patients and their covariate patterns. There are a total of 125 observations with 11 features (not counting patient id, cessation time, and relapse status) that include a good mix of quantitative, qualitative, and ordinal variables. Next, I display the distribution of cessation times.

```
# cessation times
ggplot(data=smoking, mapping=aes(x=ttr)) +
  geom_histogram(bins=20) + labs(title="Distribution of time until relapse", x="time in days un
```



```
table(smoking$ttr)
```

```
##
##   0   1   2   3   4   5   6   7   8  10  12  14  15  16  20  21  25  28  30  40
##  12   5   6   1   3   2   1   1   3   1   2   7   4   1   1   2   1   3   3   1
##  42  45  49  50  56  60  63  65  75  77  80  84 100 105 110 140 155 170 182
##   1   1   1   1   5   2   2   1   1   2   1   1   1   1   1   4   1   2  36
```

```
# proportion at the end
count(filter(smoking, ttr==182))/count(smoking)
```

```
##       n
## 1 0.288
```

```
# proportion censored
count(filter(smoking, relapse==0))/count(smoking)
```

```
##       n
## 1 0.288
```

```
# proportion at time 0
count(filter(smoking, ttr==0))/count(smoking)
```

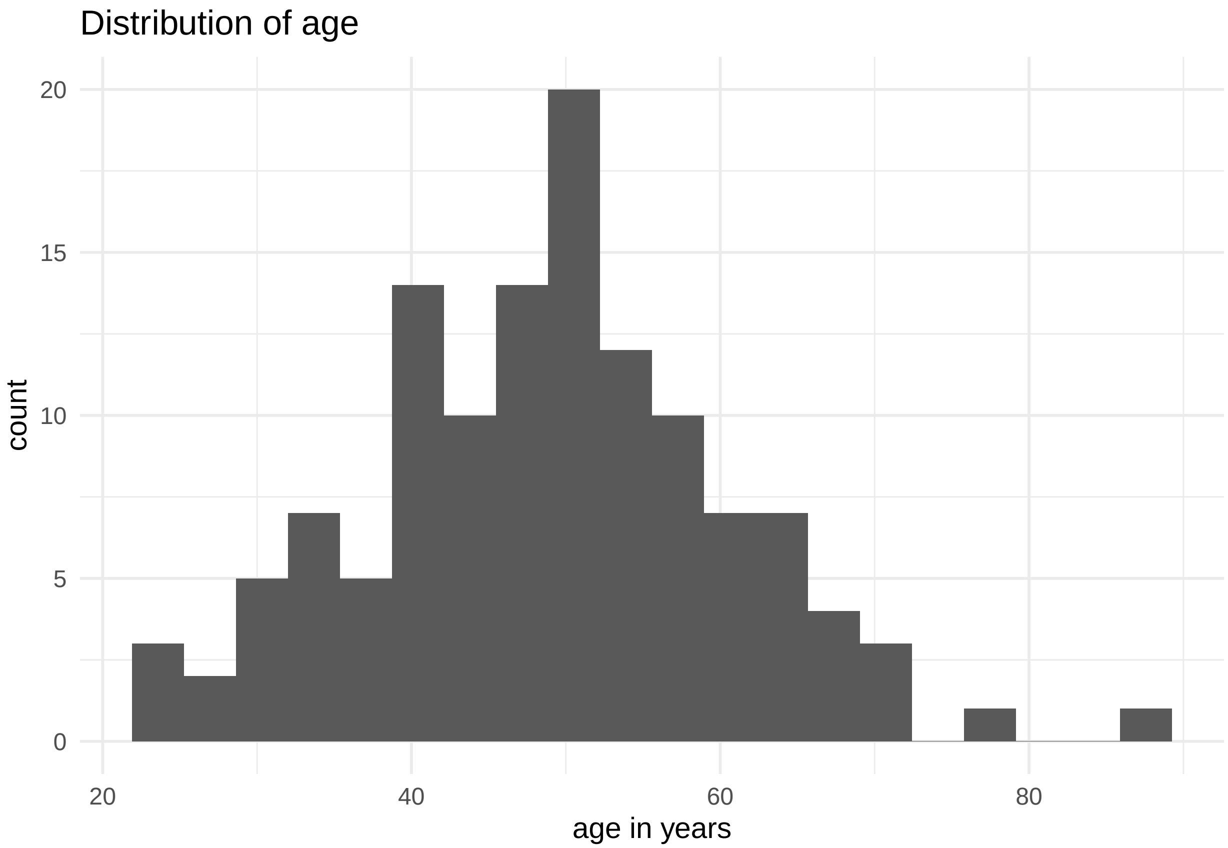
```
##       n
## 1 0.096
```

The distribution shows that in the first few days there were a lot of relapses, and the frequency gradually diminishes over 100 days. Also, there is a spike in the cessation time at the far end of the distribution. This spike represents 36 individuals, or 28% of the data, that made it to the end of the study window and were all censored, meaning we know that their cessation time is greater than 182 days. It's also worth noting that approximately 10% of the data had cessation times at 0. Next I display a handful of features.

```
# group assignment
table(smoking$grp)
```

```
##
## combination   patchOnly
##          61          64
```

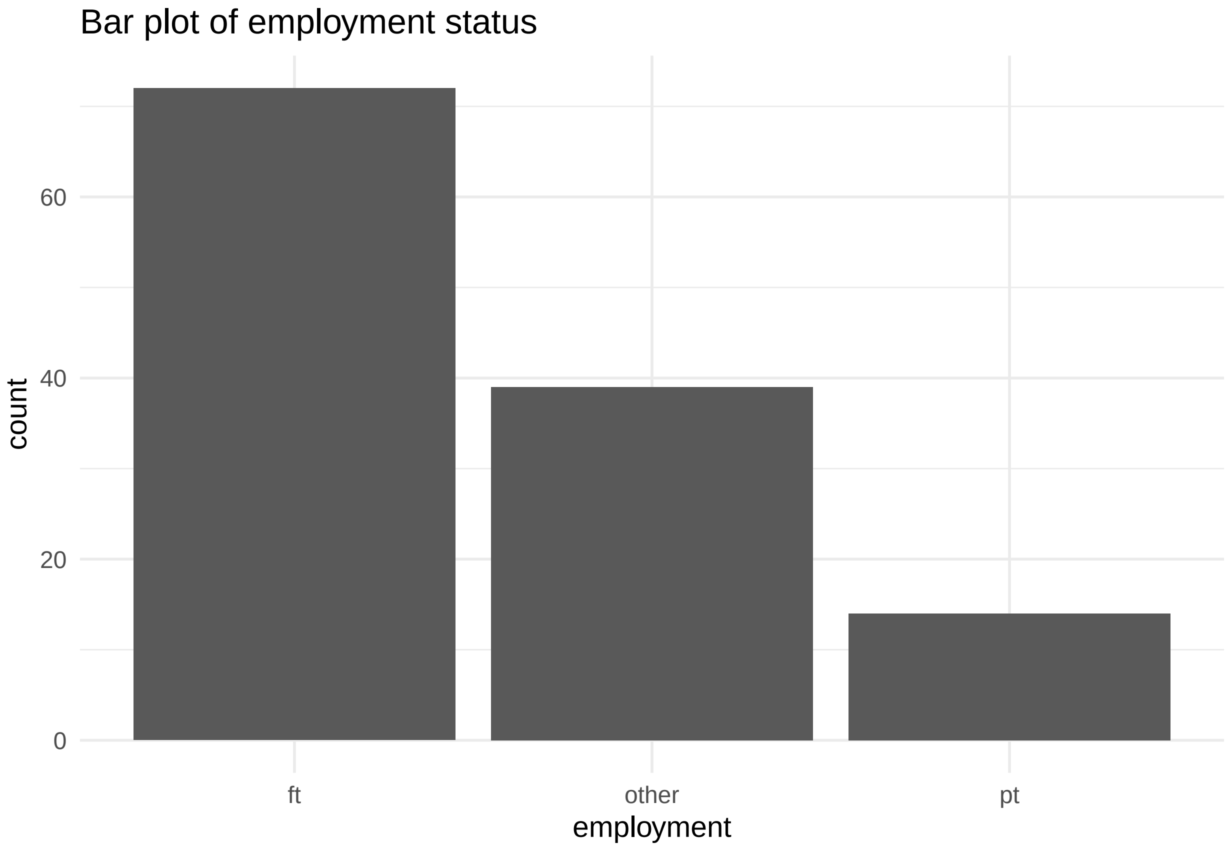
```
# age
ggplot(data=smoking, mapping=aes(x=age)) +
  geom_histogram(bins=20) + labs(title="Distribution of age",
                                  x="age in years") + theme_minimal()
```



```
# employment
table(smoking$employment)
```

```
##
##    ft other    pt
##    72    39    14
```

```
# bar plot employment
ggplot(data=smoking, mapping=aes(x=employment)) +
  geom_bar() + labs(title="Bar plot of employment status") + theme_minimal()
```



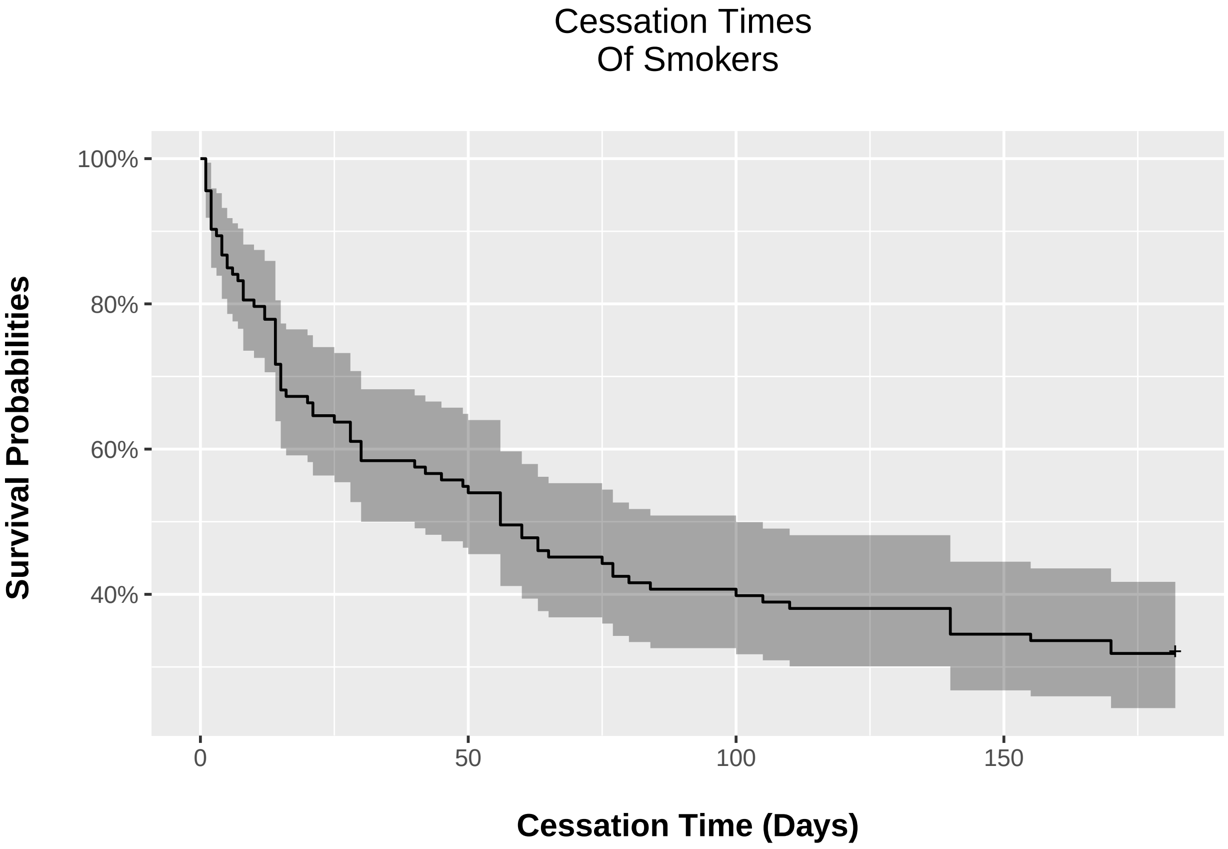
Treatment assignments are about evenly split and there is a good distribution of ages in the study. From the bar plot on employment status, we can see that the majority of subjects are full time workers, followed by other then part time.

Finally, we will look at the Kaplan-Meier survival curves. Before this, I remove all observations that have a failure time of 0. These observations probably relapsed within the first day of the study window, but their exact failure time is unknown. Since time 0 indicates the start of the study, when all subjects are in the risk set, it's not clear what a subject failing at time 0 would mean.

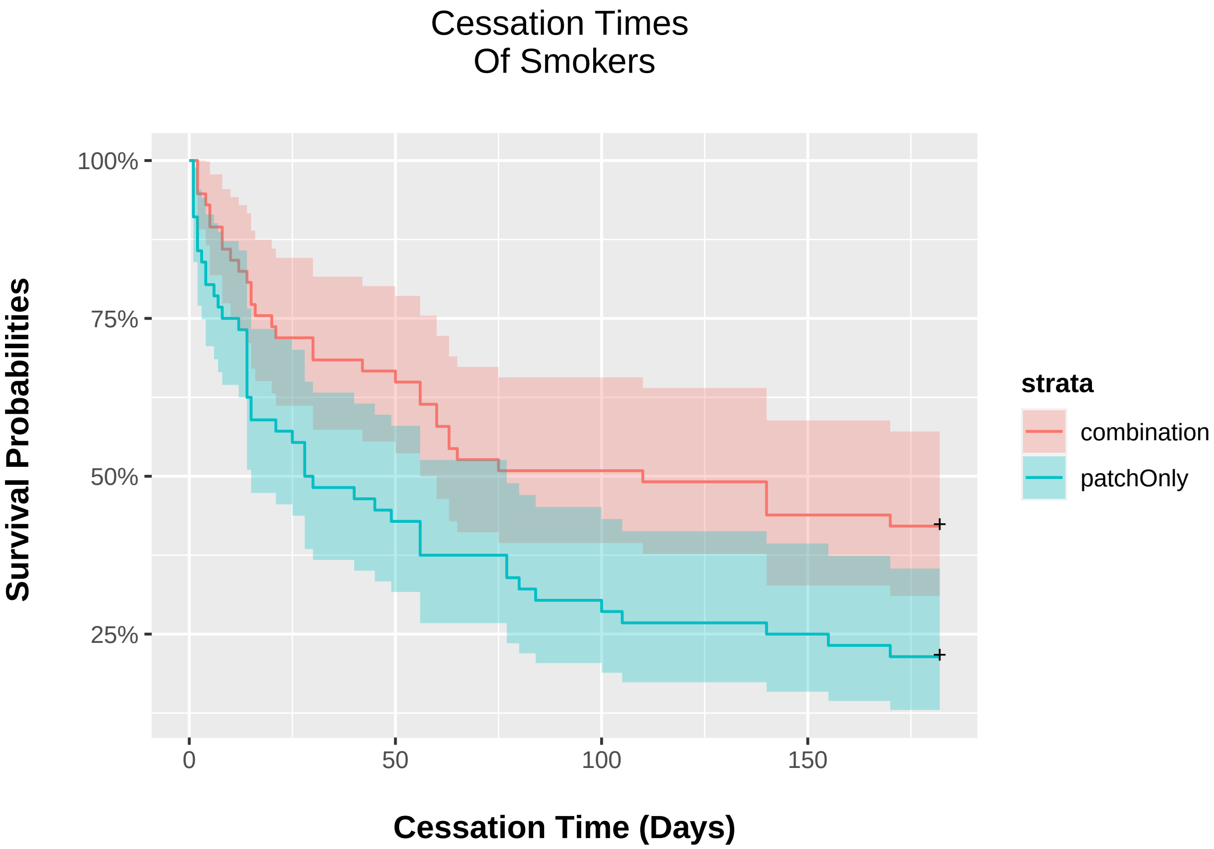
```
# remove time 0
df = filter(smoking, ttr!=0)

# create survival object
Y = Surv(df$ttr, df$relapse==1)

# intercept only
kmfit1 = survfit(Y ~ 1)
autoplot(kmfit1) +
  labs(x = "\n Cessation Time (Days) ", y = "Survival Probabilities \n",
       title = "Cessation Times \n Of Smokers \n") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title.x = element_text(face="bold", size = 12),
        axis.title.y = element_text(face="bold", size = 12))
```



```
# stratified by treatment
kmfit2 = survfit(Y ~ factor(grp), data=df)
autoplot(kmfit2) +
  labs(x = "\n Cessation Time (Days) ", y = "Survival Probabilities \n",
       title = "Cessation Times \n Of Smokers \n") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title.x = element_text(face="bold", size = 12),
        axis.title.y = element_text(face="bold", size = 12),
        legend.title = element_text(face="bold", size = 10))
```



```
# log rank test
survdiff(Y ~ factor(grp), data=df)
```

```
## Call:
## survdiff(formula = Y ~ factor(grp), data = df)
##
##                           N Observed Expected (O-E)^2/E (O-E)^2/V
## factor(grp)=combination 57       33     44.1      2.79      6.78
## factor(grp)=patchOnly   56       44     32.9      3.74      6.78
##
##  Chisq= 6.8  on 1 degrees of freedom, p= 0.009
```

The first graph shows the Kaplan-Meier survival curve for the entire data set. we can see that there is a steep decline in the early days of the study and the curve gradually flattens out. This means that initially, there was a high rate of relapse, then a decrease throughout the middle and end of the study. These observations agree with what we found in the distribution of cessation times.

The second graph shows the Kaplan-Meier survival curve stratified by treatment group. The curve for combination therapy is above the curve for patch only, indicating a longer cessation experience for the former. In other words, the patch only group had a shorter cessation time at every time point than the combination therapy. The median survival time for the combination group is 110 days, while the patch only group is 29. The code below the graph implements a log rank test on treatment group. This is a test to see if there is a significant difference in survival experience between the two groups. The null hypothesis is that there is no difference between survival curves. The test statistic is significant with a p-value of 0.009, so we reject the null hypothesis and conclude that there is a significant difference in the survival experience of the two treatment groups.

4 Cox Proportional Hazard Model

Now to fit a Cox proportional hazard (PH) model to the data. The Cox PH model is a popular survival model that is comprised of a time-dependent baseline hazard that is unspecified and an exponential function of a linear combination of covariates that is time-independent. The baseline hazard is not estimated or specified by parameter values which makes the Cox PH model semi-parametric. This model is popular because it has been shown to provide a robust approximation of survival experience when the proportional hazard assumption is nearly met. The proportional hazard assumption states that the hazard ratio between any two individuals is constant across time.

I fit multiple models and performed likelihood ratio tests to determine significant features. The result is a model with treatment group, age, and employment status as covariates to explain time until relapse. The output of the model can be seen below.

```
# cox ph model
cph_model = coxph(Y ~ grp + age + employment, data=df)
summary(cph_model)
```

```
## Call:
## coxph(formula = Y ~ grp + age + employment, data = df)
##
##   n= 113, number of events= 77
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## grppatchOnly     0.61153   1.84326  0.23388  2.615  0.00893 **
## age             -0.03098   0.96949  0.01189 -2.606  0.00915 **
## employmentother  0.63161   1.88064  0.29768  2.122  0.03385 *
## employmentpt     0.74409   2.10454  0.34582  2.152  0.03142 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## grppatchOnly       1.8433     0.5425    1.1655    2.9152
## age                0.9695     1.0315    0.9472    0.9923
## employmentother    1.8806     0.5317    1.0494    3.3705
## employmentpt       2.1045     0.4752    1.0686    4.1449
##
## Concordance= 0.631  (se = 0.033 )
## Likelihood ratio test= 16.79  on 4 df,   p=0.002
## Wald test            = 16.72  on 4 df,   p=0.002
## Score (logrank) test = 17.17  on 4 df,   p=0.002
```

5 Results

The hazard ratios for the Cox PH model are below.

```
# hazard ratios
exp(cph_model$coefficients)
```

```
##    grppatchOnly             age employmentother    employmentpt
##       1.8432570       0.9694927       1.8806447       2.1045359
```

We can interpret the regression coefficients as the hazard ratio, or the hazard for one individual divided by the hazard of another individual. For example, if we compare two individuals that have the same covariate pattern but are in two different treatment groups, the ratio of their hazards will reduce to the exponential of the coefficient associated with the treatment variable. Thus, in our case, the coefficient for treatment group is 0.61153 and the exponential of this is 1.8433. This means that the hazard for the patch only group is 1.8433 times the hazard for the combination group, controlling for age and employment status. In other words, a patch only subject who has not relapsed by a certain time has about 1.8 times the chance of relapsing at the next point in time compared to someone in the combination group. As Spruance et al. (2004) note, this corresponds to about a 64% chance of the patch only group relapsing first. The hazard ratio for a year increase in age is about 0.97. Similarly, the hazard ratio associated with other employment status vs. full-time is 1.88, and the hazard ratio associated with part-time vs. full-time is 2.10.

In summary, the model suggests that combination therapy, as opposed to just patch only, prolongs time until relapse. Also, increases in age and having a full-time job status, relative to part-time and other, are associated with longer cessation time.

6 Assumptions

The central assumption of the Cox PH model is that the hazard ratio is constant across time, or the hazards of any two individuals are proportional across time. If this assumption is violated, the model is invalid. We can assess the proportional hazard assumption with a statistical test. As Kleinbaum, Klein, et al. (2012) state, the test checks the correlation between the Schoenfeld residuals and survival time. The null hypothesis is a correlation of 0, which supports the proportional hazards assumption. The output of the test is below. The p-values for each covariate are not significant and the global test for the entire model is not significant. This means we fail to reject the null hypothesis that the proportional hazards assumption is met. In other words, the proportional hazards assumption appears to be valid.

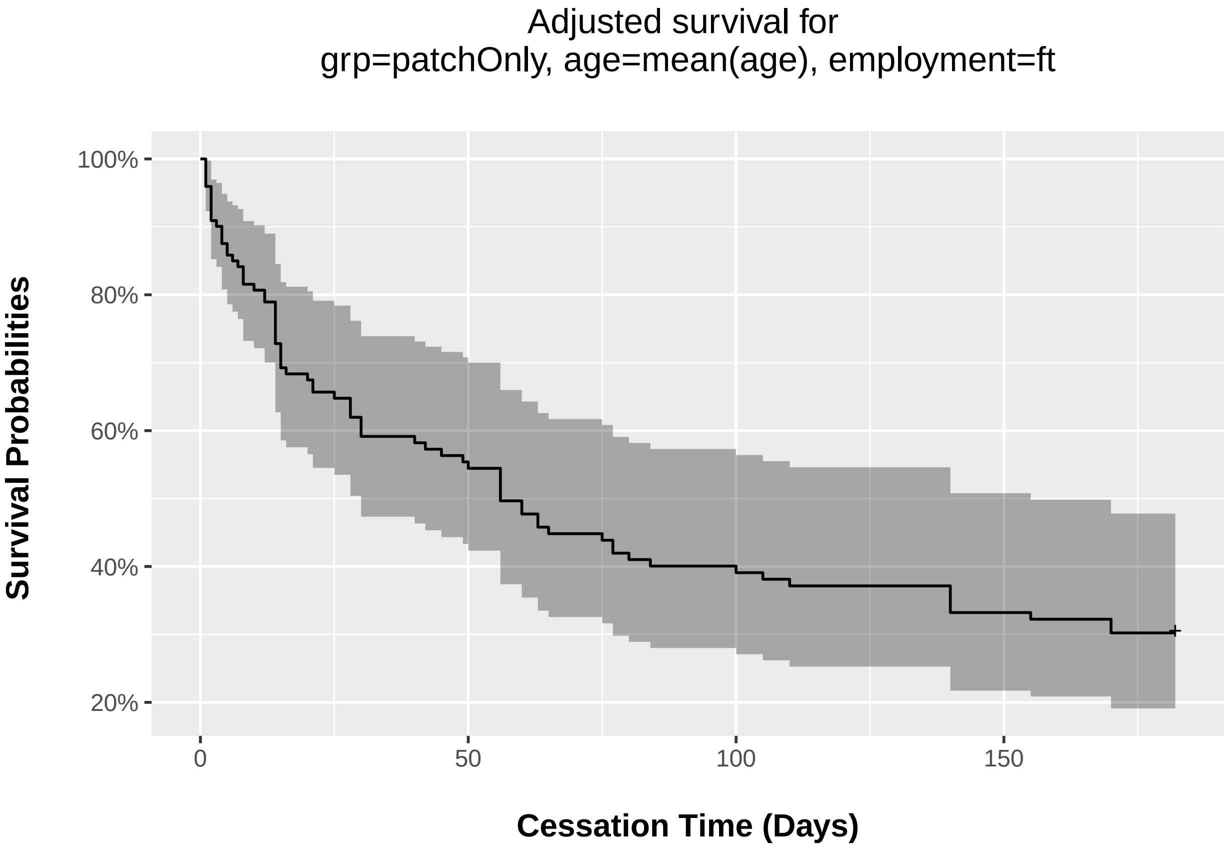
```
# ph test
cox.zph(cph_model, transform=rank)
```

```
##            chisq df    p
## grp        0.150  1 0.70
## age        0.425  1 0.51
## employment 0.406  2 0.82
## GLOBAL     0.831  4 0.93
```

7 Prediction and Survival Curves

For any covariate pattern, or individual, we can display the Cox-adjusted survival curve. Say we are interested in an individual in the patch only group with mean age and full-time employment. Their adjusted survival curve can be seen below. We can say that the median cessation time for someone with this covariate pattern is just over 50 days.

```
# adjusted survival curve
pattern1 = data.frame(grp='patchOnly', age=mean(df$age), employment='ft')
autoplot(survfit(cph_model, newdata=pattern1)) +
  labs(x = "\n Cessation Time (Days) ", y = "Survival Probabilities \n",
       title = "Adjusted survival for \n grp=patchOnly, age=mean(age), employment=ft \n") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title.x = element_text(face="bold", size = 12),
        axis.title.y = element_text(face="bold", size = 12),
        legend.title = element_text(face="bold", size = 10))
```



8 Conclusion

This analysis considered data from a randomized clinical trial on the effects of two treatments on smoking cessation, or time until relapse, among medically ill smokers. The results of my analysis indicate that triple combination therapy of nicotine patch, nicotine oral inhaler, and bupropion is more effective in extending cessation time than patch only. More precisely, there is about a 64% chance of the patch only group relapsing first, controlling for age and employment status.

9 References

Kleinbaum, David G, Mitchel Klein, et al. 2012. *Survival Analysis: A Self-Learning Text.* Vol. 3. Springer.

Spruance, Spotswood L., Julia E. Reid, Michael Grace, and Matthew Samore. 2004. "Hazard Ratio in Clinical Trials." *Antimicrobial Agents and Chemotherapy* 48 (8): 2787–92. https://doi.org/10.1128/AAC.48.8.2787-2792.2004.

Steinberg, Michael B., Shelley Greenhaus, Amy C. Schmelzer, Michelle T. Bover, Jonathan Foulds, Donald R. Hoover, and Jeffrey L. Carson. 2009. "Triple-Combination Pharmacotherapy for Medically Ill Smokers." *Annals of Internal Medicine* 150 (7): 447–54. https://doi.org/10.7326/0003-4819-150-7-200904070-00004.