Project 2

Hayley Zorkic

11/16/2020

0. Background

From the original authors of the dataset: "These are data from one of the first successful trials of adjuvant chemotherapy for colon cancer. Levamisole is a low-toxicity compound previously used to treat worm infestations in animals; 5-FU is a moderately toxic (as these things go) chemotherapy agent. There are two records per person, one for recurrence and one for death." This porject will explore the follwoing variables AT THE TIME OF DEATH from the colon dataset: Categorical (2-5 groups) - rx Treatment : Obs(ervation), Lev(amisole), Lev(amisole)+5-FU - differ Differentiation of tumour : 1=well, 2=moderate, 3=poor - extent Extent of local spread : 1=submucosa, 2=muscle, 3=serosa, 4=contiguous structures - sex (binary) : 1=male - obstruct (binary) obstruction of colon by tumour : 1=yes Numeric (more than 10 values) - age (numeric) : in years - time (numeric) days until death - nodes (numeric) number of lymph nodes with detectable cancer After omitting NA;s, 888 unique partient observations were left.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.1
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(survival)
data <- colon %>% select(id,rx,sex,age,obstruct,nodes,differ, extent, time, etype) %>% filter(etype != 
rownames(data) <- NULL
head(data)
```

```
##   id      rx sex age obstruct nodes differ extent time
## 1  1 Lev+5FU   1  43        0     5      2      3 1521
## 2  2 Lev+5FU   1  63        0     1      2      3 3087
## 3  3     Obs   0  71        0     7      2      2  963
## 4  4 Lev+5FU   0  66        1     6      2      3  293
## 5  5     Obs   1  69        0    22      2      3  659
## 6  6 Lev+5FU   0  57        0     9      2      3 1767
```

1. MANOVA

A one way MANOVA was conducted to determine the effect of 3 dependent variables (age, nodes, and time) on two levels of colon obstruction.

```r
dvs <- c("age","nodes","time")
man<-manova(cbind(age,nodes,time)~obstruct, data)

summary(man) #MANOVA
```

```
##            Df   Pillai approx F num Df den Df    Pr(>F)
## obstruct    1 0.020396   6.1351      3    884 0.0003949 ***
## Residuals 886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary.aov(man) #get univariate ANOVAs from MANOVA object
```

```
##  Response age :
##              Df Sum Sq Mean Sq F value   Pr(>F)
## obstruct      1   1128 1128.03  8.0091 0.004759 **
## Residuals   886 124786  140.84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response nodes :
##              Df  Sum Sq Mean Sq F value Pr(>F)
## obstruct      1     9.6   9.608  0.7664 0.3816
## Residuals   886 11106.7  12.536
##
##  Response time :
##              Df    Sum Sq Mean Sq F value  Pr(>F)
## obstruct      1   4407593 4407593  5.8107 0.01613 *
## Residuals   886 672059245  758532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
pairwise.t.test(data$time,data$obstruct, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data$time and data$obstruct
##
##   0
## 1 0.016
##
## P value adjustment method: none
```

```r
pairwise.t.test(data$age,data$obstruct, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data$age and data$obstruct
##
##   0
## 1 0.0048
##
## P value adjustment method: none
```

Significant differences were found among the two levels of obstruction for at least one of the dependent varibales, Pillai= 0.020396, pseudo F()= 6.1351, p=0.0003949.

Univariate ANOVAs for each dependent variable were conducted as follow-up tests to the MANOVA, using the Bonferroni method for controlling for Type 1 error rates for multiple comparisons. The univariate ANOVAs for age and time were also significant F()= 8.0091, p=0.004759, and F()= 5.8107,p=0.01613.

Post hoc analysis was performed conducting pairwise comparisons to determine if obstruction levels differed across time and age. There is a significant difference between obstruction levels in age, but not time after adjusting for multiple comparisons (bonferroni a = .05/6 = 0.008333333).

The MANOVA addumes the following: 1. Random samples, independent observations, 2. Multivariate normality of DVs (or each group), 3. Homogeneity of within-group covariance matrices ANOVA assumes equal variance for DV within each group MANOVA assumes it for each DV and equal covariance between any two DVs, 4. Linear relationships among DVs, 5. No extreme univariate or multivariate outliers, 6. No multicollinearity (i.e., DVs should not be too correlated).

If anything, I believe the time data would violate the noramlity data because there may be an exponential increase in days until death.

2. Randomization test

If we scramble our data up, we break any systematic association between group and response. On average the groups' mean responses will be the same. If we do this repeatedly and calculate an F statistic each time, we get the sampling distribution of F under the Null hypothesis that all groups have the same mean. We then compare our actual statistic to the null distribution to see if it is a plausible value in the null distribution, else all groups do not have the same mean.

```
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.5-6
```

```
#compute distances/dissimilarities
dists <- data%>%select(sex,age,nodes,differ,extent,time)%>%dist
#perform PERMANOVA on distances/dissimilarities
adonis(dists~obstruct,data)
```

```
##
## Call:
## adonis(formula = dists ~ obstruct, data = data)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## obstruct    1   4408732 4408732   5.811 0.00652  0.017 *
## Residuals 886 672195793  758686         0.99348
## Total     887 676604525                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

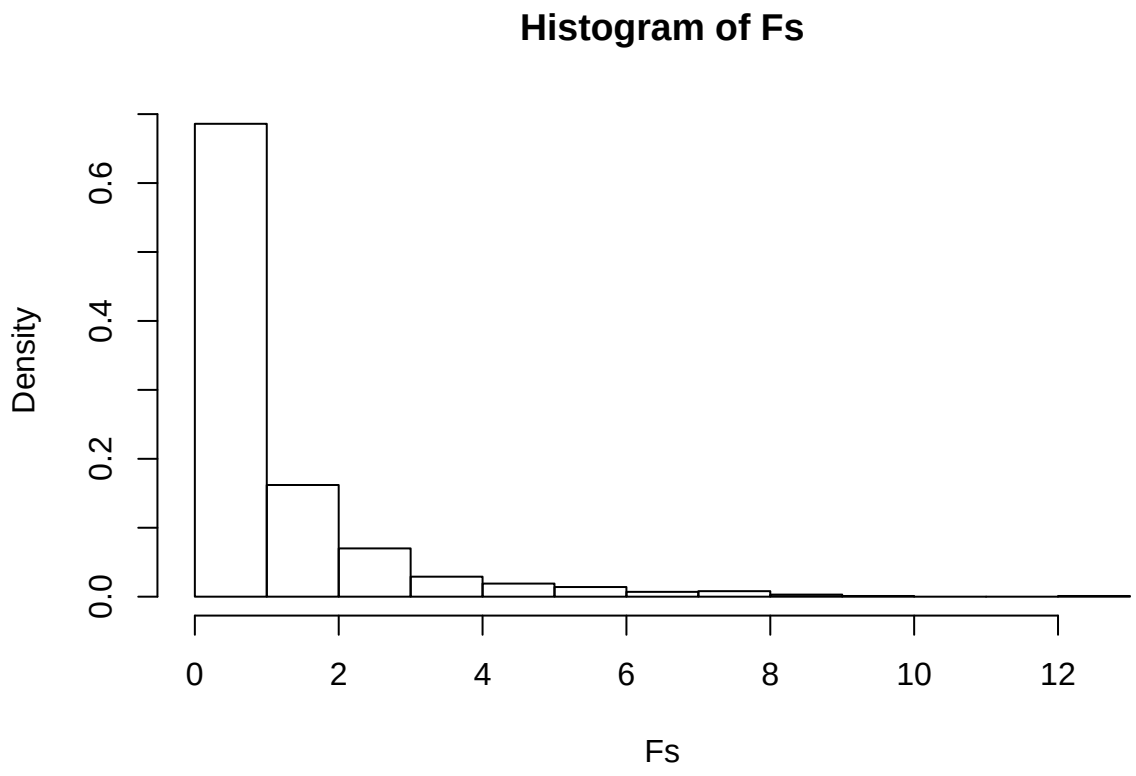
```
SST<- sum(dists^2)/888
SSW <- data%>%group_by(obstruct)%>%select(obstruct,sex,age,nodes,differ,extent,time) %>%
    do(d=dist(.[-1],"euclidean"))%>%ungroup()%>%
    summarize(sum(d[[1]]^2)/717 + sum(d[[2]]^2)/171)%>%pull
F_obs<-((SST-SSW)/1)/(SSW/886) #observed F statistic
F_obs
```

```
## [1] 5.81101
```

```
Fs<-replicate(1000,{
  new <- data%>%mutate(obstruct=sample(obstruct)) #randomly permute response variable (rx)
  SSW <- new%>%group_by(obstruct)%>%select(obstruct,sex,age,nodes,differ,extent,time) %>%
    do(d=dist(.[-1],"euclidean"))%>%ungroup()%>%
    summarize(sum(d[[1]]^2)/717 + sum(d[[2]]^2)/171)%>%pull

  ((SST-SSW)/1)/(SSW/886) #calculate new F on randomized data
})

hist(Fs,prob = T)
```




```
##Let's look at this distribution! What is the probability of getting a statistic at least as extreme as
mean(Fs>F_obs)
```

```
## [1] 0.021
```

A PERMANOVA Randomizaion Test was performed on the data- using distance matricies for partitioning distance matrices among sources of variation and fitting linear models (e.g., factors, polynomial regression) to

distance matrices; uses a permutation test with pseudo-F ratios. The p value generated by the PERMANOVA is p=0.022<0.05 so there is no significant difference in F statists generated under the null hypothesis that the obstruction groups differ by sex,age,nodes,differ,extent,and/or time, out F statistic being 5.811.

We performed a by hand PERMANOVA calulation to compare to our actual F statistic generated in the PERMANOVA from the vegan package and got a p value of p=0.014<0.05 so we reject the null hypothesis and can conclude the obstruction groups differ across the variables.

3.Linear Regression Model

A linear regression model with interaction was constructed to predict obstruction from prescription and number of nodes. All numeric variables were centered.

```
data$time_c <- data$time - mean(data$time)
data$age_c <- data$age - mean(data$age)
data$nodes_c <- data$nodes - mean(data$nodes)
fit <- lm(time_c ~ rx*nodes_c, data = data)
summary(fit)
```

```
##
## Call:
## lm(formula = time_c ~ rx * nodes_c, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1849.0  -748.9   241.9   658.4  1727.0
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -60.04      47.49  -1.264   0.2065
## rxLev                13.67      67.76   0.202   0.8402
## rxLev+5FU           175.64      68.14   2.578   0.0101 *
## nodes_c             -88.72      12.65  -7.012 4.67e-12 ***
## rxLev:nodes_c        19.56      18.46   1.060   0.2896
## rxLev+5FU:nodes_c    26.44      19.71   1.341   0.1802
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 828.6 on 882 degrees of freedom
## Multiple R-squared:  0.1048, Adjusted R-squared:  0.09973
## F-statistic: 20.65 on 5 and 882 DF,  p-value: < 2.2e-16
```

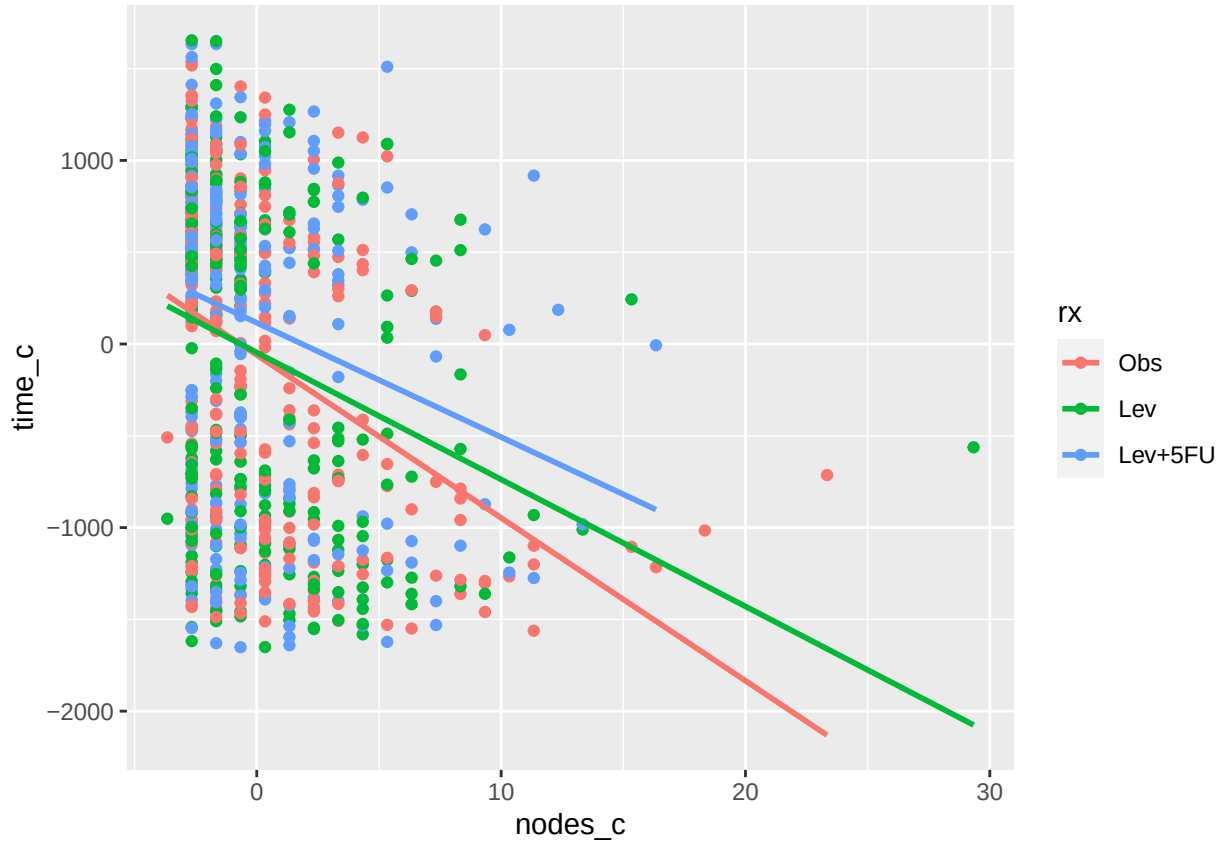
The following interpretations from the linear regression can be made, disregarding significance.

- `Intercept`: -60.04 is the predicted days from the mean time until death when rx=Obs and there is an average number of nodes.
- `rxLev`: Controlling for nodes, the predicted days from the mean time until death increase by 13.67 days for Lev prescription users.
- `rxLev+5U`: Controlling for nodes, the predicted days from the mean time until death increase by 175.64 days for Lev+5U prescription users.
- `nodes_c`: Controlling for rx, for every 1 unit increase in nodes from the average,the predicted days from the mean time until death decrease by 88.72 days.
- `rxLev:nodes_c`: The slope for nodes_c on time_c is 19.56 higher for rxLev users.

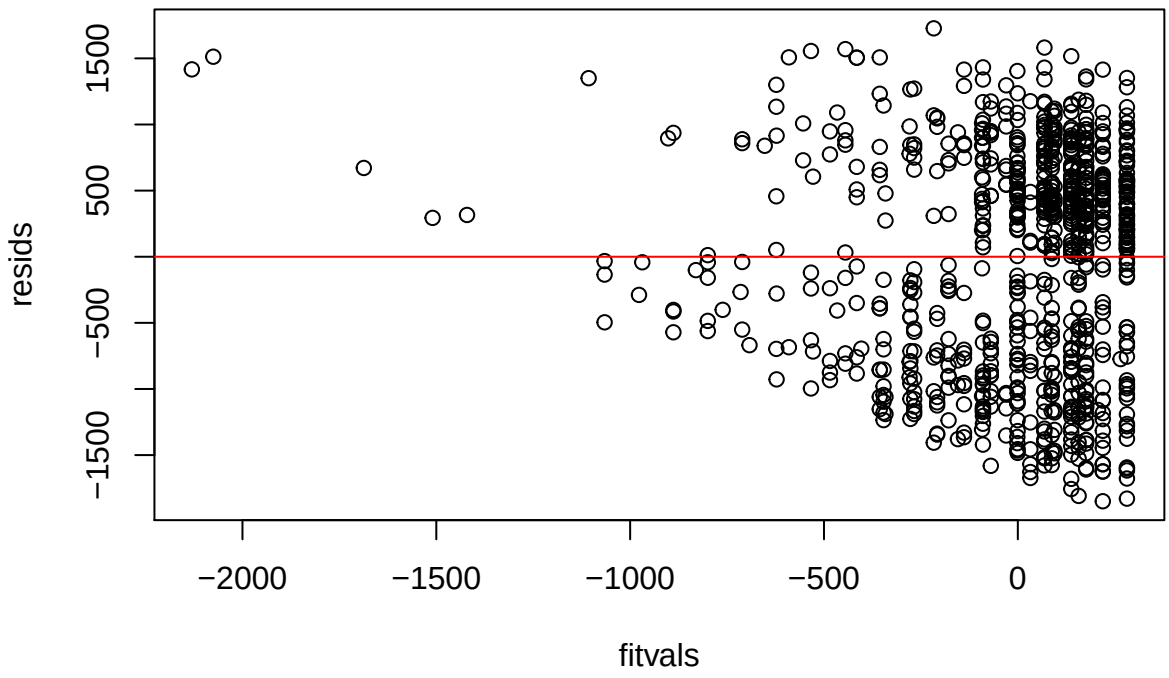
-rxLev+5U:nodes_c: The slope for nodes_ on time_c is 26.44 higher for rxLEv+5U users.

```
data %>% ggplot(aes(nodes_c,time_c,color=rx))+geom_point()+geom_smooth(method="lm", se=F)
```

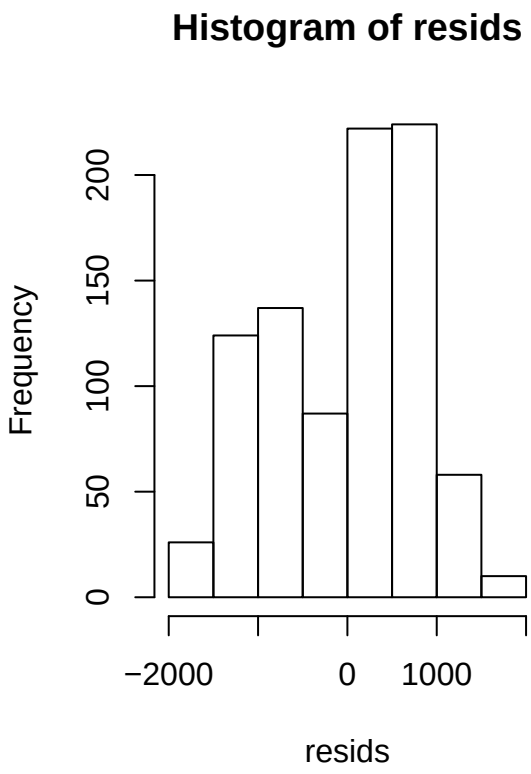
```
## `geom_smooth()` using formula 'y ~ x'
```

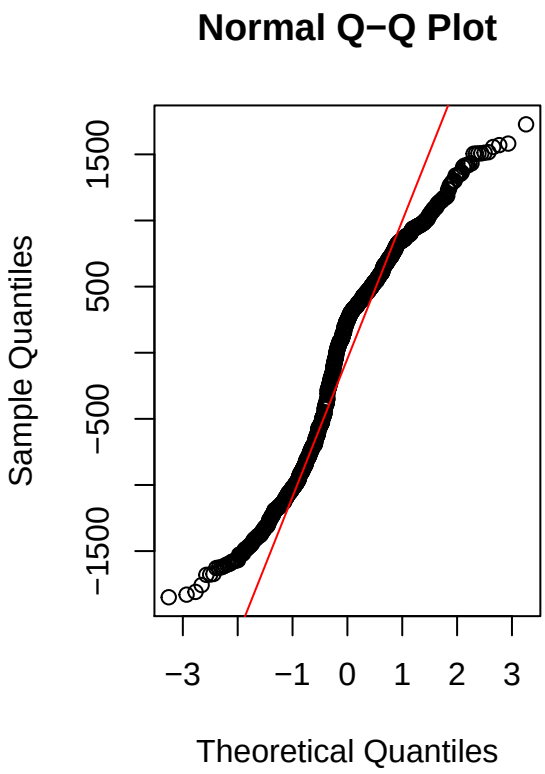



```
resids<-fit$residuals
fitvals<-fit$fitted.values
plot(fitvals,resids); abline(h=0, col='red')
```

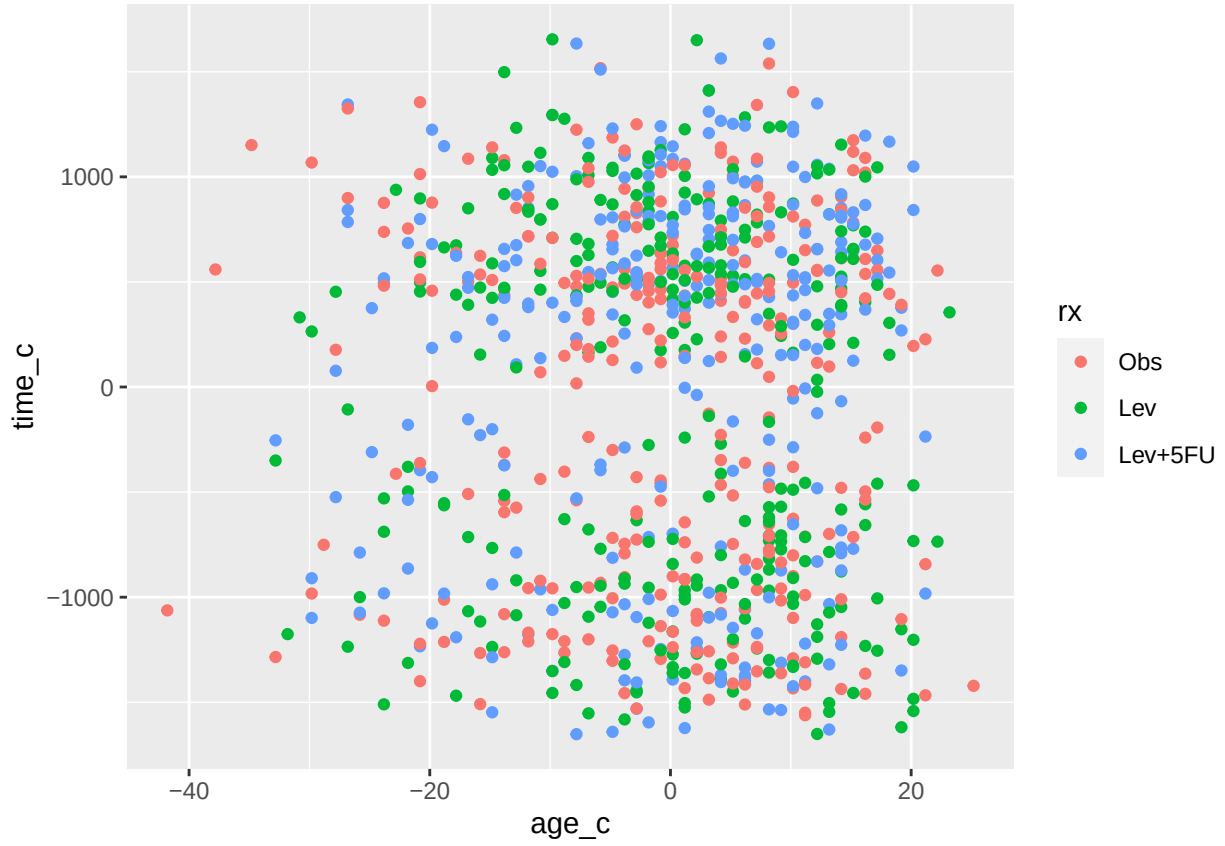



```
par(mfrow=c(1,2)); hist(resids); qqnorm(resids); qqline(resids, col='red')
```






```
ggplot(data,aes(age_c,time_c,color=rx))+geom_point()
```




```
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 11.096, df = 5, p-value = 0.04951
```

```
shapiro.test(resids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.9502, p-value < 2.2e-16
```

According to the plots, the assumptions of linearity appears to be met. Homoskedasticity had to be confirmed

with the the Breusch-Pagan test- the null hypothesis was rejected (p=0.04951<0.05) which showed that homoskedasticity is violated. Normailty was formally tested with the shapiro-wilk test which showed that normailty was violated (p<0.0000001).

```
coeftest(fit, vcov = vcovHC(fit)) #regression after adjusting standard errors for violation
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)         -60.040     45.421 -1.3219  0.186552
## rxLev                13.666     67.846  0.2014  0.840411
## rxLev+5FU           175.643     67.513  2.6016  0.009434 **
## nodes_c             -88.724     12.770 -6.9476 7.213e-12 ***
## rxLev:nodes_c        19.565     23.341  0.8382  0.402132
## rxLev+5FU:nodes_c    26.436     20.642  1.2807  0.200640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After performing a robust SE calculation with the original linear regression, the only two significant effects on the time_c are rxLev+5FU users and an increase in nodes_c- congruent with the original model.

The proportion of variation in the response variable explained by the model is 10.48%- due to chance, our numbers have some association with the outcome.

4. Bootstrapped SEs of Linear Regression Model

A Bootstrapped SE was performed on our linear regression model for predicting obstruction from prescription and number of nodes. All numeric variables were centered.

```
boot_dat<- sample_frac(data, replace=T)
samp_distn<-replicate(5000, {
boot_dat <- sample_frac(data, replace=T) #take bootstrap sample of rows
fit <- lm(time_c~rx*age_c, data=boot_dat) #fit model on bootstrap sample
coef(fit) #save coefs
})
samp_distn %>% t %>% as.data.frame %>% summarize_all(sd)
```

```
##   (Intercept)    rxLev rxLev+5FU    age_c rxLev:age_c rxLev+5FU:age_c
## 1    48.86575 72.24617  70.22366 4.310181    6.103656        5.752585
```

```
coeftest(fit)[,1:2]
```

```
##                    Estimate Std. Error
## (Intercept)       -60.04046   47.49429
## rxLev              13.66610   67.76024
## rxLev+5FU         175.64337   68.14139
## nodes_c           -88.72427   12.65309
## rxLev:nodes_c      19.56495   18.46231
## rxLev+5FU:nodes_c  26.43646   19.71237
```

```
coeftest(fit, vcov=vcovHC(fit))[,1:2]
```

```
##                    Estimate Std. Error
## (Intercept)       -60.04046   45.42066
## rxLev              13.66610   67.84649
## rxLev+5FU         175.64337   67.51271
```

```
## nodes_c           -88.72427   12.77041
## rxLev:nodes_c      19.56495   23.34097
## rxLev+5FU:nodes_c  26.43646   20.64238
```

The original SEs and the robust SEs differ significantly only at rxLev:nodes_c with robust SE having a creater Std. error than original SEs. Bootstrapped SEs were greater than Original SEs and robust SEs in intercept, rxLev, rxLev+5FU but more than half as large than Robust and Original SEs in age_c, rxLev:age_c, rxLev+5FU:age_c.

5. Logistic Regression

A logistic regression was performed without interaction to predict the binary variable obstructin from rx, age_c, and time_c. The numeric variables were centerd.

```
library(lmtest)

fit2<-glm(obstruct~rx+age_c+time_c, data=data, family="binomial")
coeftest(fit2)
```

```
##
## z test of coefficients:
##
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -1.4175e+00  1.4502e-01 -9.7749 < 2.2e-16 ***
## rxLev       -1.8633e-02  2.0592e-01 -0.0905  0.927899
## rxLev+5FU   -1.2481e-01  2.1230e-01 -0.5879  0.556603
## age_c       -1.9995e-02  6.9729e-03 -2.8676  0.004136 **
## time_c      -2.3501e-04  9.7769e-05 -2.4037  0.016229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Can tell three predictors significantly increase the probability of obstruction (significant positive
#EXPONENTIATE COEFFICIENTS TO INTERPRET:
#odds scale coefs (multiplicative): these are the ones you interpret!
coef(fit2)%>%exp%>%round(5)%>%data.frame
```

```
##                    .
## (Intercept) 0.24231
## rxLev       0.98154
## rxLev+5FU   0.88266
## age_c       0.98020
## time_c      0.99977
```

The following interpretations from the logistic regression can be made, disregarding significance.

- `Intercept`: Predicted odds of obstruction when rx=Obs, average age, and average days until death is 0.24231.
- `rxLev`: Controlling for age and time, odds of obstruction for Lev prescription users is 0.98154 times that the Obs Prescription users.
- `rxLev+5FU`: Controlling for age and time, odds of obstruction for Lev+5FU prescription users is 0.88266 times that the Obs Prescription users.
- `age_c`: Controlling for prescription and time, every year increase in age from the mean multiples the odds of obstruction multiplies odds by 0.98020.

- `time_c`: Controlling for prescription and age, every day increase in days until death from the mean multiples the odds of obstruction multiplies odds by 0.99977.

```
class_diag <- function(probs,truth){
  #CONFUSION MATRIX: CALCULATE ACCURACY, TPR, TNR, PPV
  tab<-table(factor(probs>.5,levels=c("FALSE","TRUE")),truth)
  acc=sum(diag(tab))/sum(tab)
  sens=tab[2,2]/colSums(tab)[2]
  spec=tab[1,1]/colSums(tab)[1]
  ppv=tab[2,2]/rowSums(tab)[2]
  f1=2*(sens*ppv)/(sens+ppv)

  if(is.numeric(truth)==FALSE & is.logical(truth)==FALSE) truth<-as.numeric(truth)-1

  #CALCULATE EXACT AUC
  ord<-order(probs, decreasing=TRUE)
  probs <- probs[ord]; truth <- truth[ord]

  TPR=cumsum(truth)/max(1,sum(truth))
  FPR=cumsum(!truth)/max(1,sum(!truth))

  dup<-c(probs[-1]>=probs[-length(probs)], FALSE)
  TPR<-c(0,TPR[!dup],1); FPR<-c(0,FPR[!dup],1)
  n <- length(TPR)
  auc<- sum( ((TPR[-1]+TPR[-n])/2) * (FPR[-1]-FPR[-n]) )

  data.frame(acc,sens,spec,ppv,f1,auc)
}

prob<-predict(fit2,type="response") #get predictions for each patient. prob>.5
class_diag(prob,data$obstruct)
```

```
##         acc sens spec ppv  f1      auc
## 1 0.8074324    0    1 NaN NaN 0.580754
```

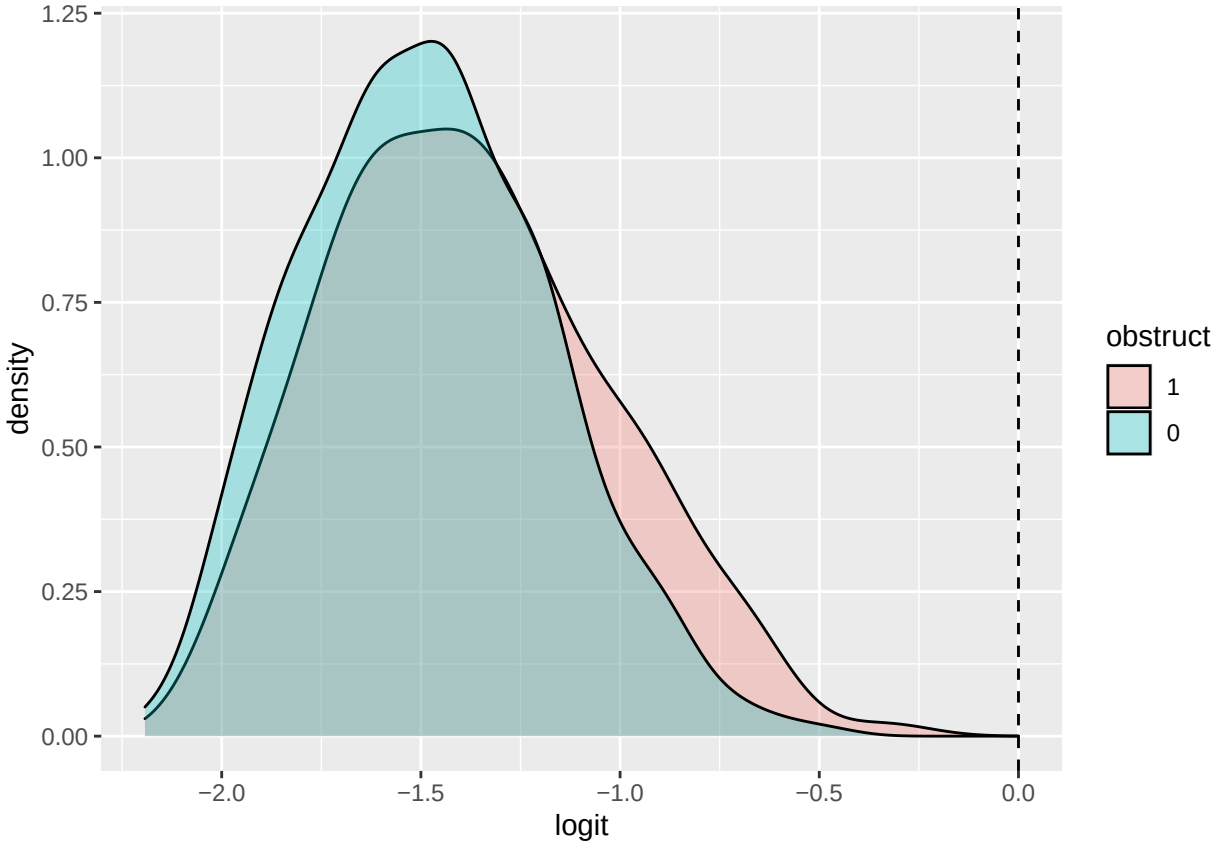
From the confusion matrix, we can see the accurcay of the model is 0.8074324, the sensitivty is 0, the specificity is 1, the prescision is NaN because the sensitivity is 0, and the auc is 0.580754... not great.

Below is the density plot of the log odds colored by obstruction binary outcome.

```
data$logit<-predict(fit2) #get predicted log-odds (logits)

#plot logit scores for each truth category

data %>% mutate(obstruct=factor(obstruct,levels=c("1","0"))) %>% #changing order so color red=malig
ggplot(aes(logit, fill=obstruct))+geom_density(alpha=.3)+
  geom_vline(xintercept=0,lty=2)
```



Below is the ROC curve and AUC calculations.

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

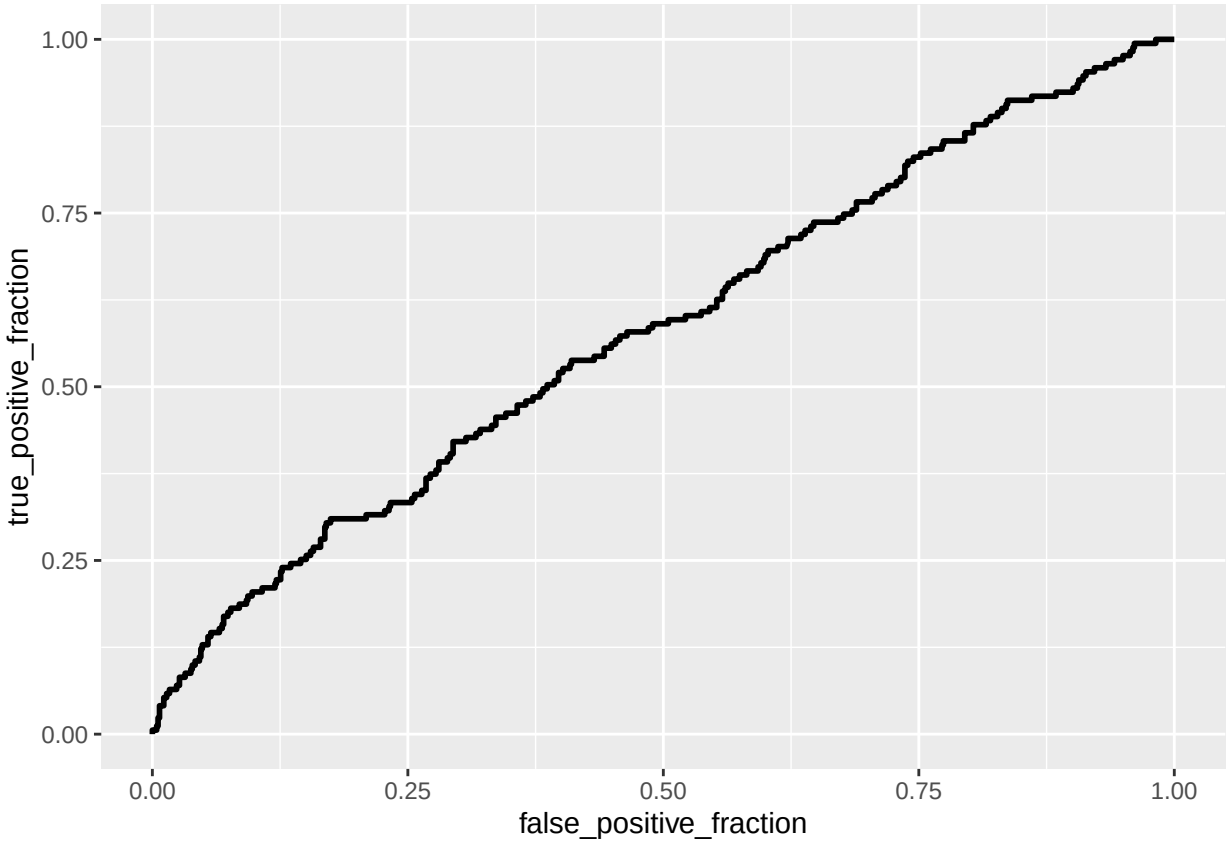
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(plotROC)#Compare with this AUC calculator!
```

```
##
## Attaching package: 'plotROC'
```

```
## The following object is masked from 'package:pROC':
##
##     ggroc
```

```
ROCplot<-ggplot(data)+geom_roc(aes(d=obstruct,m=prob), n.cuts=0)
ROCplot
```




```
calc_auc(ROCplot)
```

```
##   PANEL group      AUC
## 1     1    -1 0.580754
```

The AUC of 0.580754 is Bad! It is hard to determind obstruction from rx, age_c, and time_c.

6. Logistic Regression with ALL variables

A logistic regression was performed without interaction to predict the binary variable obstructin from all of the variables. The numeric variables were centerd.

```
library(lmtest)
fit3<-glm(obstruct~rx+sex+age_c+differ+extent+time_c+nodes_c, data=data, family="binomial")
coeftest(fit3)
```

```
##
## z test of coefficients:
##
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -2.26511508  0.68773683 -3.2936 0.0009892 ***
## rxLev       -0.01730803  0.20736421 -0.0835 0.9334804
## rxLev+5FU   -0.13828409  0.21372337 -0.6470 0.5176166
## sex         -0.17135907  0.17341041 -0.9882 0.3230691
## age_c       -0.02171535  0.00708102 -3.0667 0.0021644 **
## differ      -0.06512005  0.17143740 -0.3798 0.7040587
## extent       0.36516717  0.20025407  1.8235 0.0682248 .
```

```
## time_c      -0.00027555  0.00010514 -2.6206 0.0087768 **
## nodes_c     -0.05747001  0.02895820 -1.9846 0.0471907 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Can tell three predictors significantly increase the probability of obstruction (significant positive
#EXPONENTIATE COEFFICIENTS TO INTERPRET:
#odds scale coefs (multiplicative): these are the ones you interpret!
coef(fit3)%>%exp%>%round(5)%>%data.frame
```

```
##                    .
## (Intercept) 0.10382
## rxLev       0.98284
## rxLev+5FU   0.87085
## sex         0.84252
## age_c       0.97852
## differ      0.93695
## extent      1.44075
## time_c      0.99972
## nodes_c     0.94415
```

```
prob<-predict(fit3,type="response") #get predictions for each patient. prob>.5
class_diag(prob,data$obstruct)
```

```
##         acc sens spec ppv  f1       auc
## 1 0.8074324    0    1 NaN NaN 0.6179011
```

From the confusion matrix, the accurcay of the model is 0.8074324, the sensitivty is 0, the specificity is 1, the prescision is NaN because the sensitivity is 0, and the auc is 0.6179011... Poor.

A 10-fold CV was run with the model.

```
set.seed(1234)
k=10 #choose number of folds

data10<-data[sample(nrow(data)),] #randomly order rows
folds<-cut(seq(1:nrow(data10)),breaks=k,labels=F) #create folds

diags<-NULL
for(i in 1:k){
  ## Create training and test sets
  train<-data10[folds!=i,]
  test<-data10[folds==i,]

  truth<-test$obstruct ## Truth labels for fold i

  ## Train model on training set (all but fold i)
  fit<-glm(obstruct~rx+sex+age_c+differ+extent+time_c+nodes_c,data=train,family="binomial")

  ## Test model on test set (fold i)
  probs<-predict(fit,newdata = test,type="response")

  ## Get diagnostics for fold i
  diags<-rbind(diags,class_diag(probs,truth))
}

summarize_all(diags,mean) #average diagnostics across all k folds
```

```
##         acc sens spec ppv  f1       auc
## 1 0.8074183    0    1 NaN NaN 0.5654982
```

The accurcay of the model is 0.8074183, the sensitivty is 0, the specificity is 1, the prescision is NaN because the sensitivity is 0, and the auc is 0.5654982... Bad.

Next, a LASSO was run on the logistic regression.

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

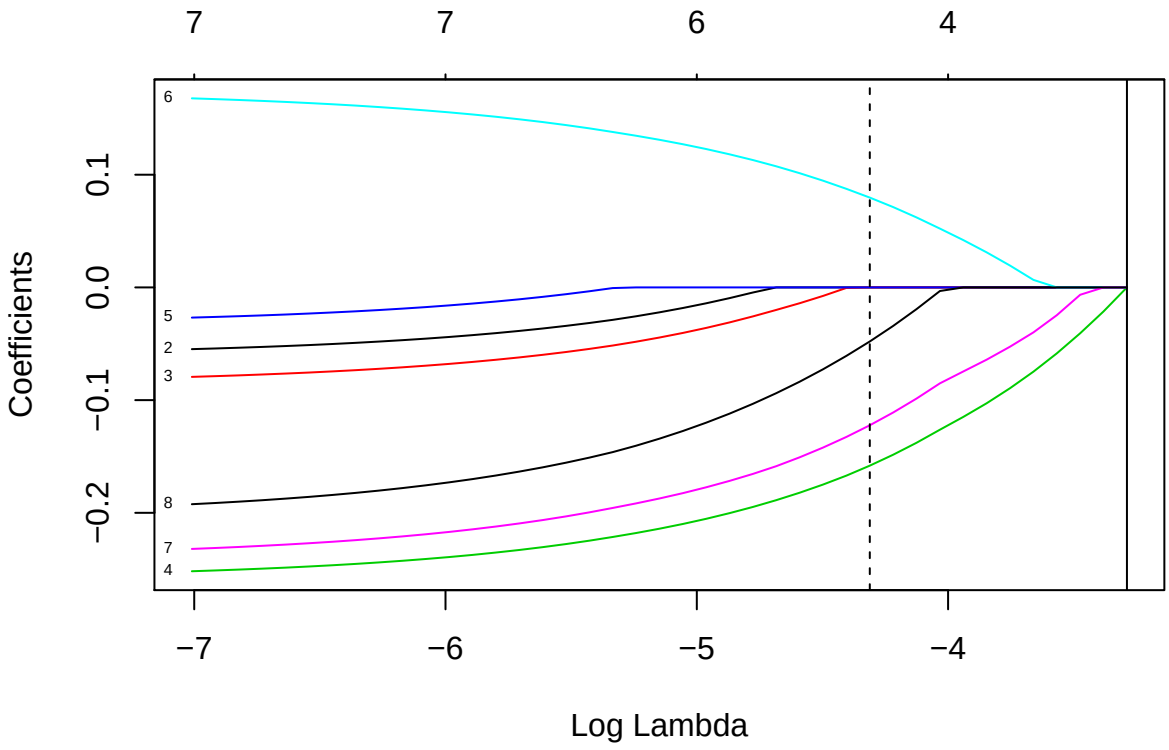
```
## Loaded glmnet 4.0-2
```

```
y<-as.matrix(data$obstruct) #grab response
x<-model.matrix(obstruct~rx+sex+age_c+differ+extent+time_c+nodes_c,data=data)[,-1] #grab predictors
x<-scale(x)
glm(y~x,family=binomial)
```

```
##
## Call:  glm(formula = y ~ x, family = binomial)
##
## Coefficients:
## (Intercept)       xrxLev   xrxLev+5FU         xsex       xage_c      xdiffer
##    -1.48575     -0.00815     -0.06483     -0.08567     -0.25873     -0.03327
##     xextent      xtime_c     xnodes_c
##     0.17472     -0.24063     -0.20345
##
## Degrees of Freedom: 887 Total (i.e. Null);  879 Residual
## Null Deviance:       870.1
## Residual Deviance: 847   AIC: 865
```

```
cv <- cv.glmnet(x,y, family="binomial") #picks an optimal value for lambda through 10-fold CV

#make a plot of the coefficients for different values of lambda
{plot(cv$glmnet.fit, "lambda", label=TRUE); abline(v = log(cv$lambda.1se)); abline(v = log(cv$lambda.mi
```




```
cv<-cv.glmnet(x,y,family="binomial")
lasso<-glmnet(x,y,family="binomial",lambda=cv$lambda.1se)
coef(lasso)
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept) -1.433412e+00
## rxLev        .
## rxLev+5FU    .
## sex          .
## age_c       -1.339574e-16
## differ       .
## extent       .
## time_c       .
## nodes_c      .
```

The only retained variable after LASSO is age_c.

The logistic regression was rerun using only the LASSO retained variable age_c.

```
#cross-validating lasso model
set.seed(1234)
k=10

datalasso<-data[sample(nrow(data)),] #randomly order rows
folds<-cut(seq(1:nrow(datalasso)),breaks=k,labels=F) #create folds

diags<-NULL
```

```
for(i in 1:k){
  train <- datalasso[folds!=i,] #create training set (all but fold i)
  test <- datalasso[folds==i,] #create test set (just fold i)
  truth <- test$obstruct #save truth labels from fold i

  fit <- glm(obstruct~age_c,
             data=train, family="binomial")
  probs <- predict(fit, newdata=test, type="response")

  diags<-rbind(diags,class_diag(probs,truth))
}

diags%>%summarize_all(mean)
```

```
##         acc sens spec ppv  f1       auc
## 1 0.8074183    0    1 NaN NaN 0.5528919
```

After performing a 10-fold CV using only the variables lasso selected: the model's out-of-sample AUC, 0.5527102, was worse than the regression above, 0.5654982.