In [1]:

```
library(boot)
library(glmnet)
```

```
Loading required package: Matrix
Loading required package: foreach
Loaded glmnet 2.0-16
```

In [2]:

```
set.seed(2019)
```

In [3]:

```
#TRAIN 1: ALL COVARIATES PLUS INTERACTION TERMS
train1 <- read.csv("trainC.csv")
test1 <- read.csv("testC.csv")
train1 <- subset(train1, select = -c(sessionDate, trialNum, timeSinceKetamine, animalName))
test1 <- subset(test1, select = -c(sessionDate, trialNum, timeSinceKetamine, animalName))

#TRAIN 2: ALL COVARIATES NO INTERACTION TERMS
train2 <- subset(train1, select = c(totalCellNum,gender,genotype,weight_g,ketamine_day,
                                    correlationScore,lickAccuracy,lickNumber,avgFR,
                                    avgSingleCellVariance,varianceFR,avgTrialSpeed,
                                    varianceSpeed,medianCellDepth,ketBool))
test2 <- subset(test1, select = c(totalCellNum,gender,genotype,weight_g,ketamine_day,
                                    correlationScore,lickAccuracy,lickNumber,avgFR,
                                    avgSingleCellVariance,varianceFR,avgTrialSpeed,
                                    varianceSpeed,medianCellDepth,ketBool))
```

Model Generation and Test Error Estimation

Basic Logistic Regression Model with Interaction Terms

In [4]:

```
k = 10
n = length(train1[,1])
fsize = round(n/k)
rmse = rep(0,k)
zoloss = rep(0,k)
for (i in 1:(k-1)){
    # Get train and validation sets
    df_train <- train1[-(((i-1)*fsize+1):(i*fsize)),]
    df_val <- train1[((i-1)*fsize+1):(i*fsize),]
    # Fit model on training and make predictions on validation
    model_cv <- glm(ketBool ~ ., data=df_train, family='binomial')
    lr_pred_lo <- predict(model_cv,df_val) # lo : log odds
    num_val = length(df_val$ketBool)
    lr_pred = rep(0,num_val)
    actual = rep(0,num_val)
    for (j in 1:num_val){
        if (lr_pred_lo[j]>0){
            lr_pred[j]=1
        }
    actual[j] = df_val$ketBool[j]
    }
    # Compute 0-1 loss for each observation
    lr_loss = abs(lr_pred-actual) # loss is 0 if NB_pred=actual, 1 otherwise
    # Compute mean 0-1 loss on the val set
    zoloss[i] = mean(lr_loss)
}
df_train <- train1[-(((k-1)*fsize+1):n),]
df_val <- train1[((k-1)*fsize+1):n,]
# Fit model on training and make predictions on validation
model_cv <- glm(ketBool ~ ., data=df_train, family='binomial')
lr_pred_lo <- predict(model_cv,df_val) # lo : log odds
num_val = length(df_val$ketBool)
lr_pred = rep(0,num_val)
actual = rep(0,num_val)
for (j in 1:num_val){
    if (lr_pred_lo[j]>0){
        lr_pred[j]=1
    }
    actual[j] = df_val$ketBool[j]
}
lr_loss = abs(lr_pred-actual)
zoloss[k] = mean(lr_loss)
test_error_est = mean(zoloss)

cat("======================================================================\n")
cat("Logistic Regression Model with Interaction Terms\n\n")
cat("Zero-One Loss (10-fold Cross-Validation Average):",test_error_est,"\n")
cat("Accuracy (10-fold Cross-Validation Average):",1-test_error_est,"\n")
cat("======================================================================\n")

# Train now on entire training set to get model for prediction
model1 <- glm(ketBool ~ ., data=train1, family='binomial')
```

```
======================================================================
Logistic Regression Model with Interaction Terms

Zero-One Loss (10-fold Cross-Validation Average): 0.09182746
Accuracy (10-fold Cross-Validation Average): 0.9081725
======================================================================
```

Basic Logistic Regression without Interaction Terms

In [5]:

```
k = 10
n = length(train2[,1])
fsize = round(n/k)
rmse = rep(0,k)
zoloss = rep(0,k)
for (i in 1:(k-1)){
    # Get train and validation sets
    df_train <- train2[-(((i-1)*fsize+1):(i*fsize)),]
    df_val <- train2[((i-1)*fsize+1):(i*fsize),]
    # Fit model on training and make predictions on validation
    model_cv <- glm(ketBool ~ ., data=df_train, family='binomial')
    lr_pred_lo <- predict(model_cv,df_val) # lo : log odds
    num_val = length(df_val$ketBool)
    lr_pred = rep(0,num_val)
    actual = rep(0,num_val)
    for (j in 1:num_val){
        if (lr_pred_lo[j]>0){
            lr_pred[j]=1
        }
    actual[j] = df_val$ketBool[j]
    }
    # Compute 0-1 loss for each observation
    lr_loss = abs(lr_pred-actual) # loss is 0 if NB_pred=actual, 1 otherwise
    # Compute mean 0-1 loss on the val set
    zoloss[i] = mean(lr_loss)
}
df_train <- train2[-(((k-1)*fsize+1):n),]
df_val <- train2[((k-1)*fsize+1):n,]
# Fit model on training and make predictions on validation
model_cv <- glm(ketBool ~ ., data=df_train, family='binomial')
lr_pred_lo <- predict(model_cv,df_val) # lo : log odds
num_val = length(df_val$ketBool)
lr_pred = rep(0,num_val)
actual = rep(0,num_val)
for (j in 1:num_val){
    if (lr_pred_lo[j]>0){
        lr_pred[j]=1
    }
    actual[j] = df_val$ketBool[j]
}
lr_loss = abs(lr_pred-actual)
zoloss[k] = mean(lr_loss)
test_error_est = mean(zoloss)

cat("=====================================================================\n")
cat("Logistic Regression Model without Interaction Terms\n\n")
cat("Zero-One Loss (10-fold Cross-Validation Average):",test_error_est,"\n")
cat("Accuracy (10-fold Cross-Validation Average):",1-test_error_est,"\n")
cat("=====================================================================\n")

# Train now on entire training set to get model for prediction
model2 <- glm(ketBool ~ ., data=train2, family='binomial')
```

```
=====================================================================
Logistic Regression Model without Interaction Terms

Zero-One Loss (10-fold Cross-Validation Average): 0.1413709
Accuracy (10-fold Cross-Validation Average): 0.8586291
=====================================================================
```

Look at Coefficients on TRAIN

Model 1 (including interaction terms) Summary

In [10]:

```
summary(model1)
```

```
Call:
glm(formula = ketBool ~ ., family = "binomial", data = train1)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.9532  -0.2677   0.0100   0.2355   5.0333  

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)                           -3.07543    0.84032  -3.660 0.000252 ***
totalCellNum                           1.48468    0.36237   4.097 4.18e-05 ***
gender                                 5.27077    0.80163   6.575 4.86e-11 ***
genotype                               4.70327    0.73513   6.398 1.58e-10 ***
weight_g                               0.12584    0.35738   0.352 0.724760    
ketamine_day                          -0.71834    0.25786  -2.786 0.005340 ** 
correlationScore                      -2.30386    0.96069  -2.398 0.016479 *  
lickAccuracy                          -2.59838    0.62801  -4.137 3.51e-05 ***
lickNumber                            -0.51051    0.76272  -0.669 0.503289    
avgFR                                  4.72545    1.78614   2.646 0.008154 ** 
avgSingleCellVariance                  7.85496    2.25356   3.486 0.000491 ***
varianceFR                             1.58919    1.02274   1.554 0.120216    
avgTrialSpeed                         -0.91611    0.73195  -1.252 0.210715    
varianceSpeed                          1.67179    1.06811   1.565 0.117539    
medianCellDepth                        1.26773    0.30867   4.107 4.01e-05 ***
totalCellNumxCorrelationScore         -2.45892    0.29758  -8.263  < 2e-16 ***
totalCellNumxLickAccuracy              0.24686    0.16550   1.492 0.135803    
totalCellNumxLickNumber                0.08498    0.19849   0.428 0.668571    
totalCellNumxAvgFR                     1.43740    0.56874   2.527 0.011493 *  
totalCellNumxAvgSingleCellVariance    -0.67645    0.40254  -1.680 0.092871 .  
totalCellNumxVarianceFR               -0.86265    0.16684  -5.171 2.33e-07 ***
totalCellNumxAvgTrialSpeed            -0.94312    0.24154  -3.905 9.44e-05 ***
totalCellNumxVarianceSpeed             0.75321    0.28168   2.674 0.007495 ** 
genderxCorrelationScore               -0.08490    0.24236  -0.350 0.726094    
genderxLickAccuracy                    0.28144    0.12668   2.222 0.026304 *  
genderxLickNumber                     -0.14458    0.10581  -1.366 0.171828    
genderxAvgFR                          -2.35801    0.80157  -2.942 0.003264 ** 
genderxAvgSingleCellVariance           0.54791    0.54906   0.998 0.318326    
genderxVarianceFR                      0.08090    0.30108   0.269 0.788160    
genderxAvgTrialSpeed                  -1.05481    0.29317  -3.598 0.000321 ***
genderxVarianceSpeed                  -1.15447    0.38986  -2.961 0.003064 ** 
genotypexCorrelationScore             -0.33733    0.22293  -1.513 0.130233    
genotypexLickAccuracy                  0.56838    0.16717   3.400 0.000674 ***
genotypexLickNumber                   -0.52894    0.14448  -3.661 0.000251 ***
genotypexAvgFR                         0.92673    0.72838   1.272 0.203262    
genotypexAvgSingleCellVariance        -5.99671    0.85149  -7.043 1.89e-12 ***
genotypexVarianceFR                   -0.22840    0.33708  -0.678 0.498033    
genotypexAvgTrialSpeed                -0.49684    0.32787  -1.515 0.129686    
genotypexVarianceSpeed                -0.61546    0.44262  -1.391 0.164377    
weight_gxCorrelationScore              3.16988    0.79975   3.964 7.38e-05 ***
weight_gxLickAccuracy                  0.94502    0.50304   1.879 0.060297 .  
weight_gxLickNumber                    0.04100    0.56794   0.072 0.942451    
weight_gxAvgFR                        -1.81136    1.68183  -1.077 0.281473    
weight_gxAvgSingleCellVariance        -1.28619    1.45427  -0.884 0.376470    
weight_gxVarianceFR                   -0.93719    0.97722  -0.959 0.337537    
weight_gxAvgTrialSpeed                 0.54390    0.57514   0.946 0.344310    
weight_gxVarianceSpeed                -2.35374    0.78367  -3.003 0.002669 ** 
ketamine_dayxCorrelationScore          0.41956    0.33708   1.245 0.213237    
ketamine_dayxLickAccuracy             -0.20682    0.20144  -1.027 0.304539    
ketamine_dayxLickNumber                0.45125    0.22458   2.009 0.044502 *  
ketamine_dayxAvgFR                     1.44940    0.96651   1.500 0.133715    
ketamine_dayxAvgSingleCellVariance    -0.92388    0.94597  -0.977 0.328747    
ketamine_dayxVarianceFR               -0.76557    0.45440  -1.685 0.092027 .  
ketamine_dayxAvgTrialSpeed             0.24416    0.32816   0.744 0.456857    
ketamine_dayxVarianceSpeed             1.47173    0.38611   3.812 0.000138 ***
medianCellDepthxCorrelationScore      -1.44980    0.32376  -4.478 7.53e-06 ***
medianCellDepthxLickAccuracy           0.36188    0.22533   1.606 0.108280    
medianCellDepthxLickNumber            -0.08501    0.26385  -0.322 0.747295    
medianCellDepthxAvgFR                 -2.05059    0.75163  -2.728 0.006368 ** 
medianCellDepthxAvgSingleCellVariance -2.85366    0.94725  -3.013 0.002590 ** 
medianCellDepthxVarianceFR             0.20696    0.27064   0.765 0.444430    
medianCellDepthxAvgTrialSpeed          1.63901    0.31478   5.207 1.92e-07 ***
medianCellDepthxVarianceSpeed         -1.95645    0.45961  -4.257 2.07e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5538.9  on 3996  degrees of freedom
Residual deviance: 1794.6  on 3934  degrees of freedom
AIC: 1920.6
```

```
Number of Fisher Scoring iterations: 7
```

Model 2 (not including interaction terms) Summary

In [11]:

```
summary(model2)
```

```
Call:
glm(formula = ketBool ~ ., family = "binomial", data = train2)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.1024  -0.4780   0.0845   0.4688   3.8002  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -0.64251    0.15630  -4.111 3.94e-05 ***
totalCellNum          -0.03728    0.06442  -0.579 0.562759    
gender                 0.46578    0.13882   3.355 0.000793 ***
genotype               0.05546    0.11283   0.492 0.623009    
weight_g              -0.46373    0.06503  -7.131 9.94e-13 ***
ketamine_day           0.16121    0.04486   3.594 0.000326 ***
correlationScore      -1.50123    0.07377 -20.351  < 2e-16 ***
lickAccuracy          -0.80960    0.05888 -13.750  < 2e-16 ***
lickNumber            -0.60947    0.06549  -9.306  < 2e-16 ***
avgFR                  1.77153    0.12885  13.749  < 2e-16 ***
avgSingleCellVariance -1.32595    0.11463 -11.567  < 2e-16 ***
varianceFR            -0.23810    0.05966  -3.991 6.58e-05 ***
avgTrialSpeed         -0.18518    0.05614  -3.298 0.000973 ***
varianceSpeed         -0.99420    0.07504 -13.249  < 2e-16 ***
medianCellDepth       -0.04057    0.05383  -0.754 0.451108    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5538.9  on 3996  degrees of freedom
Residual deviance: 2741.0  on 3982  degrees of freedom
AIC: 2771

Number of Fisher Scoring iterations: 6
```

Test Performance

In [6]:

```
lr_pred_lo <- predict(model1,test1) # lo : log odds
num_val = length(test1$ketBool)
lr_pred = rep(0,num_val)
actual = rep(0,num_val)
for (j in 1:num_val){
    if (lr_pred_lo[j]>0){
        lr_pred[j]=1
    }
    actual[j] = test1$ketBool[j]
}
lr_loss = abs(lr_pred-actual)
zoloss[k] = mean(lr_loss)
test_error_est = mean(zoloss)

cat("======================================================================\n")
cat("Logistic Regression Model with Interaction Terms\n\n")
cat("Zero-One Loss (Test Set):",test_error_est,"\n")
cat("Accuracy (Test Set):",1-test_error_est,"\n")
cat("======================================================================\n")
```

```
======================================================================
Logistic Regression Model with Interaction Terms

Zero-One Loss (Test Set): 0.13475 
Accuracy (Test Set): 0.86525 
======================================================================
```

In [8]:

```r
lr_pred_lo <- predict(model2,test2) # lo : log odds
num_val = length(test2$ketBool)
lr_pred = rep(0,num_val)
actual = rep(0,num_val)
for (j in 1:num_val){
    if (lr_pred_lo[j]>0){
        lr_pred[j]=1
    }
    actual[j] = test2$ketBool[j]
}
lr_loss = abs(lr_pred-actual)
zoloss[k] = mean(lr_loss)
test_error_est = mean(zoloss)

cat("=====================================================================\n")
cat("Logistic Regression Model without Interaction Terms\n\n")
cat("Zero-One Loss (Test Set):",test_error_est,"\n")
cat("Accuracy (Test Set):",1-test_error_est,"\n")
cat("=====================================================================\n")
```

```
=====================================================================
Logistic Regression Model without Interaction Terms

Zero-One Loss (Test Set): 0.14175
Accuracy (Test Set): 0.85825
=====================================================================
```

Look at Coefficients on TEST

With Interaction Terms

In [15]:

```
model1_test <- glm(ketBool ~ ., data=test1, family='binomial')
summary(model1_test)
```

```
Warning message:
“glm.fit: fitted probabilities numerically 0 or 1 occurred”
```

```
Call:
glm(formula = ketBool ~ ., family = "binomial", data = test1)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.0289  -0.1986   0.0016   0.1849   2.9694  

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)                           -3.65227    2.05234  -1.780 0.075147 .  
totalCellNum                           0.26300    0.88535   0.297 0.766426    
gender                                 7.81598    1.83003   4.271 1.95e-05 ***
genotype                               7.42112    1.86461   3.980 6.89e-05 ***
weight_g                               1.15238    0.83398   1.382 0.167037    
ketamine_day                          -1.58289    0.63604  -2.489 0.012823 *  
correlationScore                      -2.20880    2.19214  -1.008 0.313647    
lickAccuracy                          -2.24522    1.44600  -1.553 0.120494    
lickNumber                             0.33636    1.85504   0.181 0.856112    
avgFR                                  4.23584    3.76048   1.126 0.259992    
avgSingleCellVariance                  9.97080    5.09635   1.956 0.050411 .  
varianceFR                             3.36081    2.66836   1.260 0.207848    
avgTrialSpeed                          0.29209    1.81798   0.161 0.872356    
varianceSpeed                          3.03594    2.45716   1.236 0.216627    
medianCellDepth                        1.65360    0.72813   2.271 0.023146 *  
totalCellNumxCorrelationScore         -3.01461    0.66719  -4.518 6.23e-06 ***
totalCellNumxLickAccuracy             -0.01844    0.42563  -0.043 0.965452    
totalCellNumxLickNumber                0.85883    0.46080   1.864 0.062355 .  
totalCellNumxAvgFR                    -0.72389    1.28868  -0.562 0.574299    
totalCellNumxAvgSingleCellVariance     2.13417    1.06527   2.003 0.045134 *  
totalCellNumxVarianceFR               -1.20751    0.36359  -3.321 0.000897 ***
totalCellNumxAvgTrialSpeed            -1.15633    0.56101  -2.061 0.039289 *  
totalCellNumxVarianceSpeed             2.58590    0.57912   4.465 8.00e-06 ***
genderxCorrelationScore               -0.49791    0.55891  -0.891 0.373005    
genderxLickAccuracy                    0.08964    0.28202   0.318 0.750598    
genderxLickNumber                      0.50079    0.34721   1.442 0.149218    
genderxAvgFR                          -4.79261    1.83800  -2.608 0.009120 ** 
genderxAvgSingleCellVariance           2.02599    1.25764   1.611 0.107191    
genderxVarianceFR                      0.60497    0.64171   0.943 0.345811    
genderxAvgTrialSpeed                  -1.38870    0.69665  -1.993 0.046219 *  
genderxVarianceSpeed                  -2.08394    1.04008  -2.004 0.045109 *  
genotypexCorrelationScore             -0.65555    0.55225  -1.187 0.235209    
genotypexLickAccuracy                  0.28092    0.36612   0.767 0.442899    
genotypexLickNumber                   -0.37848    0.60850  -0.622 0.533948    
genotypexAvgFR                        -1.08513    1.96100  -0.553 0.580020    
genotypexAvgSingleCellVariance        -6.50830    2.23454  -2.913 0.003585 ** 
genotypexVarianceFR                   -0.04429    0.81545  -0.054 0.956689    
genotypexAvgTrialSpeed                -0.48904    0.76063  -0.643 0.520263    
genotypexVarianceSpeed                -0.02729    1.12641  -0.024 0.980672    
weight_gxCorrelationScore              4.20517    1.80688   2.327 0.019949 *  
weight_gxLickAccuracy                  0.53872    1.01541   0.531 0.595730    
weight_gxLickNumber                   -2.76258    1.46563  -1.885 0.059443 .  
weight_gxAvgFR                         3.28156    3.49283   0.940 0.347467    
weight_gxAvgSingleCellVariance        -5.86363    3.35396  -1.748 0.080417 .  
weight_gxVarianceFR                   -4.58480    1.94465  -2.358 0.018391 *  
weight_gxAvgTrialSpeed                -0.06083    1.38059  -0.044 0.964858    
weight_gxVarianceSpeed                -5.94149    1.73747  -3.420 0.000627 ***
ketamine_dayxCorrelationScore         -0.10723    0.80114  -0.134 0.893524    
ketamine_dayxLickAccuracy              0.31969    0.45310   0.706 0.480469    
ketamine_dayxLickNumber                0.37617    0.53723   0.700 0.483798    
ketamine_dayxAvgFR                     3.33903    2.03621   1.640 0.101042    
ketamine_dayxAvgSingleCellVariance    -0.31757    2.03221  -0.156 0.875822    
ketamine_dayxVarianceFR               -0.62787    1.00364  -0.626 0.531581    
ketamine_dayxAvgTrialSpeed             0.07342    0.73297   0.100 0.920208    
ketamine_dayxVarianceSpeed             1.03936    0.87252   1.191 0.233567    
medianCellDepthxCorrelationScore      -1.59720    0.70336  -2.271 0.023158 *  
medianCellDepthxLickAccuracy           0.32039    0.49388   0.649 0.516519    
medianCellDepthxLickNumber             1.08690    0.77763   1.398 0.162200    
medianCellDepthxAvgFR                 -4.70330    1.56983  -2.996 0.002735 ** 
medianCellDepthxAvgSingleCellVariance -2.55433    2.04819  -1.247 0.212353    
medianCellDepthxVarianceFR             0.90155    0.98434   0.916 0.359726    
medianCellDepthxAvgTrialSpeed          1.31766    0.76205   1.729 0.083792 .  
medianCellDepthxVarianceSpeed         -1.42238    1.01890  -1.396 0.162715    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.19  on 999  degrees of freedom
Residual deviance:  399.57  on 937  degrees of freedom
AIC: 525.57
```

```
Number of Fisher Scoring iterations: 8
```

Without Interaction Terms

In [16]:

```
model2_test <- glm(ketBool ~ ., data=test2, family='binomial')
summary(model2_test)
```

```
Call:
glm(formula = ketBool ~ ., family = "binomial", data = test2)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.52524  -0.51562   0.05817   0.49094   3.03239  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -0.11054    0.29360  -0.377 0.706540    
totalCellNum          -0.20103    0.12503  -1.608 0.107873    
gender                 0.38001    0.26237   1.448 0.147507    
genotype               0.28291    0.21514   1.315 0.188518    
weight_g              -0.21089    0.13190  -1.599 0.109851    
ketamine_day          -0.07733    0.08284  -0.934 0.350543    
correlationScore      -1.27870    0.13062  -9.789  < 2e-16 ***
lickAccuracy          -0.77094    0.10958  -7.035 1.99e-12 ***
lickNumber            -0.69481    0.14704  -4.725 2.30e-06 ***
avgFR                  1.96396    0.25970   7.562 3.96e-14 ***
avgSingleCellVariance -1.21420    0.22392  -5.423 5.88e-08 ***
varianceFR            -0.54385    0.14694  -3.701 0.000215 ***
avgTrialSpeed         -0.22271    0.11054  -2.015 0.043934 *  
varianceSpeed         -0.79791    0.14073  -5.670 1.43e-08 ***
medianCellDepth       -0.14587    0.10636  -1.371 0.170241    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.19  on 999  degrees of freedom
Residual deviance:  712.67  on 985  degrees of freedom
AIC: 742.67

Number of Fisher Scoring iterations: 6
```

In [ ]: