BIOST544_HW1

Aaron Wolf

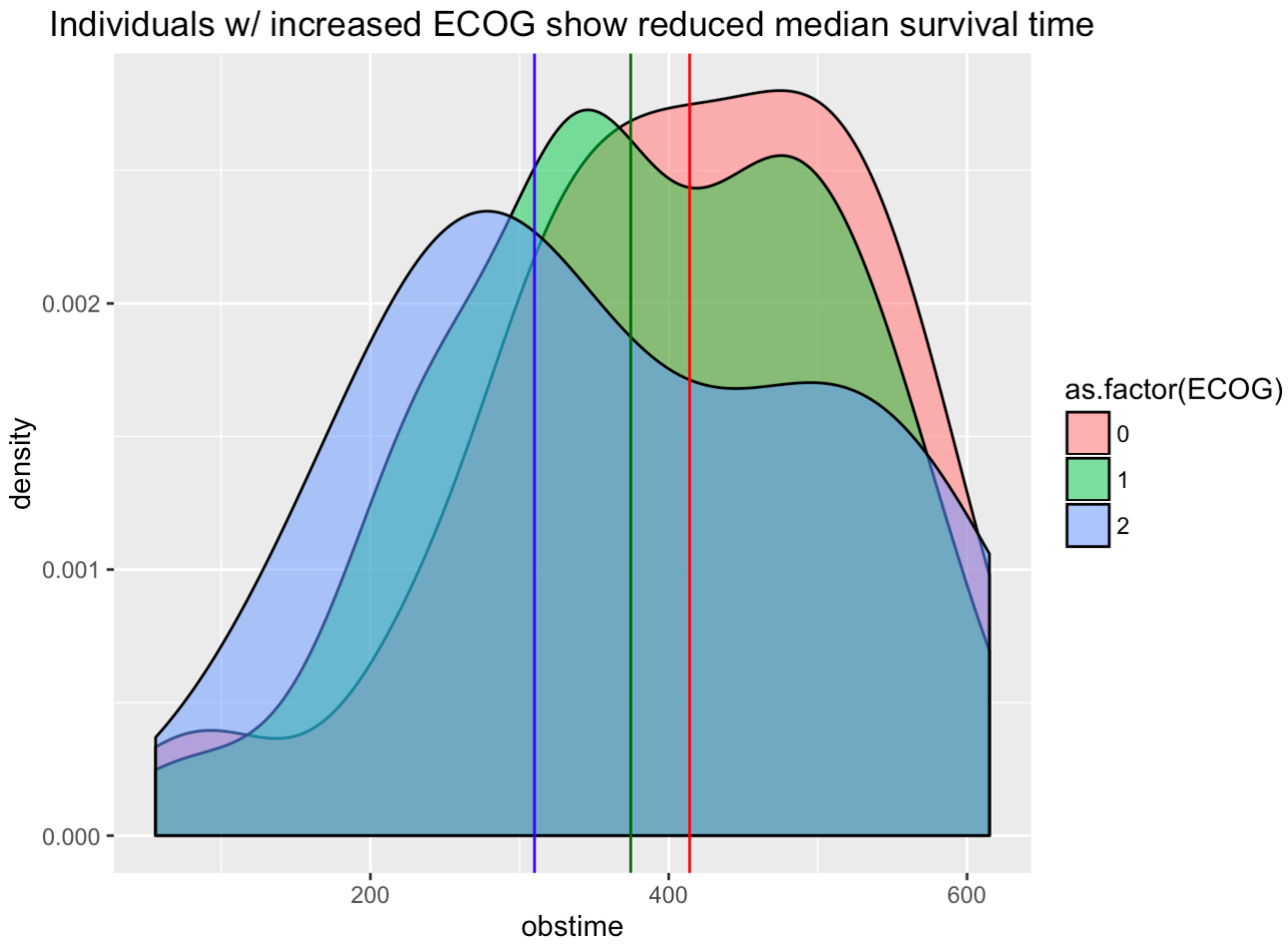
January 8, 2017

```
nsclc <- as.data.table(read.table('~/Documents/Dropbox/2016:2017/BIOST544/Data/nsclc-modified.txt'))
```

1)

a. Increased ECOG score at baseline has decreased median survival time (as measured by observation time). Relationship demonstrated graphically.

```
ggplot(data=nsclc) +
  geom_density(aes(x=obstime, fill=as.factor(ECOG)), alpha=0.5) +
  geom_vline(xintercept = median(filter(nsclc, ECOG==0)$obstime), color='red') +
  geom_vline(xintercept = median(filter(nsclc, ECOG==1)$obstime), color='darkgreen') +
  geom_vline(xintercept = median(filter(nsclc, ECOG==2)$obstime), color='blue') +
  ggtitle('Individuals w/ increased ECOG show reduced median survival time')
```

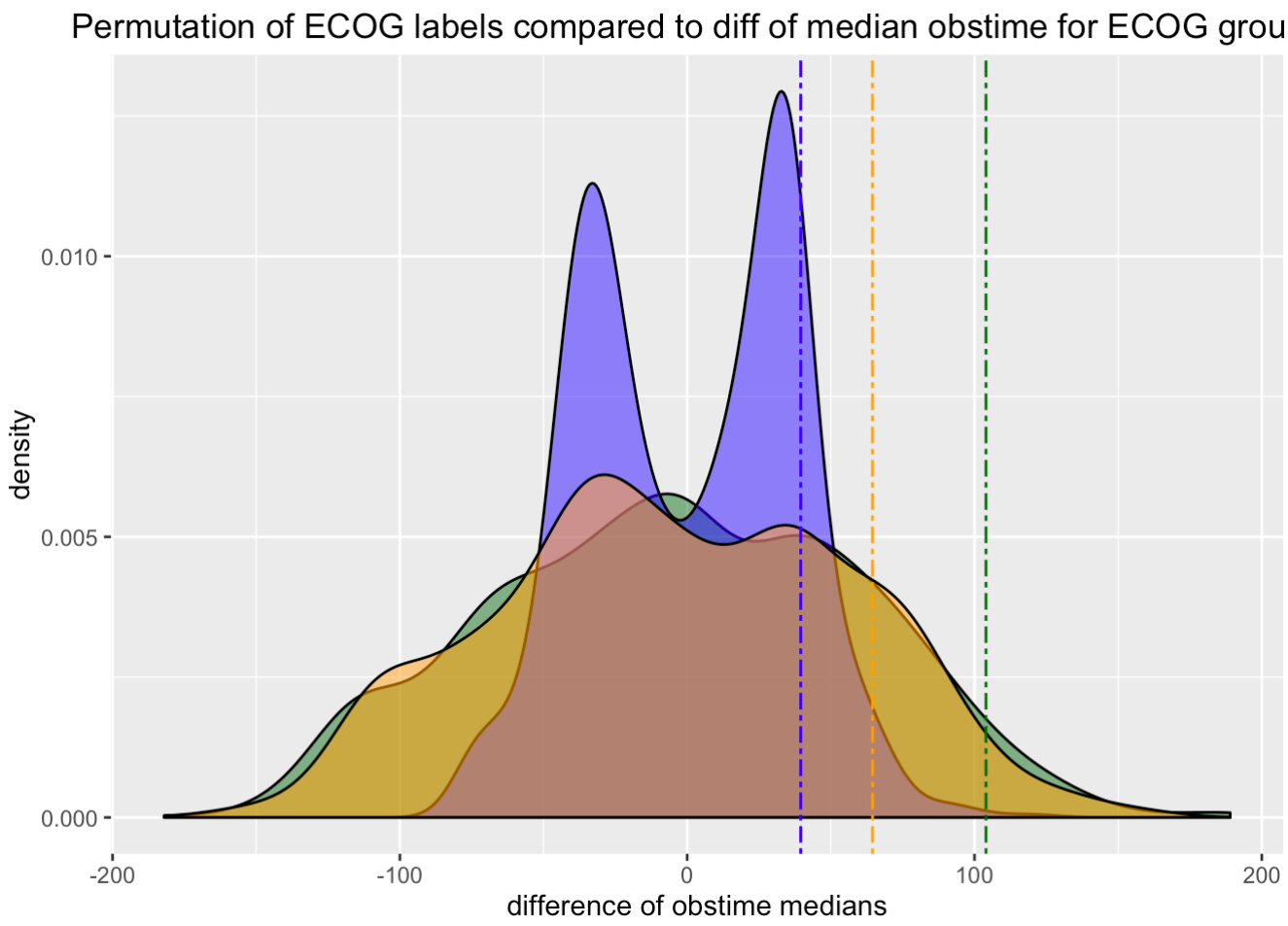


Permute ECOG labels, check difference of median obstimes

```
permute_data.fn = function(data){
  permutation <- sample(1:nrow(data), replace=FALSE)
  permutation.data <- data
  permutation.data$ECOG = data$ECOG[permutation]

  median.obstime.0_2 = with(permutation.data,
       (median(obstime[ECOG==0]) - median(obstime[ECOG==2]))
       )
  median.obstime.0_1 = with(permutation.data,
       (median(obstime[ECOG==0]) - median(obstime[ECOG==1]))
       )
  median.obstime.1_2 = with(permutation.data,
       (median(obstime[ECOG==1]) - median(obstime[ECOG==2]))
       )

  return(c(median.obstime.0_2, median.obstime.0_1, median.obstime.1_2))
}


permuted.diff_median <- as.data.table(t(replicate(n = 1000, expr = permute_data.fn(data =
nsclc))))
setnames(permuted.diff_median , c('ECOG_0.ECOG_2','ECOG_0.ECOG_1','ECOG_1.ECOG_2'))
```

Graphical representaion of difference in obstime for ECOG groups after permutation. After permutation, all ECOG groups show centering at 0 for median difference in obstimes.

```
ggplot(data=permuted.diff_median) +
  geom_density(aes(x=ECOG_0.ECOG_2), fill='darkgreen', alpha=0.5, show.legend = TRUE) +
  geom_density(aes(x=ECOG_0.ECOG_1), fill='blue', alpha=0.5) +
  geom_density(aes(x=ECOG_1.ECOG_2), fill='orange', alpha=0.5) +
  geom_vline(xintercept = (median(filter(nsclc, ECOG==0)$obstime) - median(filter(nsclc,
ECOG==2)$obstime)), color="darkgreen", linetype='twodash') +
  geom_vline(xintercept = (median(filter(nsclc, ECOG==0)$obstime) - median(filter(nsclc,
ECOG==1)$obstime)), color="blue", linetype='twodash') +
  geom_vline(xintercept = (median(filter(nsclc, ECOG==1)$obstime) - median(filter(nsclc,
ECOG==2)$obstime)), color="orange", linetype='twodash') +
  xlab(label = 'difference of obstime medians') +
  ggtitle('Permutation of ECOG labels compared to diff of median obstime for ECOG groups') +
  scale_color_manual(values = c(ECOG_0.2="red", ECOG_0.1="blue", ECOG_1.2="orange"))
```



I highlihghted on the graph above the position of the median difference in obstime for the empirical data (i.e. non-simulated). I then calcualte the tail probabilities for these medians relative to the distribution of median obstime differences for the simulated data. I find that the difference in obstime is greatest between groups ECOG.0 and ECOG.2 (~96% of simulated data is less that the empirical measure).

```
#Calculate tail probabilities
mean(permuted.diff_median$ECOG_0.ECOG_2 <= (median(filter(nsclc, ECOG==0)$obstime) - median(filte
r(nsclc, ECOG==2)$obstime)))
mean(permuted.diff_median$ECOG_0.ECOG_1 <= (median(filter(nsclc, ECOG==0)$obstime) - median(filte
r(nsclc, ECOG==1)$obstime)))
mean(permuted.diff_median$ECOG_1.ECOG_2 <= (median(filter(nsclc, ECOG==1)$obstime) - median(filte
r(nsclc, ECOG==2)$obstime)))
```

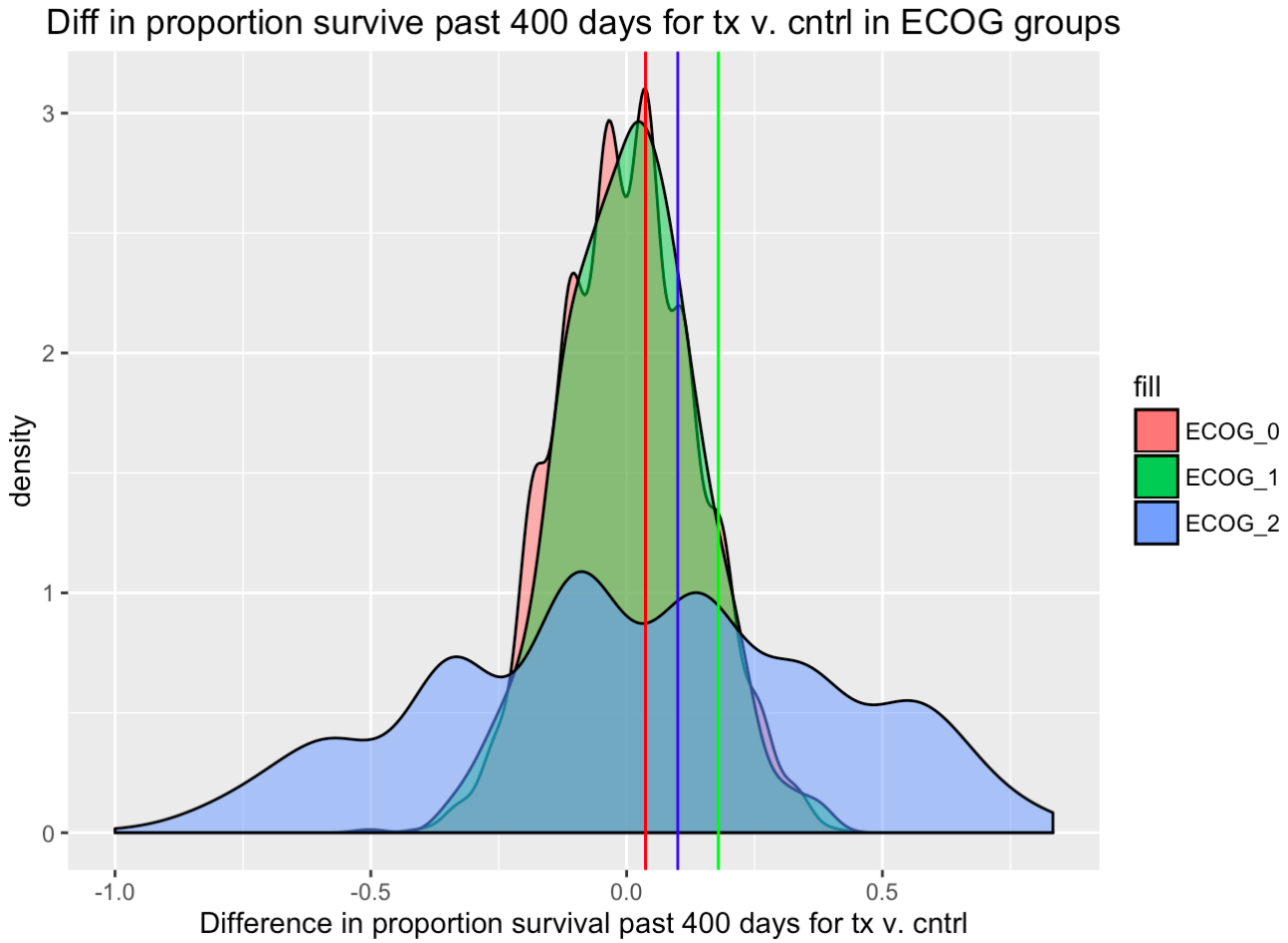
```
## [1] 0.964
## [1] 0.884
## [1] 0.843
```

b. How does ECOG score affect treatment outcome, i.e. is treatment better than control only for ECOG_0? I began by measuring the difference in survival proportion between treatment and control for each ECOG group. Then I simulated multiple trials, permuting tx label within ECOG groups so as to maintain within-group effect of ECOG and assess whether tx out performs control in each group. To assess whether the empirical data (difference in proportion survival btwn tx and control) show evidence of tx out performing control I then calculated tail probabilities comparing the empirical data to the simulated data. ECOG.1 shows the greatest difference in tx v. control performance, with the tx group having a higher proportion of survival past 400 days.

```
## What is the proportion of individuals that survived in ECOG_0 group
#nsclc %>% filter(ECOG==0) %>% summarise(mean(survival.past.400))
## If you were ECOG_0 and recieved treatment, you did slightly better than the group as a whole
#nsclc %>% filter(ECOG==0, tx==1) %>% summarise(mean(survival.past.400))
## If you were ECOG_0 and did not recieved placebo, you did slightly worse than the group as a wh
ole
#nsclc %>% filter(ECOG==0, tx==0) %>% summarise(mean(survival.past.400))

prop.survival = nsclc %>% group_by(ECOG, tx) %>% summarise(prop.survival.past.400 = mean(surviva
l.past.400)) %>% arrange(desc(tx))

diff.prop.survival.ECOG_0.empir = filter(prop.survival, ECOG==0 & tx==1)$prop.survival.past.400 -
                            filter(prop.survival, ECOG==0 & tx==0)$prop.survival.past.400

diff.prop.survival.ECOG_1.empir = filter(prop.survival, ECOG==1 & tx==1)$prop.survival.past.400 -
                            filter(prop.survival, ECOG==1 & tx==0)$prop.survival.past.400

diff.prop.survival.ECOG_2.empir = filter(prop.survival, ECOG==2 & tx==1)$prop.survival.past.400 -
                            filter(prop.survival, ECOG==2 & tx==0)$prop.survival.past.400

######################
diff.prop.survival.fn = function(data){
  #The permutation here shuffles the tx-label WITHIN a given ECOG group. ECOG group effect is mai
ntained, but within group difference between tx and placebo is examined. Is tx better than placeb
o in ECOG_0? Is this also true in ECOG_1 and ECOG_2?
  permutation.ECOG_0 <- sample(1:nrow(filter(data, ECOG==0)), replace=FALSE)
  permutation.ECOG_1 <- sample(1:nrow(filter(data, ECOG==1)), replace=FALSE)
  permutation.ECOG_2 <- sample(1:nrow(filter(data, ECOG==2)), replace=FALSE)

  permutation.data.ECOG_0 <- filter(data, ECOG==0)
  permutation.data.ECOG_1 <- filter(data, ECOG==1)
  permutation.data.ECOG_2 <- filter(data, ECOG==2)

  permutation.data.ECOG_0$tx = permutation.data.ECOG_0$tx[permutation.ECOG_0]
  permutation.data.ECOG_1$tx = permutation.data.ECOG_0$tx[permutation.ECOG_1]
  permutation.data.ECOG_2$tx = permutation.data.ECOG_0$tx[permutation.ECOG_2]

  permuted.data = rbind(permutation.data.ECOG_0, permutation.data.ECOG_1,
permutation.data.ECOG_2)

  dt.prop.survival = permuted.data %>% group_by(ECOG, tx) %>% summarise(prop.survival.past.400 =
mean(survival.past.400))

  diff.survival.ECOG_0 = filter(dt.prop.survival, ECOG==0 & tx==1)$prop.survival.past.400 -
                            filter(dt.prop.survival, ECOG==0 & tx==0)$prop.survival.past.400

  diff.survival.ECOG_1 = filter(dt.prop.survival, ECOG==1 & tx==1)$prop.survival.past.400 -
                              filter(dt.prop.survival, ECOG==1 & tx==0)$prop.survival.past.400

  diff.survival.ECOG_2 = filter(dt.prop.survival, ECOG==2 & tx==1)$prop.survival.past.400 -
                              filter(dt.prop.survival, ECOG==2 & tx==0)$prop.survival.past.400

  #dfdf.survival.ECOG.0_2 = diff.survival.ECOG_0 - diff.survival.ECOG_2
  return(c(diff.survival.ECOG_0, diff.survival.ECOG_1, diff.survival.ECOG_2))
}
######################
```

```
permutated.diff_survival = as.data.table(t(as.data.table(replicate(n = 1000, expr = diff.prop.sur
vival.fn(nsclc)))))
setnames(permutated.diff_survival, c('diff.survival.ECOG_0','diff.survival.ECOG_1','diff.surviva
l.ECOG_2'))
```

```
ggplot() +
  geom_density(data=permutated.diff_survival, aes(x=diff.survival.ECOG_0, fill='ECOG_0'),
alpha=0.5) +
  geom_density(data=permutated.diff_survival, aes(x=diff.survival.ECOG_1, fill='ECOG_1'),
alpha=0.5) +
  geom_density(data=permutated.diff_survival, aes(x=diff.survival.ECOG_2, fill='ECOG_2'),
alpha=0.5) +
  geom_vline(xintercept = diff.prop.survival.ECOG_0.empir, color='red') +
  geom_vline(xintercept = diff.prop.survival.ECOG_1.empir, color = 'green') +
  geom_vline(xintercept = diff.prop.survival.ECOG_2.empir, color = 'blue') +
  ggtitle(label = 'Diff in proportion survive past 400 days for tx v. cntrl in ECOG groups') +
  xlab(label ='Difference in proportion survival past 400 days for tx v. cntrl')
```



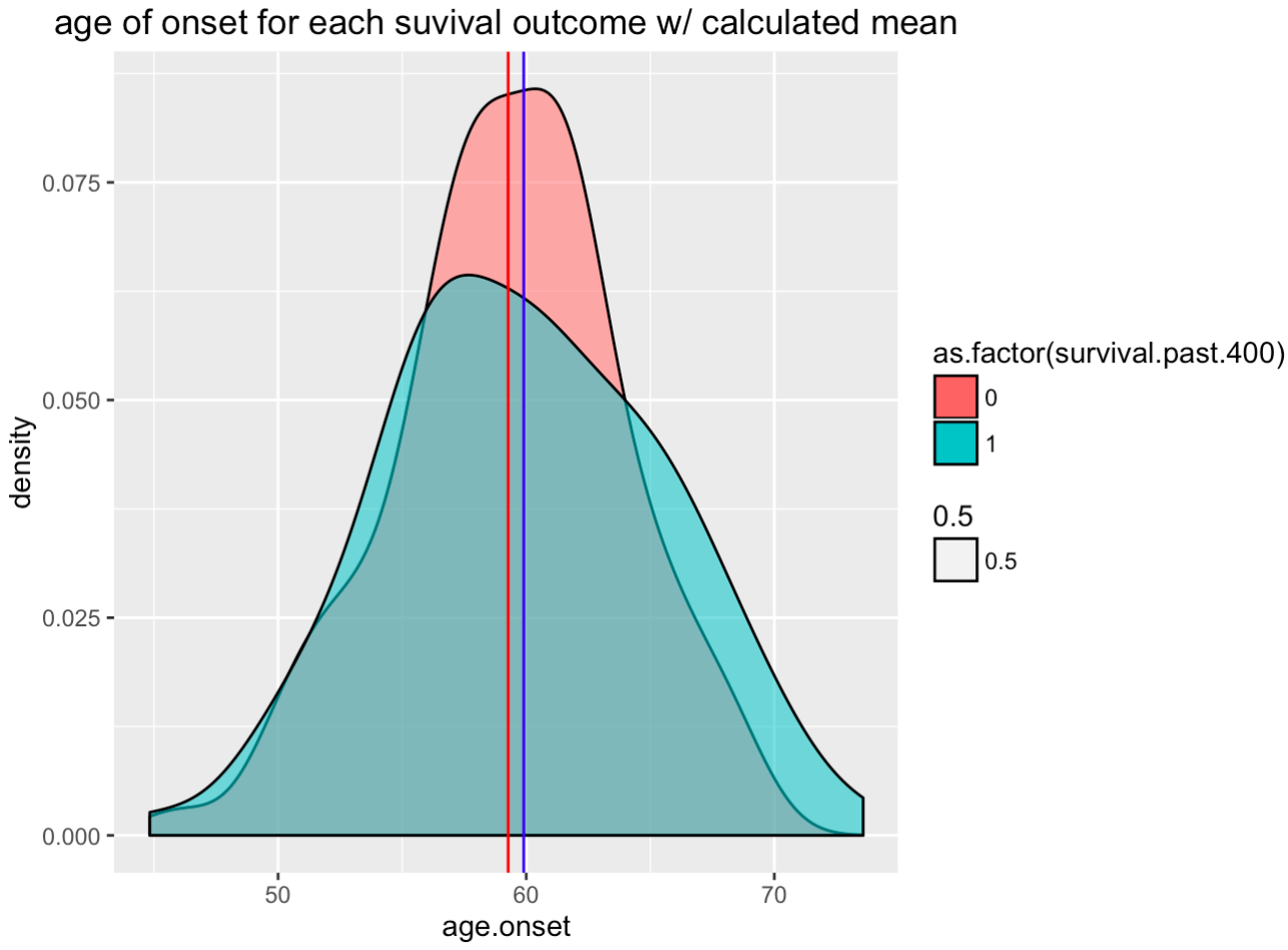

```
#Calculate tail probabilities
mean(permutated.diff_survival$diff.survival.ECOG_0 <= diff.prop.survival.ECOG_0.empir)
mean(permutated.diff_survival$diff.survival.ECOG_1 <= diff.prop.survival.ECOG_1.empir)
mean(permutated.diff_survival$diff.survival.ECOG_2 <= diff.prop.survival.ECOG_2.empir)
```

```
## [1] 0.713
## [1] 0.899
## [1] 0.596
```

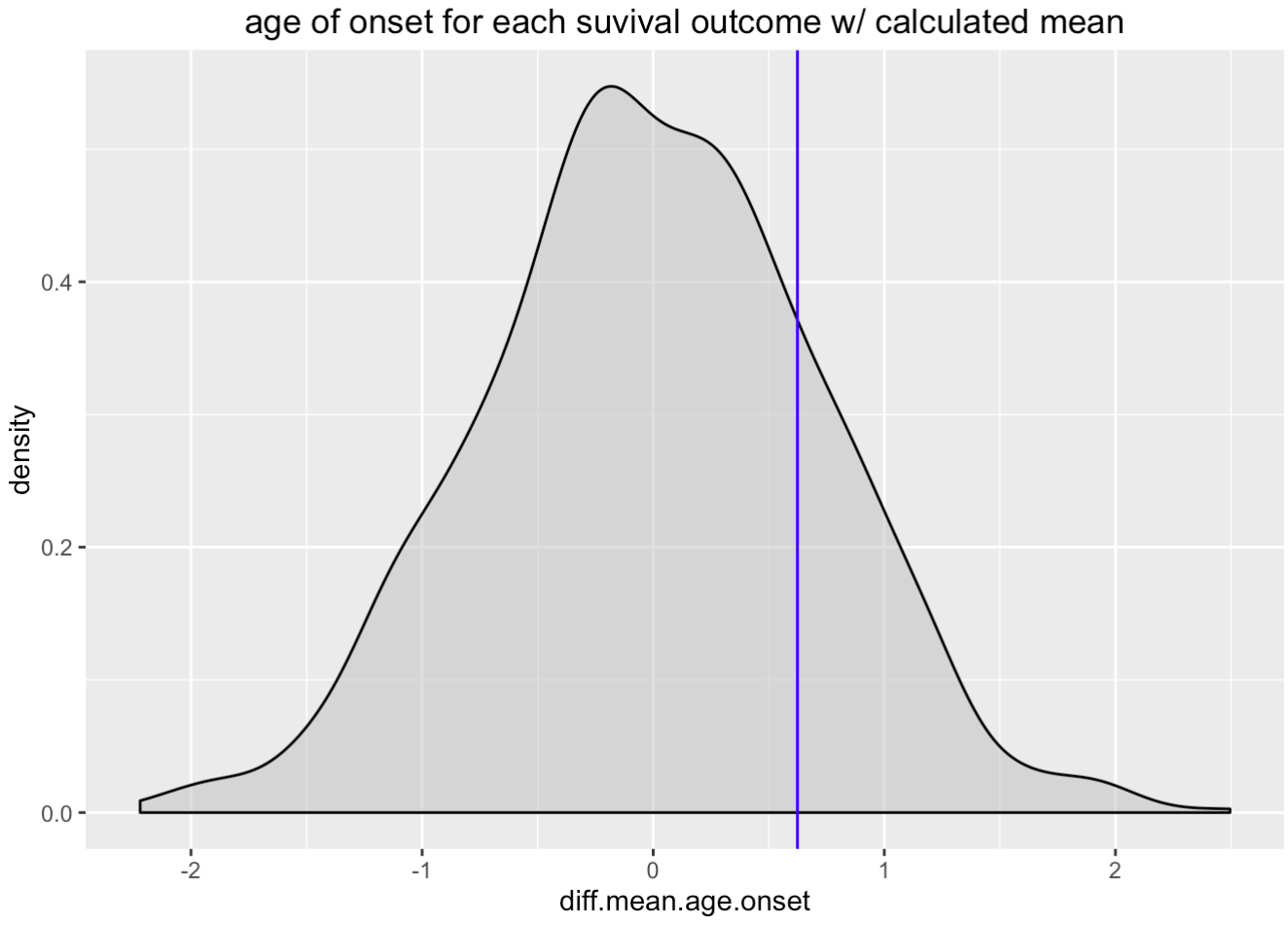
2)

a. Is age of onset associated with outcome? I plot the relationship between age of onset and survival outcome and survival time in a variety of formats. I begin by graphing the difference in mean age of onset for survivors and non-survivors. There appears to be some small difference in mean age of onset for these two groups. I ran a permutation test, permuting the survivor/non-survivor status, and recalculating the difference in mean age of onset for these categories. I compared the distribution of simulated differences to the empirical difference. The empirical difference in age of onset for survivor v. non-survivor is greater than only ~80% of the simulated data. A boxplot of survival status v. age of onset also shows that individuals that survive past 400 days have a slightly higher median age of onset.

```
# Age of onset is calculated as age (yrs) when patient began trial, minus the time (mo) since dia
gnosis
nsclc[,age.onset:= round(age - (durdis/12), 2)]

#Plot age of onset for each suvival outcome
#calculate mean age of onset for each outcome -> are these significantly different?
ggplot() +
  geom_density(data=nsclc, aes(x=age.onset, fill=as.factor(survival.past.400), alpha=0.5))+
  geom_vline(xintercept=mean(nsclc[survival.past.400=='1']$age.onset), color='blue') +
  geom_vline(xintercept=mean(nsclc[survival.past.400=='0']$age.onset), color='red') +
  ggtitle(label = 'age of onset for each suvival outcome w/ calculated mean' )
```



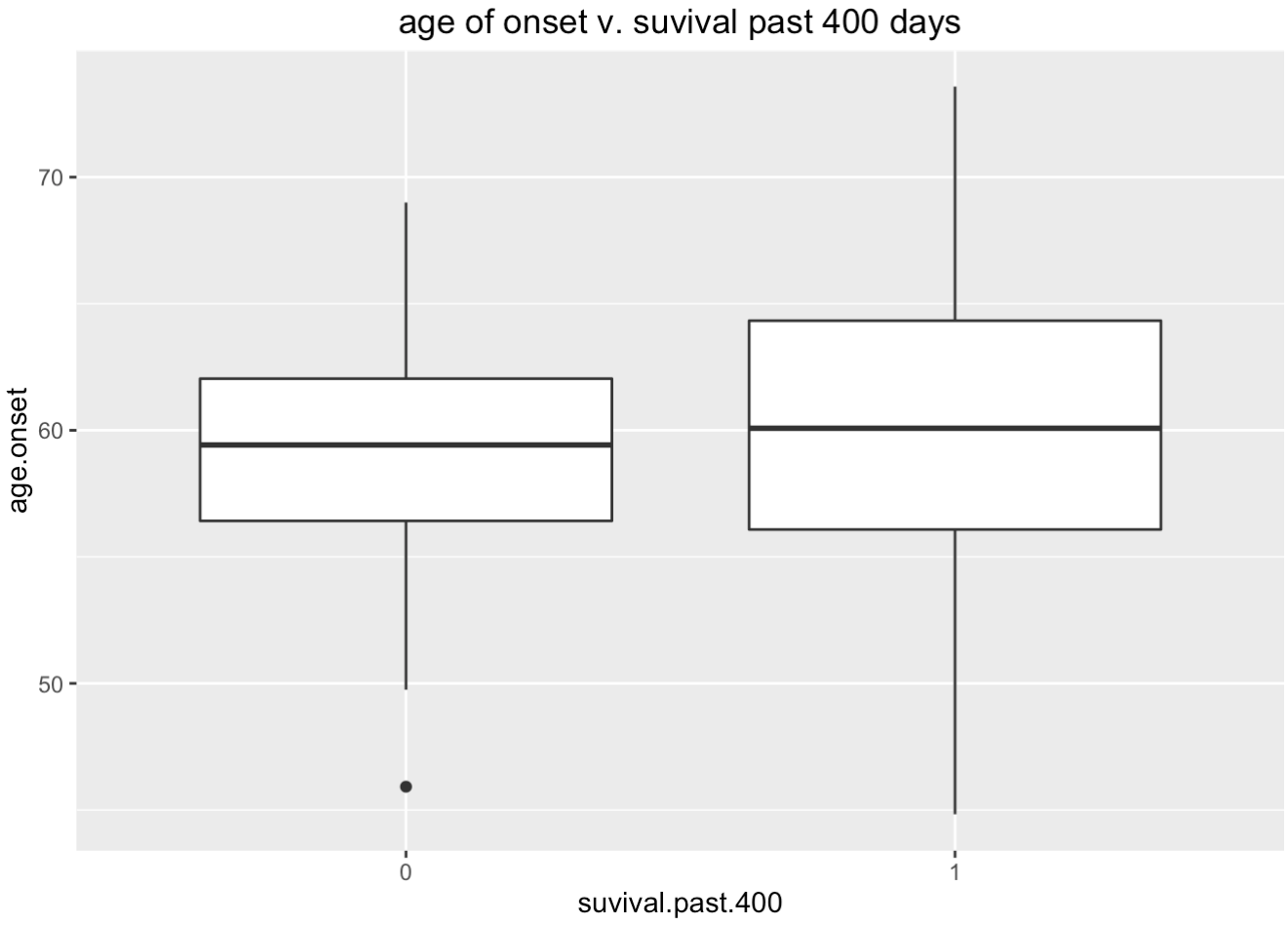
```
#calculate difference in mean age of onset for survivors and non-survivors
#permute status of survival, recalculate this difference
#compare simulated difference in mean age to empirical difference in mean age

diff_mean.age.onset.empir = mean(nsclc[survival.past.400=='1']$age.onset) -
                            mean(nsclc[survival.past.400=='0']$age.onset)

diff.mean.age.fn = function(data){
  permutation <- sample(1:nrow(data), replace=FALSE)
  permutation.data <- data
  permutation.data$survival.past.400 = data$survival.past.400[permutation]

  diff.mean.age.onset = mean(permutation.data[survival.past.400=='1']$age.onset) -
                        mean(permutation.data[survival.past.400=='0']$age.onset)

  return(diff.mean.age.onset)
}

permuted.diff_mean.age.onset = as.data.table(replicate(n = 1000, expr = diff.mean.age.fn(nsclc)))
setnames(permuted.diff_mean.age.onset, c('diff.mean.age.onset'))

ggplot() +
  geom_density(data=permuted.diff_mean.age.onset, aes(x=diff.mean.age.onset), fill='grey',
alpha=0.5)+
  geom_vline(xintercept=diff_mean.age.onset.empir, color='blue') +
  ggtitle(label = 'age of onset for each suvival outcome w/ calculated mean' )
```




```
mean(permuted.diff_mean.age.onset$diff.mean.age.onset <= diff_mean.age.onset.empir)
```

```
## [1] 0.81
```

```
#Plot age of onset v. survival status
#Those that survive past 400 days show a slightly higher median age of onset
ggplot(data=nsclc)+
  geom_boxplot(aes(x=as.factor(survival.past.400), y=age.onset)) +
  xlab(label = 'suvival.past.400') +
  ggtitle('age of onset v. suvival past 400 days')
```



Plotting age of onset v. survival status and fitting a non-linear regression line shows a slight association. Individuals diagnosed at age ~60yrs may have lower likelihood of survival, and those diagnosed closer to age ~70yrs may have slightly higher likelihood of survival.

```
############################
#Plot age of onset v. suvival status
#I see no sign of any relationship between these variables.
#The spearman correlation between these variables is weak.
ggplot(data=nsclc, aes(x=age.onset, y=survival.past.400)) +
  geom_point() +
  geom_smooth() +
  ggtitle('age of onset v. suvival status')
```

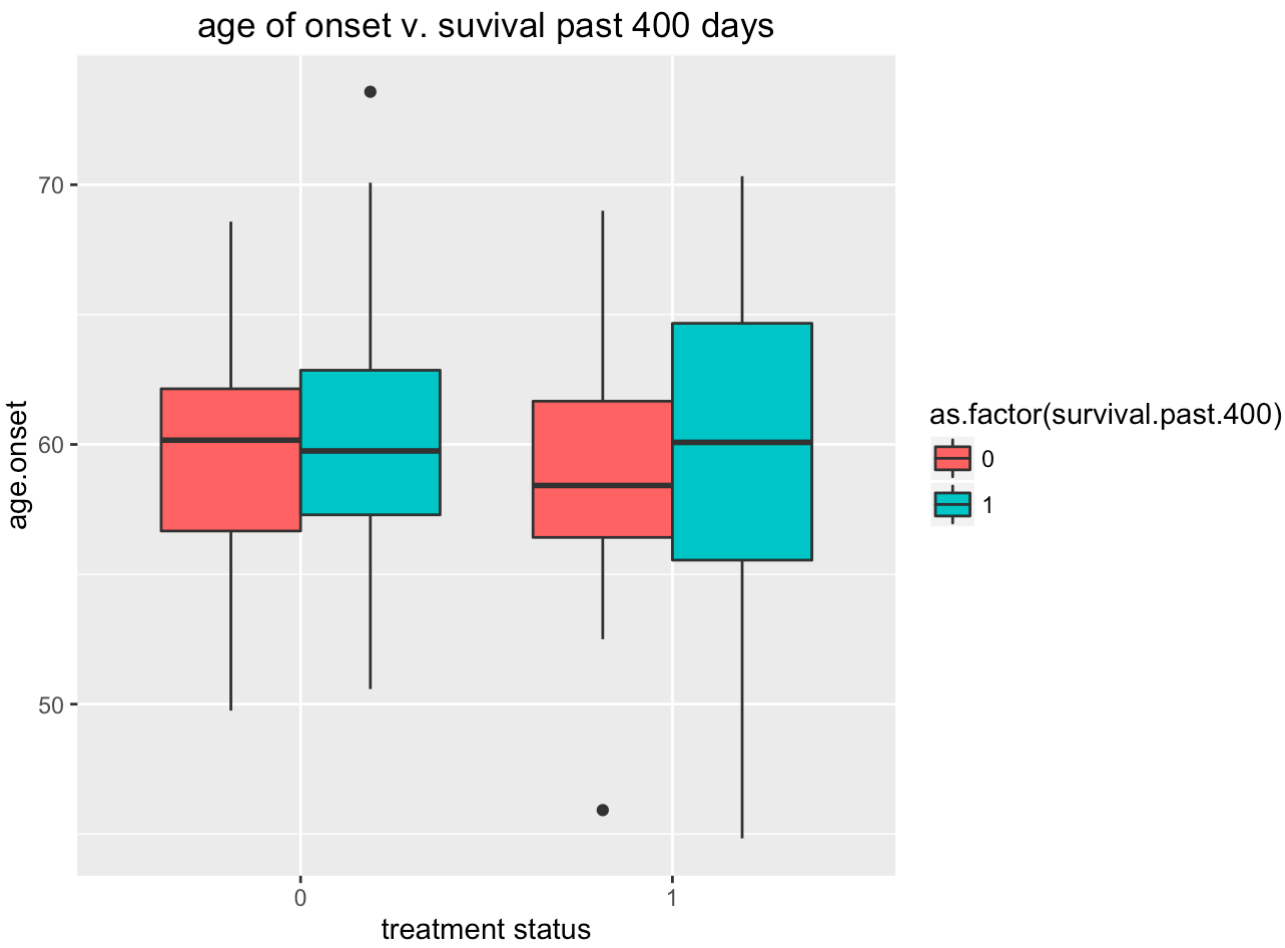

```
cor.test(y = nsclc$survival.past.400, x = nsclc$age.onset, method = 'spearman')
```

```
## Warning in cor.test.default(y = nsclc$survival.past.400, x =
## nsclc$age.onset, : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  nsclc$age.onset and nsclc$survival.past.400
## S = 1060193, p-value = 0.5612
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.04264076
```

b. Does the association btwn age of onset and survival status change depending on treatment status? Can we make any kind of predictions from this association? Based on a boxplot of the age of onset for survivors v. non-survivors, stratified by treatment-status, it appears that those who were on the new treatment and survived had a higher median age of onset than non-survivors. This trend was reversed for the control group, and was much weaker. The spearman correlation between age of onset and survival status is slightly higher for the new treatment group than for the control treatment group. The regression on the data suggests that for the new treatment, the lowest likelihood of survival past 400 days occurs if an individual is diagnosed between ages 55-63 yrs old. At age of onset 55-60 yrs old, individuals given the control treatment are more likely to survive than those given the new treatment.

```
ggplot(data=nsclc)+
  geom_boxplot(aes(x=as.factor(tx), y=age.onset, fill=as.factor(survival.past.400))) +
  xlab(label = 'treatment status') +
  ggtitle('age of onset v. suvival past 400 days')
```




```
#It appears like both the treatment group and the control group have significant outliers for age
 of diagnosis. I've decided to compare them based on difference in median age rather than differe
nce in mean age for this reason.

ggplot(data=nsclc, aes(x=age.onset, y=survival.past.400, color=as.factor(tx))) +
  geom_point() +
  geom_smooth() +
  ggtitle('age of onset v. suvival status')
```


```
cor.test(x = nsclc[tx=='1']$age.onset, y = nsclc[tx=='1']$survival.past.400, method = 'spearman')
```

```
## Warning in cor.test.default(x = nsclc[tx == "1"]$age.onset, y = nsclc[tx
## == : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  nsclc[tx == "1"]$age.onset and nsclc[tx == "1"]$survival.past.400
## S = 150805.7, p-value = 0.7064
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.03852923
```

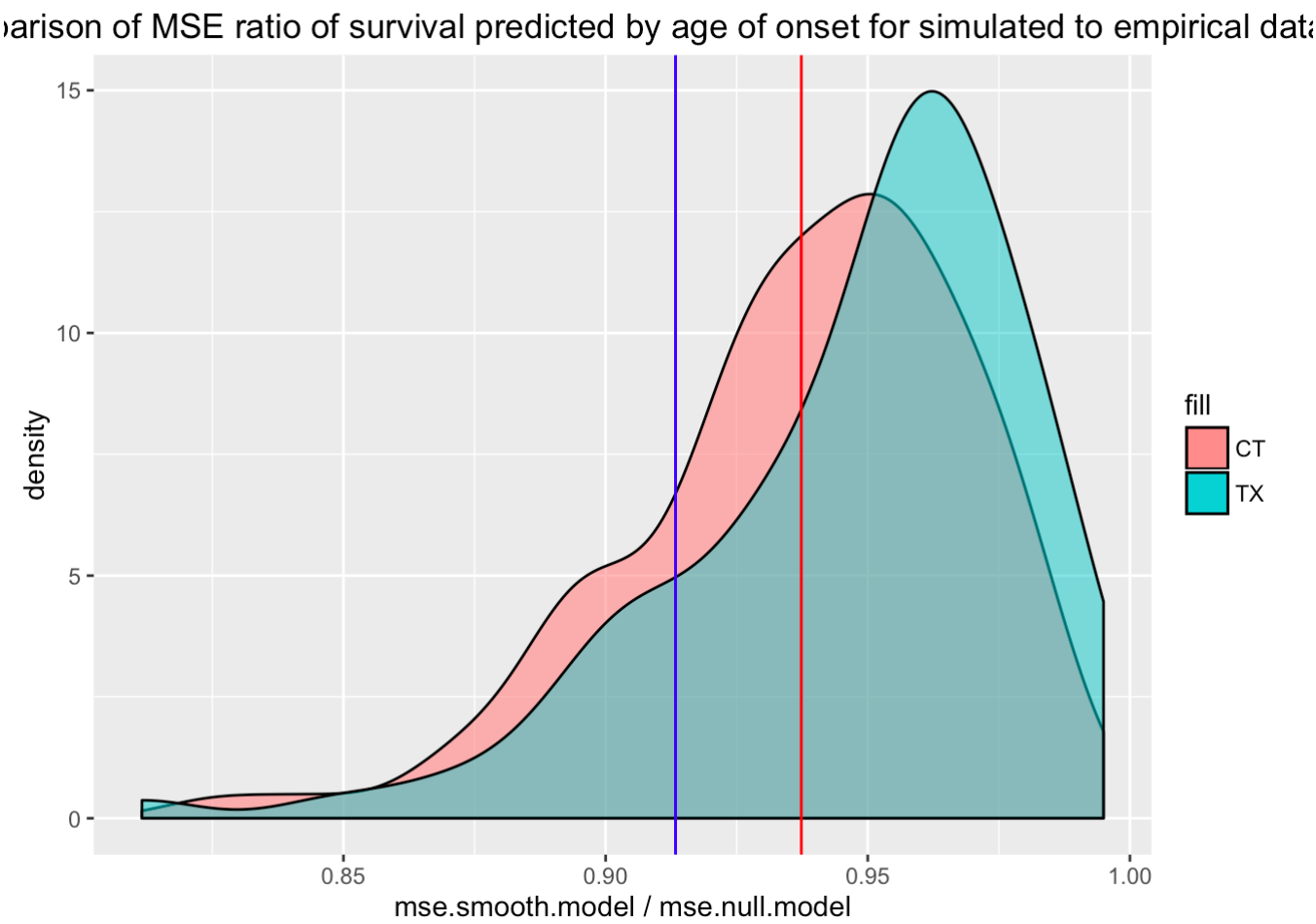
```
cor.test(x = nsclc[tx=='0']$age.onset, y = nsclc[tx=='0']$survival.past.400, method = 'spearman')
```

```
## Warning in cor.test.default(x = nsclc[tx == "0"]$age.onset, y = nsclc[tx
## == : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  nsclc[tx == "0"]$age.onset and nsclc[tx == "0"]$survival.past.400
## S = 117802.7, p-value = 0.7767
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.03031083
```

I ran a permutation test examining the ratio of mean squared error for non-linear models predicting survival status either using age of onset as a variable or not. I stratified by treatment status. Comparing the distribution of simulated to empirical data showed that for neither treatment status did the model including age of onset significantly do better than the null model.

```
fit.survive.TX.null = lm(as.numeric(survival.past.400) ~ 1, data=nsclc[tx=='1'])
fit.survive.TX.smooth = loess(as.numeric(survival.past.400) ~ age.onset, data=nsclc[tx=='1'])
MSE.TX.ratio.empir = mean((fit.survive.TX.smooth$residuals)^2) / mean((fit.survive.TX.null$residu
als)^2)

fit.survive.CT.null = lm(as.numeric(survival.past.400) ~ 1, data=nsclc[tx=='0'])
fit.survive.CT.smooth = loess(as.numeric(survival.past.400) ~ age.onset, data=nsclc[tx=='0'])
MSE.CT.ratio.empir = mean((fit.survive.CT.smooth$residuals)^2) / mean((fit.survive.CT.null$residu
als)^2)

diff.MSE.age.treat.fn = function(data){

  permutation.TX <- sample(1:nrow(filter(data, tx==1)), replace=FALSE)
  permutation.CT <- sample(1:nrow(filter(data, tx==0)), replace=FALSE)

  permutation.data.TX <- filter(data, tx==1)
  permutation.data.CT <- filter(data, tx==0)

  permutation.data.TX$survival.past.400 = permutation.data.TX$survival.past.400[permutation.TX]
  permutation.data.CT$survival.past.400 = permutation.data.CT$survival.past.400[permutation.CT]

  permuted.data = rbind(permutation.data.TX, permutation.data.CT)

  fit.survive.TX.null = lm(as.numeric(survival.past.400) ~ 1, data=permuted.data[tx=='1'])
  fit.survive.TX.smooth = loess(as.numeric(survival.past.400) ~ age.onset, data=permuted.data[tx=
='1'])
  MSE.TX.ratio = mean((fit.survive.TX.smooth$residuals)^2) /
mean((fit.survive.TX.null$residuals)^2)

  fit.survive.CT.null = lm(as.numeric(survival.past.400) ~ 1, data=permuted.data[tx=='0'])
  fit.survive.CT.smooth = loess(as.numeric(survival.past.400) ~ age.onset, data=permuted.data[tx=
='0'])
  MSE.CT.ratio = mean((fit.survive.CT.smooth$residuals)^2) /
mean((fit.survive.CT.null$residuals)^2)

  return(c(MSE.CT.ratio, MSE.TX.ratio))
}

p.MSE.ratio = as.data.table(t(as.data.table(replicate(100, expr = diff.MSE.age.treat.fn(data = ns
clc)))))
setnames(p.MSE.ratio, c('MSE.CT.ratio', 'MSE.TX.ratio'))

ggplot() +
  geom_density(data=p.MSE.ratio, aes(x=MSE.CT.ratio, fill='CT'), alpha=0.5) +
  geom_density(data=p.MSE.ratio, aes(x=MSE.TX.ratio, fill='TX'), alpha=0.5) +
  geom_vline(xintercept = MSE.CT.ratio.empir, color='red') +
  geom_vline(xintercept = MSE.TX.ratio.empir, color='blue') +
  xlab('mse.smooth.model / mse.null.model') +
  ggtitle(label = 'Comparison of MSE ratio of survival predicted by age of onset for simulated to
 empirical data')
```




```
mean(p.MSE.ratio$MSE.CT.ratio >= MSE.CT.ratio.empir)
```

```
## [1] 0.55
```

```
mean(p.MSE.ratio$MSE.TX.ratio >= MSE.TX.ratio.empir)
```

```
## [1] 0.83
```