STAT 426 Assignment 12

Due Tuesday, December 7, 11:59 pm.

Submit through Moodle.

Name: Brianna Diaz

Netid: bdiaz22

Submit your work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown. Be sure to show your work.

Problem 1. (10 pts)

The data file "endometrial.txt" is from a study of endometrial cancer. The variables are: NV = neovasculation (1 = present, 0 = absent); PI = pulsatility index of arteria uteria, EH = endometrium height, and HG = histology evaluation (0 = low grade, 1 = high grade).

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

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

```
library(readr)
library(glmnet)
```

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

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

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

```
## Loaded glmnet 4.1-3
```

```
cancer <- read.table("endometrial.txt", header =  TRUE)
```

a) (2 pts) Fit and summarize an additive logistic regression model with HG as the response variable and the other variables as predictors, using ordinary maximum likelihood.

```
prob1_a <- glm(HG ~ ., family = "binomial", data = cancer )

summary(prob1_a)
```

```
##
## Call:
## glm(formula = HG ~ ., family = "binomial", data = cancer)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.50137  -0.64108  -0.29432   0.00016   2.72777
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.30452    1.63730   2.629 0.008563 **
## NV           18.18556 1715.75089   0.011 0.991543
## PI           -0.04218    0.04433  -0.952 0.341333
## EH           -2.90261    0.84555  -3.433 0.000597 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 104.903  on 78  degrees of freedom
## Residual deviance:  55.393  on 75  degrees of freedom
## AIC: 63.393
##
## Number of Fisher Scoring iterations: 17
```

b) (2 pts) Try using the drop1 function to simplify the model. Is there a better model than the three variable model, according to AIC and/or the likelihood ratio tests?

```
drop1(prob1_a)
```

```
## Single term deletions
##
## Model:
## HG ~ NV + PI + EH
##        Df Deviance    AIC
## <none>      55.393 63.393
## NV      1   64.751 70.751
## PI      1   56.378 62.378
## EH      1   75.154 81.154
```

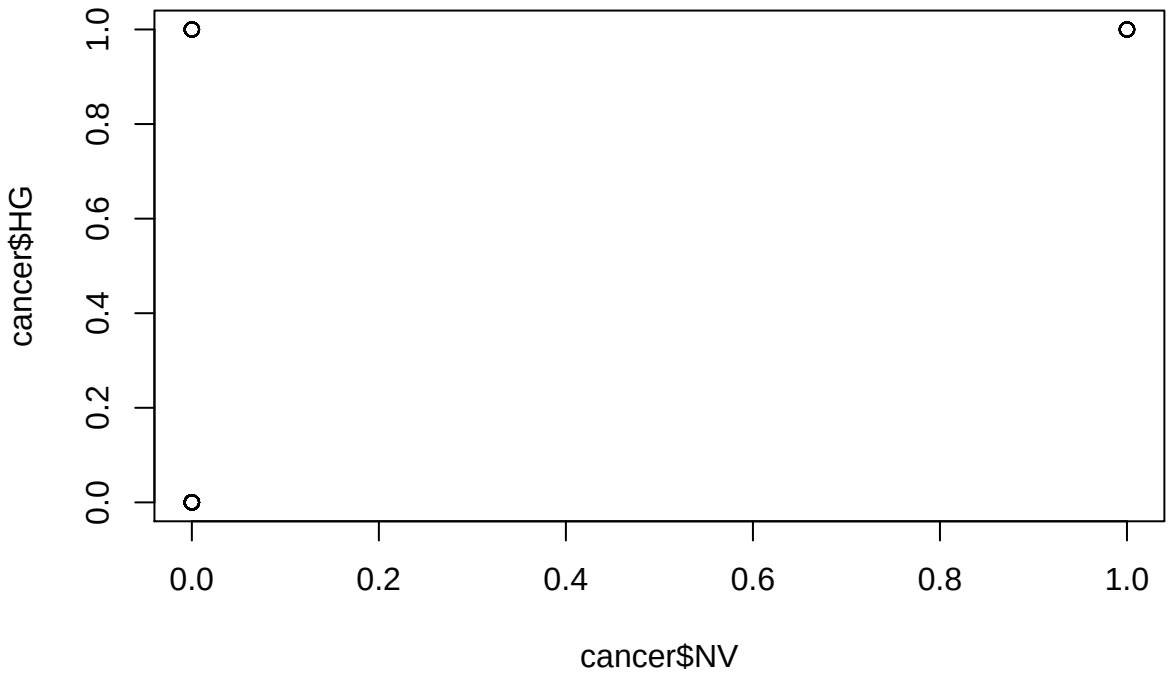
```
drop1(prob1_a, test = "LRT")
```

```
## Single term deletions
##
## Model:
## HG ~ NV + PI + EH
##        Df Deviance    AIC     LRT  Pr(>Chi)
## <none>      55.393 63.393
## NV      1   64.751 70.751  9.3576  0.002221 **
## PI      1   56.378 62.378  0.9851  0.320934
## EH      1   75.154 81.154 19.7606 8.777e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#No there is not a better model
```

c) (2 pts) Make a scatter plot of the response versus NV. What is this situation called? Does it indicate a problem for the estimated coefficient of NV and its standard error? If so, what is the problem?

```
plot(cancer$NV, cancer$HG )
```



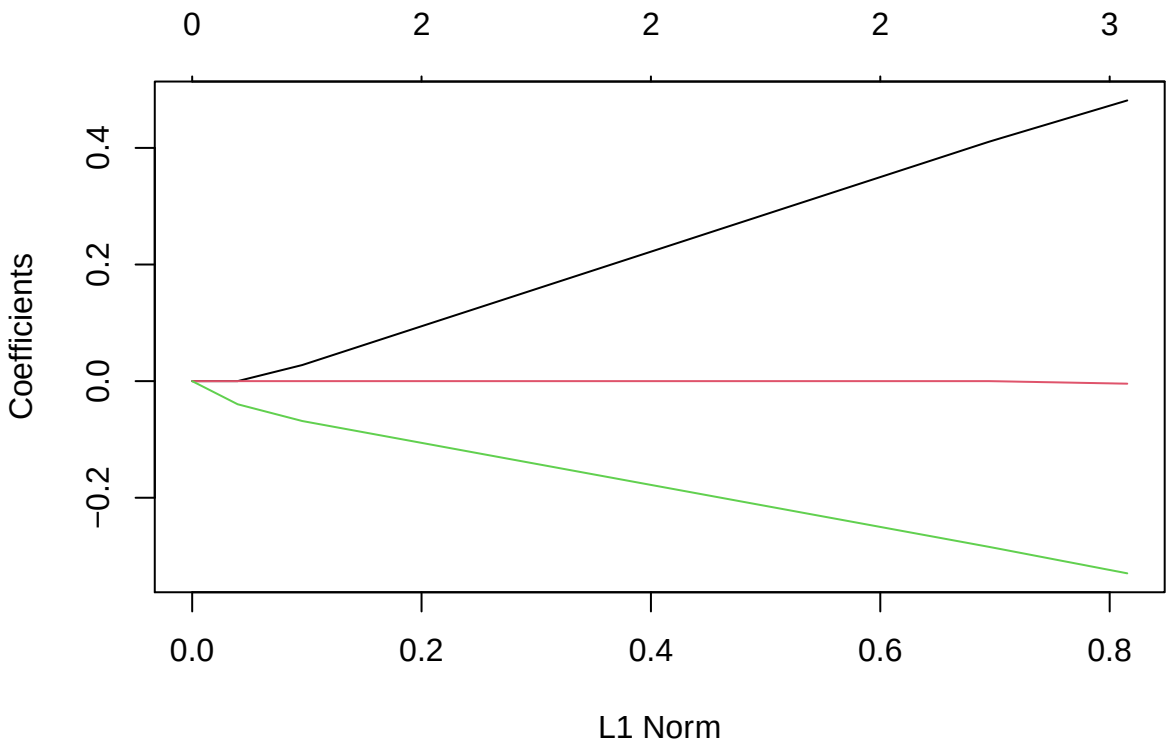

```
#There is really no patter in the graph. Therefore the coefficients and the standard e
```

d) (2 pts) Use the lasso penalty to fit the logistic model for HG using all three predictors. Display the coefficient plot.

```
X = as.matrix(cancer[,-4])
y = cancer[,4]
modlasso <- glmnet(X,y)

plot(modlasso, lable = TRUE, lwd = 1)
```

```
## Warning in plot.window(...): "lable" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "lable" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "lable" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "lable" is not a
## graphical parameter

## Warning in box(...): "lable" is not a graphical parameter
```

```
## Warning in title(...): "lable" is not a graphical parameter
```



e) (2 pts) Show the coefficient estimates for $\lambda = 0.25$. How are these different from the coefficients in the maximum likelihood fit? What causes the magnitude of the NV coefficient to be so much smaller using the lasso than it was using unconstrained maximum likelihood?

```
coef(modlasso, s = 0.25)
```

```
## 4 x 1 sparse Matrix of class "dgCMatrix"
##                      s1
## (Intercept)  0.47745019
## NV           0.02035102
## PI           .
## EH          -0.06081456
```

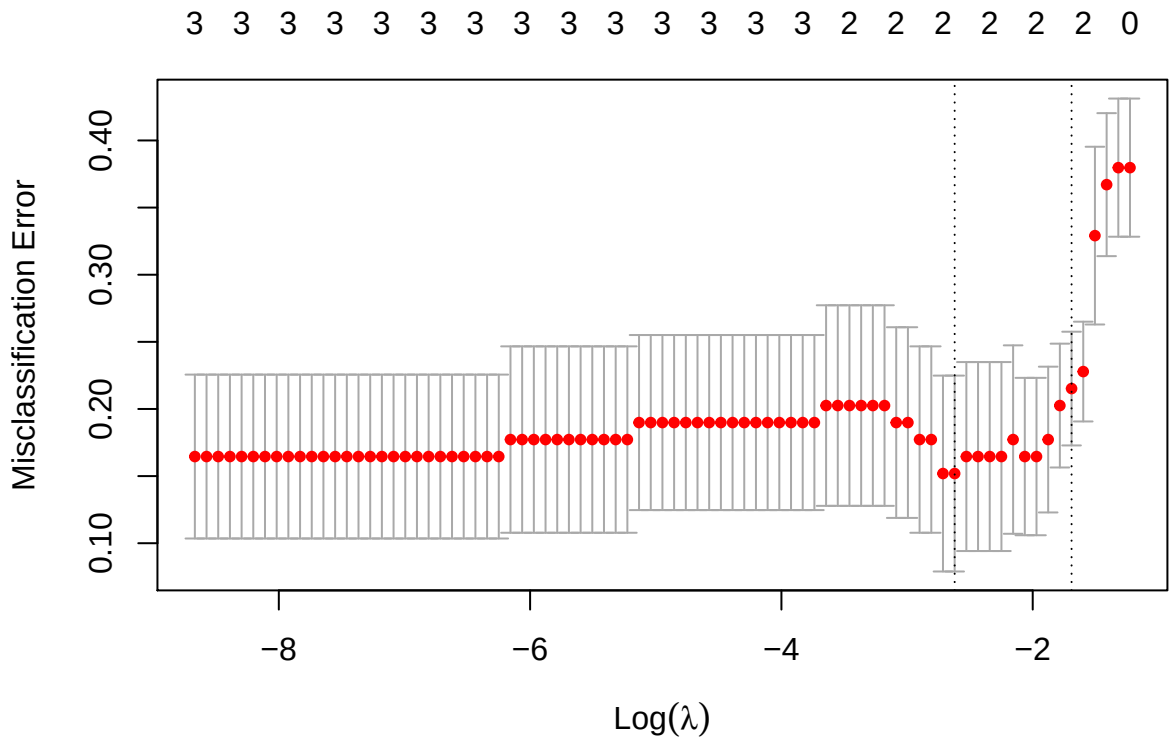
```
#These coefficients are much smaller than max likelihood fit.

#This is because there are only a few predictors effecting the response
```

Problem 2 (6 pts)

This problem also refers to the endometrial cancer data from Problem 1.

a) (2 pts) Use **5-fold** cross validation with classification error as the performance measure (see the cv.glmnet documentation for how to select the number of "folds" – the default is 10-fold). Show the plot of misclassification error versus λ or $\log(\lambda)$. Note: in order to make your results reproducible, set the seed before running the cross-validation function.

```
set.seed(1121)
prob2_a <- cv.glmnet(X,y, family = "binomial", type.measure = "class", nfolds = 5)
plot(prob2_a)
```



b) (2 pts) Show the value of λ that minimized the misclassification error using 5-fold cross-validation. Also show the coefficients of the corresponding model. Which, if any, variables were eliminated?

```
prob2_a$lambda.min
```

```
## [1] 0.07272276
```

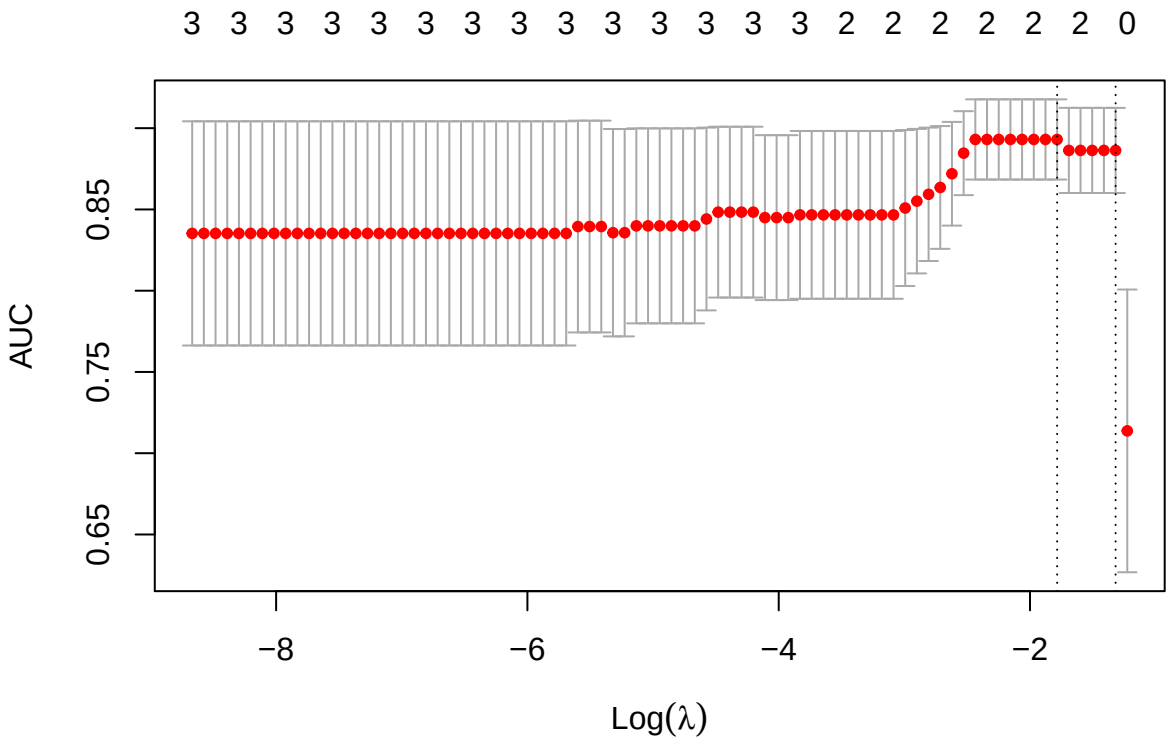
```
coef(prob2_a, s = prob2_a$lambda.min)
```

```
## 4 x 1 sparse Matrix of class "dgCMatrix"
##                      s1
## (Intercept)  1.651154
## NV           1.482165
## PI           .
## EH          -1.509018
```

```
# The PI variable was eliminated.
```

c) (2 pts) Now use 5-fold cross validation with "auc" as the performance measure (see documentation for how to do this). Show the plot of AUC versus λ or $\log(\lambda)$. Which value of λ maximizes auc?

```
set.seed(1121)
prob2_c <- cv.glmnet(X,y, family = "binomial", type.measure = "auc", nfolds = 5)
plot(prob2_c)
```




```
prob2_a$lambda.min
```

```
## [1] 0.07272276
```

```
# auc is maximized at 0.07272276
```

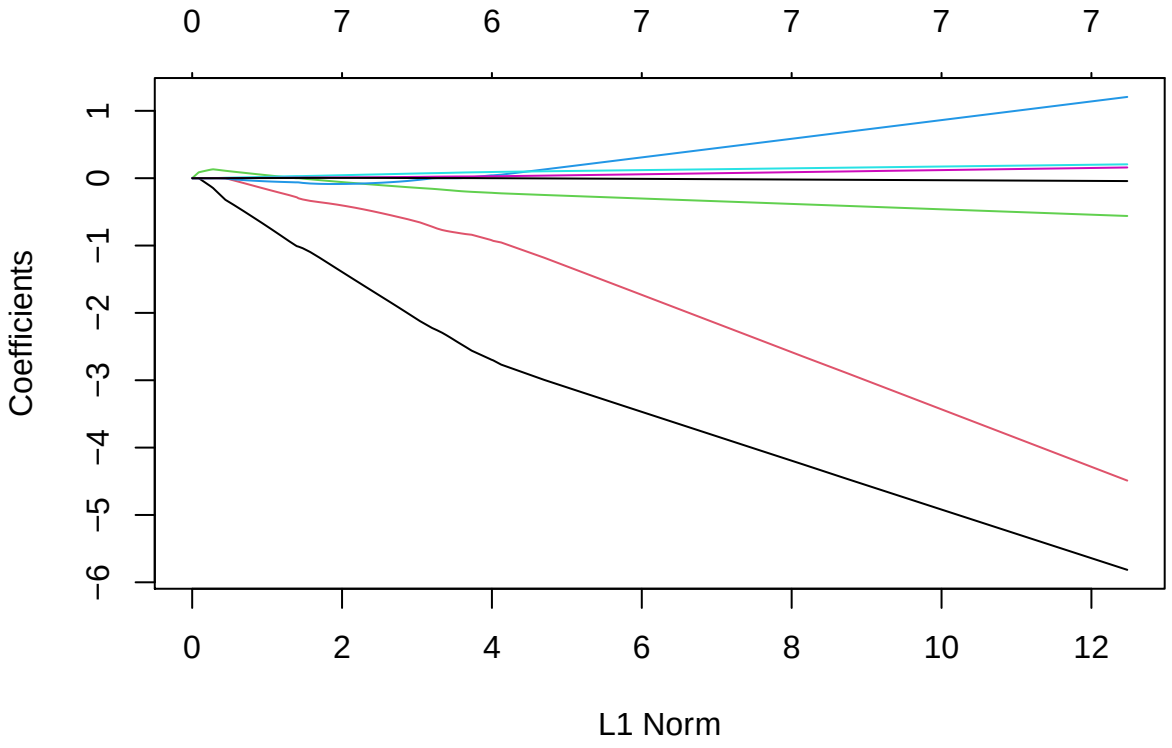
Problem 3 (4 pts)

The data file "horseshoe.txt" is from a study of horseshoe crab mating patterns. The relevant variables are the `color` index, `spine` length, `width`, and `weight` of the female, and the response `satell`, the number of males attached.

```
crab <- read.table('horseshoe.txt', header = TRUE)
```

a) (2 pts) Use the elastic net method with $\alpha = 0.7$ to fit a Poisson log-linear model with numerical variables color, spine, width and weight. Show the coefficient plot versus L1 norm, and the estimated coefficients for $\lambda = 0.1$.

```
XX <- model.matrix(~color*spine*width, data = crab)
yy <- crab$satell
wt = crab$weight
mfit <- glmnet(XX, yy, weights = wt, family = "poisson", alpha = 0.7)

plot(mfit)
```



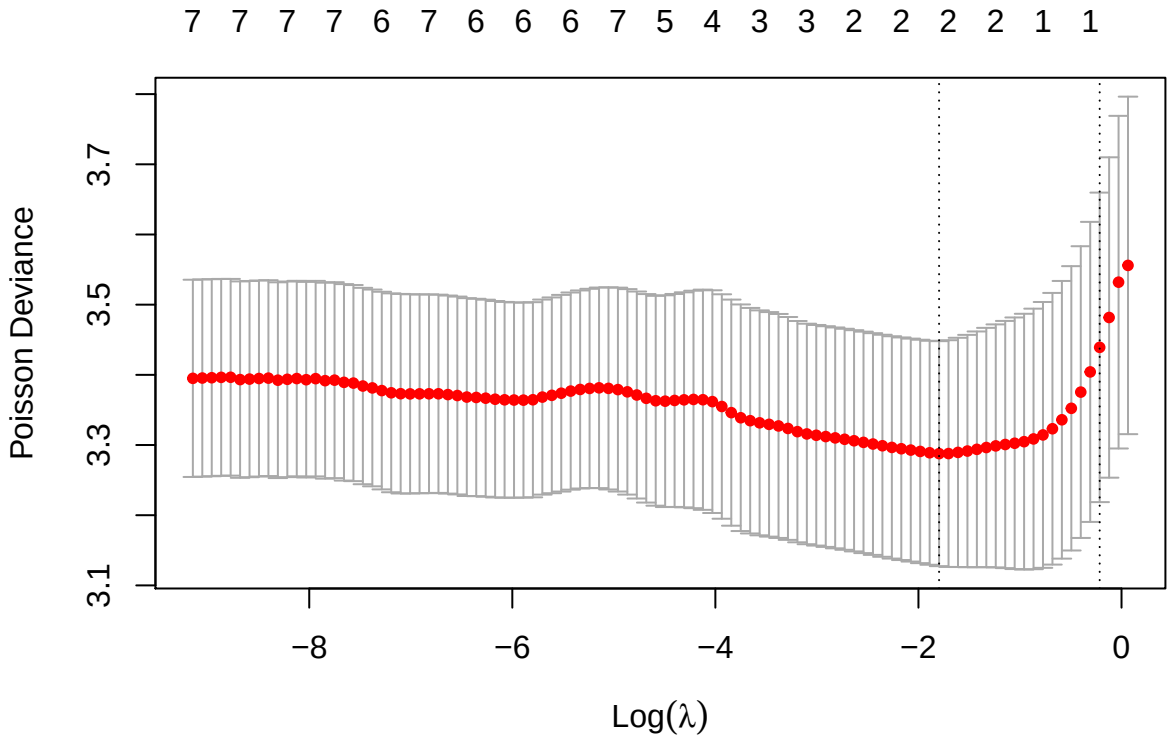

```
coef(mfit, s = 0.1)
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                              s1
## (Intercept)       -1.8579313
## (Intercept)        .
## color             -0.1192929
## spine              .
## width              0.1261913
## color:spine        .
```

```
## color:width        .
## spine:width        .
## color:spine:width  .
```

b) (2 pts) Use 10-fold cross validation with performance measure "deviance" to select the best value for λ. Display the performance graph (deviance versus λ or $\log(\lambda)$, as well as the value of λ that minimizes cross-validation deviance, and the corresponding coefficient estimates for the model. Which, if any, predictors were eliminated?

```
prob3_b <- cv.glmnet(XX, yy, weights = wt, family = "poisson", type.measure = "deviance"
plot(prob3_b)
```




```
prob3_b$lambda.min
```

```
## [1] 0.1658556
```

```
coef(prob3_b, s=prob3_b$lambda.min)
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                              s1
## (Intercept)       -1.73368178
## (Intercept)        .
## color             -0.08372958
```

```
## spine              .
## width              0.11739108
## color:spine        .
## color:width        .
## spine:width        .
## color:spine:width  .
```

```
# All the interactions, and spine were eliminated.
```