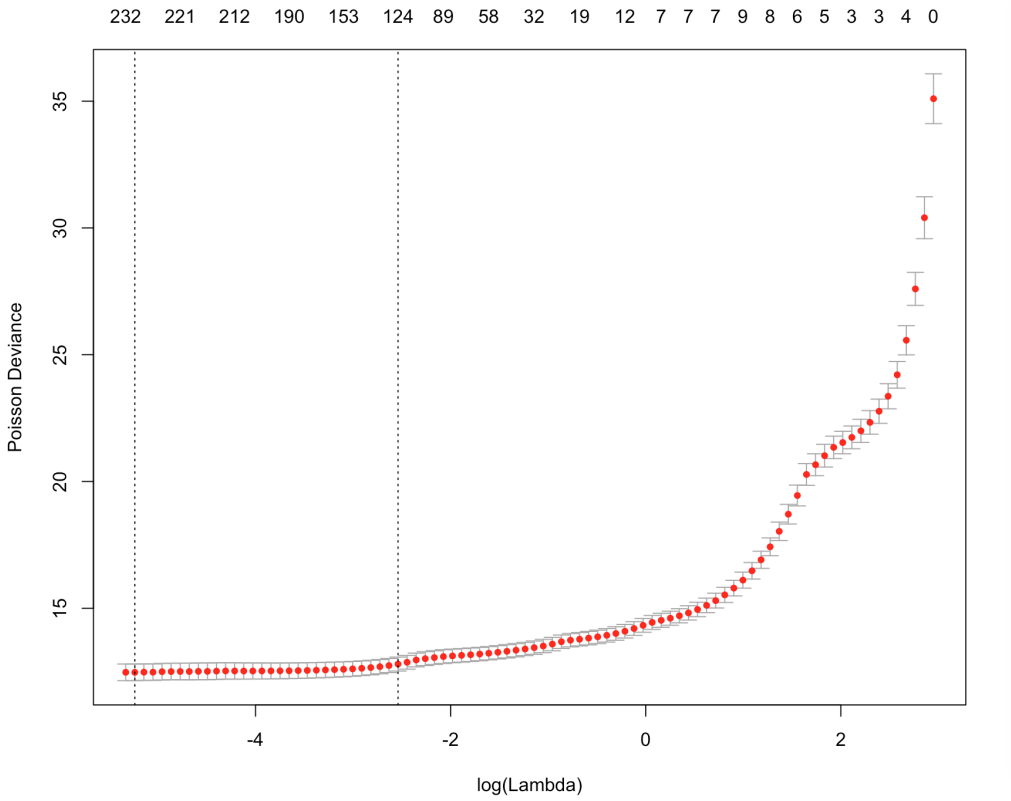
CS498 AML HW7

Huiyun Wu (hwu63)
Yidi Yang (yyang160)

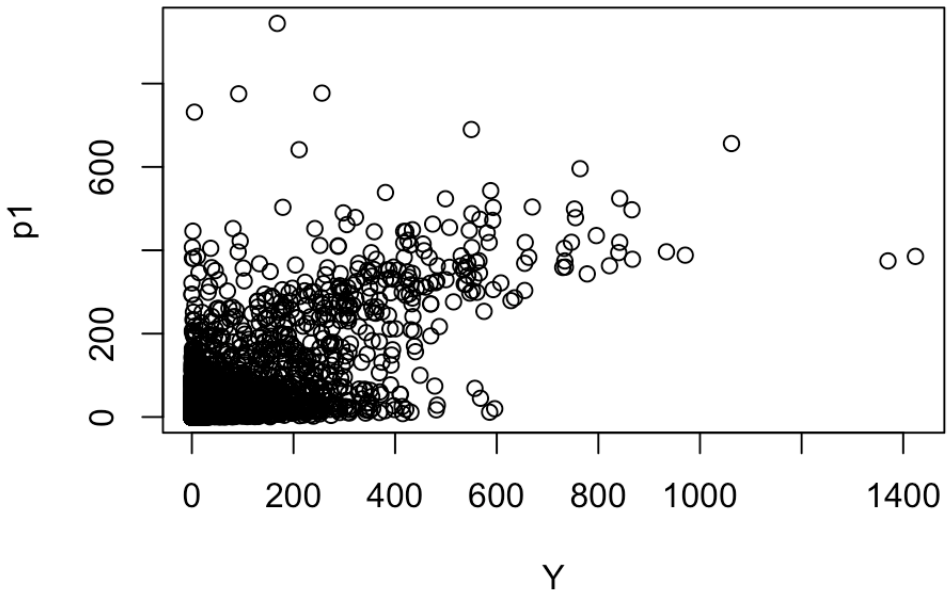
12.3

1. Plot of the cross-validated deviance of the model.



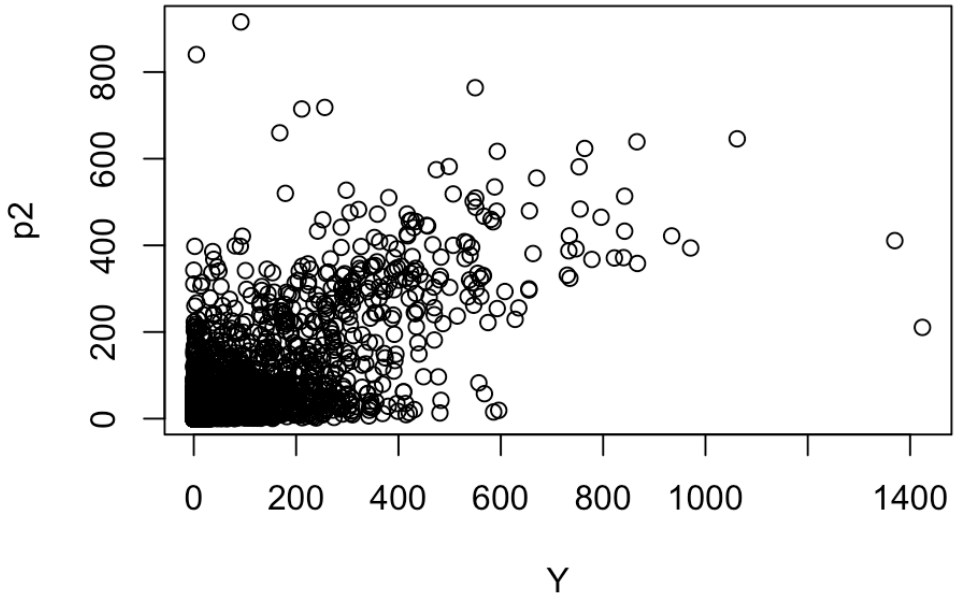
2. Scatter plots of true values vs predicted values on training data.

a. Lambda = "lambda.1se" = 0.07892024.



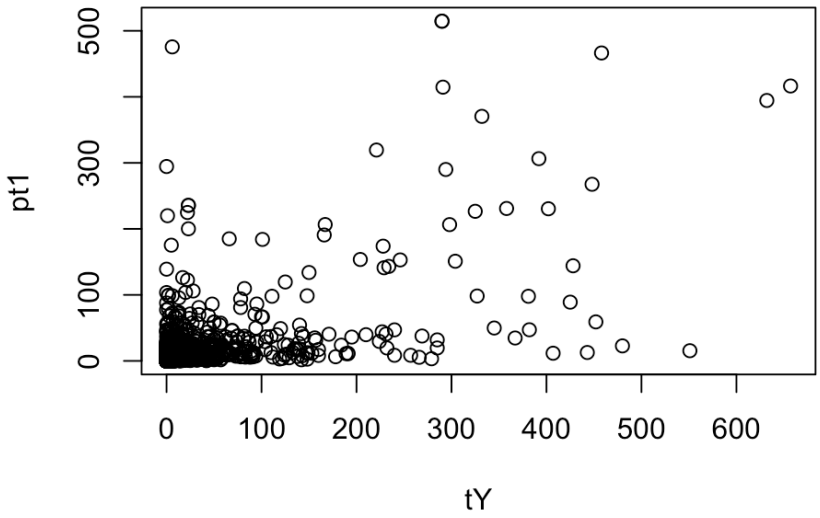
b.

c. Lambda = "lambda.min" = 0.005314608.

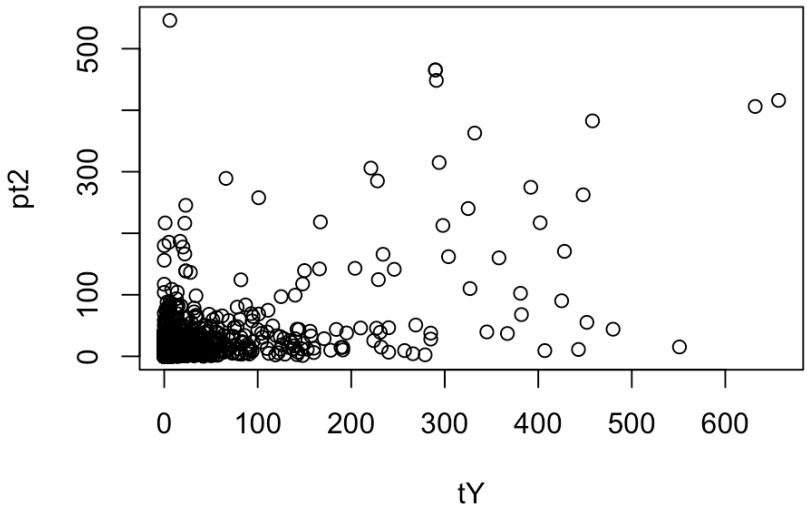


3. Scatter plots of true values vs predicted values on testing data.

a. Lambda = "lambda.1se" = 0.07892024.



b. Lambda = "lambda.min" = 0.005314608.



Performance of the model:

The model performs not ideally. According to the plots above, the scales of the true values and the predicted ones differ and there are many obvious points that are predicted wrong.

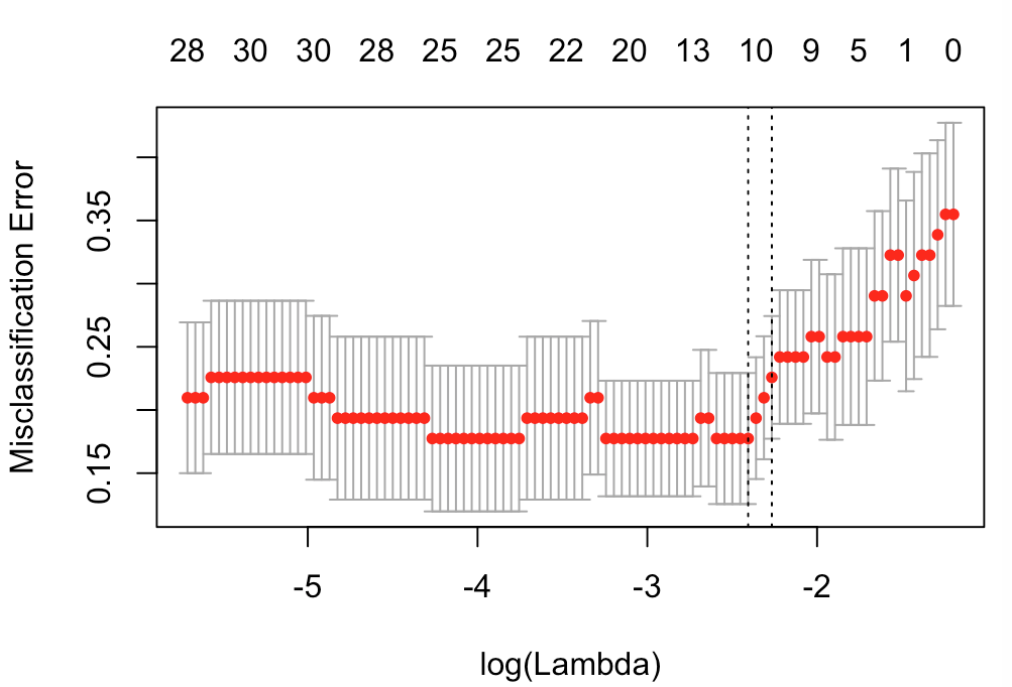
Why is the model difficult?

The true values contain too many zeros so that the training data has bias and the dependent variable is less relevant to the features. Those that has non-zero values are sparse and look like outliers on the plots, which may also influence the reliability of the model.

12.4

4. Plot of the (mis)classification error of the model against the regularization variable.

(family = “binomial”, type.measure = “class”)

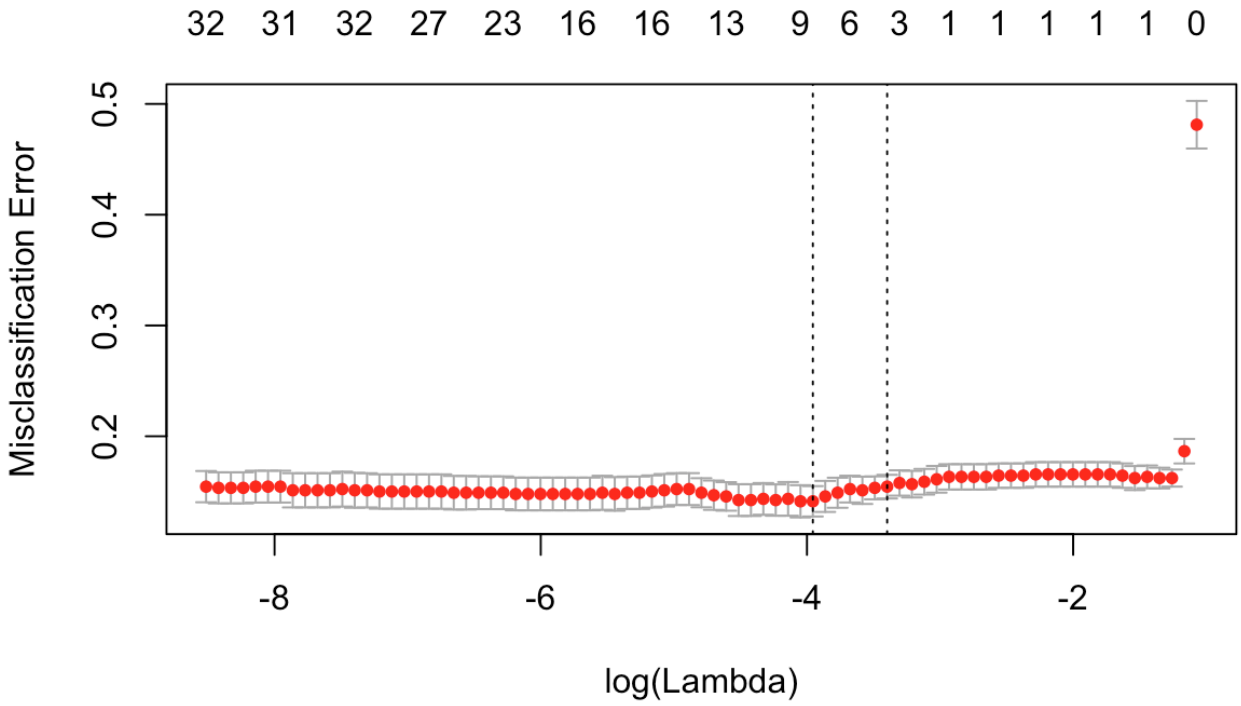


Compare with the baseline prediction:

The common type in this case is tumor. The accuracy of baseline prediction is 0.6451613, with 40 tumor samples and 22 normal ones. The accuracy of the logistic model using lambda = 0.07485259 (lambda.min) is 0.9354839. The logistic model has significantly higher accuracy than the baseline prediction model does.

5. Plot of the classification error of the model against the regularization variable.

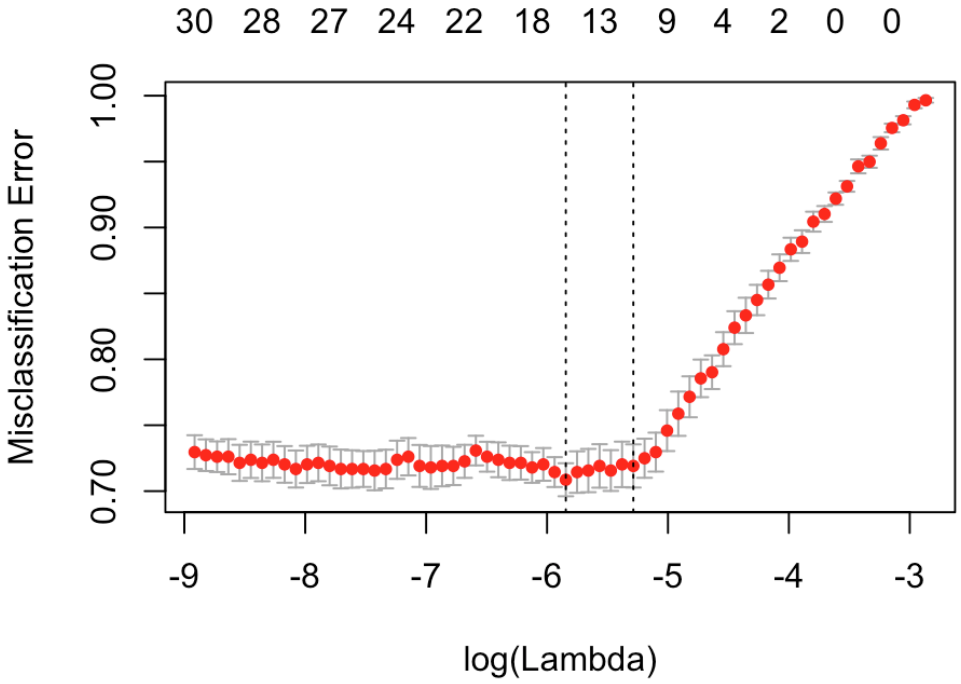
(family = “binomial”, alpha=1, type.measure = “class”)



Compare with the baseline prediction:

The common type in this case is male. The accuracy of baseline prediction is 0.50883. The accuracy of the logistic model with lasso using lambda = 0.01202526 (lambda.min) is 0.8509934. The logistic model has significantly higher accuracy than the baseline prediction model does.

6. Plot of the classification error of the model against the regularization variable.



Compare with the baseline prediction:

Since there are multiple common types, we choose "BXD100" at random to do the comparison. The accuracy of baseline prediction using class "BXD100" (predicting all the instances as BXD100) is 0.02331002. The accuracy of the logistic model with lasso using lambda = 0.01202526 (lambda.min) is 0.6130536. The logistic model has significantly higher accuracy than the baseline.

7. Code Screenshots.

```r
library(glmnet)
library(readr)
# -----------------12.3--------------------
blogData <- read_csv("~/Downloads/BlogFeedback/blogData_train.csv", col_names = FALSE)
X <- as.matrix(blogData[,c(1:280)])
Y <- as.matrix(blogData[,281])
cv <- cv.glmnet(X, Y, family="poisson")
plot(cv)

p1 <- predict(cv, X, s='lambda.1se', type='response')
p2 <- predict(cv, X, s='lambda.min', type = 'response')
plot(Y, p1)
plot(Y, p2)

setwd("~/Downloads/BlogFeedback/")
testfiles <- list.files(pattern = "*00.csv")
testdata <- lapply(testfiles, function(i){
  read.csv(i, header=FALSE)
})
library(data.table)
testdata <- rbindlist(testdata)
tX <- as.matrix(testdata[,c(1:280)])
tY <- as.matrix(testdata[,281])
pt1 <- predict(cv, tX, s='lambda.1se', type='response')
pt2 <- predict(cv, tX, s='lambda.min', type='response')
plot(tY, pt1)
plot(tY, pt2)
```

```r
# ----------------12.4-------------------
genedata <- read.table("~/Desktop/CS498AML/HW7/gene.txt", quote = "\"")
tissues <- read.table("~/Desktop/CS498AML/HW7/tissues.txt", quote="\"")
genedata <- as.matrix(t(genedata))
tissues[tissues >0 ] = 1
tissues[tissues <0] = 0
table(tissues)
tissues <- as.matrix(tissues)
logis <- cv.glmnet(genedata, tissues, type.measure = "class", family="binomial")
plot(logis)
blacc <- 40/62
print(blacc)
results <- predict(logis, genedata, s="lambda.min")
results[results > mean(results)] = 1
results[results < mean(results)] = 0
mean(results==tissues)
```

(Continued)

```r
#------------------12.5------------------
micedata <- read_csv("~/Downloads/Crusio1.csv", col_names = TRUE)
micedata <- micedata[,c(2,4:41)]
micedata <- micedata[complete.cases(micedata),]
micedataX <- as.matrix(micedata[,c(2:39)])
micedataY <- as.matrix(micedata[,1])

micedataY[micedataY == 'f'] = 0
micedataY[micedataY == 'm'] = 1
table(micedataY)
logislasso <- cv.glmnet(micedataX, micedataY, type.measure = "class", alpha = 1, family="binomial")
plot(logislasso)
gender <- predict(logislasso, micedataX)
gender <- ifelse(gender > mean(gender), 1, 0)
mean(gender==micedataY,alpha=1)

print(logislasso$lambda.min)

micedata <- read_csv("~/Downloads/Crusio1.csv", col_names = TRUE)
micedata <- micedata[,c(1,4:41)]
micedata <- micedata[complete.cases(micedata),]
micedata <- micedata[micedata$strain %in% names(table(micedata$strain))[table(micedata$strain) > 9], ]

micedataY <- as.matrix(micedata[,1])
micedataX <- as.matrix(micedata[,c(2:39)])
micedataY <- as.matrix(micedataY[micedataY %in% names(table(micedataY))[table(micedataY) > 9], ])

logislasso <- cv.glmnet(micedataX, micedataY, type.measure = "class", alpha= 1, family="multinomial")
plot(logislasso)
mean(predict(logislasso, micedataX, s="lambda.min", type="class") == micedataY, alpha=1)
table(micedataY)
print(logislasso$lambda.min)
```

(End)