Homework 4 MC Solutions

ISYE 6414 Instructor

In this problem, we will return to the study of the relationship between geographic access to healthcare services and severe health outcomes for pediatric asthma. Particularly, we will perform variable selection to identify which predicting variables are selected for explaining the variations in the proportion of the emergency department (ED) visits encountered by children in the state of Georgia.

Reference: Garcia, E., Serban, N., Swann, J., Fitzpatrick, A. (2015)"A study of the Impact of Geographic Access on Severe Health Outcomes for Pediatric Asthma", Journal of Allergy and Clinical Immunology, 136(3):610-8.

The response data and the predicting variables are observed for all counties in Georgia. Counties form a contiguous geographic division of a state in the United States, for example, Georgia has 159 different counties. The first column in the data file (County) specifies the names of the counties.

- The response variable:

ED visits: Number of ED visits for children in each county in 2010

- Candidate explanatory variables are the following in the data file:

A5.9: A binary variable specifying whether children ages 5 to 9 have had ED visits for each county

A10.14: A binary variable specifying whether children ages 10 to 14 have had ED visits for each county

No.Hospitals: Number of hospitals in each county

PercentLessHS: Percentage of the population without a high school degree in each county

PercentHS: Percentage of the population with a high school degree only in each county

MedianIncome: Median household income in each county

SpecDist: The average travel distance to an asthma specialist within each county (a measure of access)

PedDist: The average travel distance to a pediatrician within each county (a measure of access)

Use the following code to get started for this problem. Please make sure that you are using R version 3.6.X and above, i.e. version 4.X is also acceptable.

```
# Ensure that the sampling type is correct
RNGkind(sample.kind="Rejection")

# Reading the data:
data = read.csv("GA_EDVisits.csv",header=TRUE)
data = na.omit(data)

# Get names of the column
names = colnames(data)
attach(data)
```

```
# Standardized predictors - use these variables in your modeling in addition
# to the predictors A5.9, A10.14
sAvgDistS = scale(log(SpecDist))
sAvgDistP = scale(log(PedDist))
sMedianIncome = scale(MedianIncome)
sNumHospitals = scale(No.Hospitals)
sPercentLessHS = scale(PercentLessHS)
sPercentHS = scale(PercentHS)

# Define interaction terms
DistA5.9 = sAvgDistS*A5.9
DistA10.14 = sAvgDistS* A10.14
DistIncome = sAvgDistS*sMedianIncome
DistLessHS = sAvgDistS*sPercentLessHS
DistHS = sAvgDistS*sPercentHS
DistPA5.9 = sAvgDistP*A5.9
DistPA10.14 = sAvgDistP* A10.14
DistPIncome = sAvgDistP*sMedianIncome
DistPLessHS = sAvgDistP*sPercentLessHS
DistPHS = sAvgDistP*sPercentHS

# Define final data frame
X = data.frame(A5.9, A10.14, sAvgDistS, sAvgDistP, sMedianIncome,sPercentLessHS,
               sPercentHS,sNumHospitals,DistA5.9, DistA10.14, DistIncome,
               DistLessHS, DistHS, DistPA5.9,
               DistPA10.14, DistPIncome,DistPLessHS, DistPHS)

# Set Seed to 100
set.seed(100)
```

Model Setup:

- Define *model1* as a Poisson regression with ED.visits as the response and all predicting variables in X, except for the 10 interaction terms.
- Define *model2* s a Poisson regression with ED.visits as the response and all predicting variables in X.

Note: Some of the questions below may have more than one correct answer. Select all that are correct.

Question 21

Fit *model1* and *model2*. Which of the following is true?

- ☐ All regression coefficients are statistically significant at the alpha level 0.01 in model1
- ☐ All regression coefficients for the main effects (i.e. non-interaction terms) are statistically significant at the alpha level of 0.05 in model2
- ☐ All regression coefficients for the interaction terms are statistically significant at the alpha level of 0.05 in model2
- ☒ The AIC of model2 is less than the AIC of model1

```r
model1 = glm(ED.visits~., family="poisson", data=X[,1:8])
summary(model1)
```

```
##
## Call:
## glm(formula = ED.visits ~ ., family = "poisson", data = X[, 1:8])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -15.2681   -2.9139   -0.7568    1.6152   21.8397
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.884785   0.023579  79.935  < 2e-16 ***
## A5.9            1.381304   0.023274  59.350  < 2e-16 ***
## A10.14          0.896530   0.024688  36.314  < 2e-16 ***
## sAvgDistS      -0.013103   0.011705  -1.119    0.263
## sAvgDistP      -0.622280   0.013035 -47.738  < 2e-16 ***
## sMedianIncome  -0.062438   0.012108  -5.157 2.51e-07 ***
## sPercentLessHS -0.256131   0.015824 -16.186  < 2e-16 ***
## sPercentHS     -0.274635   0.015677 -17.519  < 2e-16 ***
## sNumHospitals   0.246244   0.003165  77.808  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 50167.7  on 455  degrees of freedom
## Residual deviance:  9415.9  on 447  degrees of freedom
## AIC: 11285
##
## Number of Fisher Scoring iterations: 5
```

```r
model2=glm(data$ED.visits~., family="poisson", data=X)
summary(model2)
```

```
##
## Call:
## glm(formula = data$ED.visits ~ ., family = "poisson", data = X)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -14.831   -2.769   -1.030    1.698   18.007
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.926077   0.033529  57.445  < 2e-16 ***
## A5.9            1.238610   0.037526  33.007  < 2e-16 ***
## A10.14          0.825022   0.039581  20.844  < 2e-16 ***
## sAvgDistS      -0.188698   0.029599  -6.375 1.83e-10 ***
## sAvgDistP      -0.375824   0.029194 -12.873  < 2e-16 ***
## sMedianIncome  -0.023043   0.021137  -1.090 0.275635
```

```
## sPercentLessHS -0.201507   0.017560 -11.475  < 2e-16 ***
## sPercentHS     -0.123806   0.019640  -6.304 2.90e-10 ***
## sNumHospitals   0.264129   0.004560  57.925  < 2e-16 ***
## DistA5.9        0.004448   0.031580   0.141 0.887999
## DistA10.14      0.067756   0.033340   2.032 0.042125 *
## DistIncome     -0.390545   0.017087 -22.856  < 2e-16 ***
## DistLessHS     -0.100253   0.017830  -5.623 1.88e-08 ***
## DistHS         -0.468708   0.018077 -25.928  < 2e-16 ***
## DistPA5.9      -0.121907   0.030453  -4.003 6.25e-05 ***
## DistPA10.14    -0.112413   0.032239  -3.487 0.000489 ***
## DistPIncome     0.424868   0.019439  21.856  < 2e-16 ***
## DistPLessHS     0.285962   0.021364  13.385  < 2e-16 ***
## DistPHS         0.429543   0.015802  27.183  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 50168  on 455  degrees of freedom
## Residual deviance:  8095  on 437  degrees of freedom
## AIC: 9984.1
##
## Number of Fisher Scoring iterations: 5
```

```
AIC(model2) < AIC(model1)
```

```
## [1] TRUE
```

Question 22

Next you will perform a forward-backward stepwise regression using AIC.

Let the minimum and starting model be the model with no interactions (model1) and the full model be the one with interactions (model2).

Based on this stepwise regression procedure, which of the following are true? (Select all that are true)

- ☒ The first two interaction terms entering the model were DistPHS and DistHS
- ☐ Once each interaction term entered the model, it was never discarded from the model
- ☐ There are regression coefficients for the interactions terms in the final selected model that are not statistically significant at the alpha level of 0.01
- ☒ All regression coefficients associated to the interactions terms in the final selected model are statistically significant at the alpha level of 0.05

```
mod3 = step(model1, scope=list(lower=model1,upper=model2), direction="both", trace=FALSE)
summary(mod3)
```

```
##
## Call:
## glm(formula = ED.visits ~ A5.9 + A10.14 + sAvgDistS + sAvgDistP +
##     sMedianIncome + sPercentLessHS + sPercentHS + sNumHospitals +
##     DistPHS + DistHS + DistIncome + DistPIncome + DistPLessHS +
```

```
##     DistLessHS + DistPA5.9 + DistPA10.14 + DistA10.14, family = "poisson",
##     data = X[, 1:8])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -14.818   -2.773   -1.032    1.709   18.006
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.92690    0.03300  58.386  < 2e-16 ***
## A5.9            1.23757    0.03678  33.649  < 2e-16 ***
## A10.14          0.82420    0.03914  21.060  < 2e-16 ***
## sAvgDistS      -0.18519    0.01602 -11.560  < 2e-16 ***
## sAvgDistP      -0.37789    0.02523 -14.976  < 2e-16 ***
## sMedianIncome  -0.02304    0.02114  -1.090  0.27576
## sPercentLessHS -0.20150    0.01756 -11.475  < 2e-16 ***
## sPercentHS     -0.12380    0.01964  -6.303 2.91e-10 ***
## sNumHospitals   0.26413    0.00456  57.925  < 2e-16 ***
## DistPHS         0.42957    0.01580  27.186  < 2e-16 ***
## DistHS         -0.46875    0.01808 -25.934  < 2e-16 ***
## DistIncome     -0.39057    0.01709 -22.858  < 2e-16 ***
## DistPIncome     0.42488    0.01944  21.857  < 2e-16 ***
## DistPLessHS     0.28597    0.02136  13.385  < 2e-16 ***
## DistLessHS     -0.10027    0.01783  -5.624 1.87e-08 ***
## DistPA5.9      -0.11931    0.02424  -4.923 8.53e-07 ***
## DistPA10.14    -0.11034    0.02870  -3.845  0.00012 ***
## DistA10.14      0.06422    0.02192   2.929  0.00340 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 50167.7  on 455  degrees of freedom
## Residual deviance:  8095.1  on 438  degrees of freedom
## AIC: 9982.1
##
## Number of Fisher Scoring iterations: 5
```

```
#Steps taken in the search
mod3$anova
```

```
##             Step Df     Deviance Resid. Df Resid. Dev       AIC
## 1               NA           NA       447   9415.874 11284.908
## 2      + DistPHS -1 280.56608679       446   9135.308 11006.342
## 3       + DistHS -1 350.47982510       445   8784.828 10657.862
## 4   + DistIncome -1 183.83964730       444   8600.988 10476.023
## 5  + DistPIncome -1 294.85481425       443   8306.133 10183.168
## 6  + DistPLessHS -1 146.39675819       442   8159.737 10038.771
## 7   + DistLessHS -1  30.97742213       441   8128.759 10009.794
## 8     + DistA5.9 -1  17.27556206       440   8111.484  9994.518
## 9    + DistPA5.9 -1   4.35195875       439   8107.132  9992.166
## 10 + DistPA10.14 -1   7.94144350       438   8099.190  9986.225
## 11  + DistA10.14 -1   4.14646216       437   8095.044  9984.078
## 12    - DistA5.9  1   0.01984193       438   8095.064  9982.098
```

```
# Checking with interaction terms were statistically significant at alpha 0.05
alpha <- 0.05
p_values <- summary(mod3)$coef[,4][10:18]
p_values <= alpha
```

```
##     DistPHS      DistHS  DistIncome DistPIncome DistPLessHS  DistLessHS
##        TRUE        TRUE        TRUE        TRUE        TRUE        TRUE
##   DistPA5.9 DistPA10.14  DistA10.14
##        TRUE        TRUE        TRUE
```

Question 23

Which of the following is true regarding the final selected model from the previous question?

- ☒ All but one of the interaction terms are included in the final model
- ☐ All regression coefficients in the final model are significant at the alpha level of 0.05
- ☐ Residual deviance for the final model is lower than the residual deviance for model2
- ☐ The AIC of the final model is at least 80 points lower than the AIC of model2

All but one of the interaction terms (DistA5.9) are included in the final model.

Question 24

Perform Lasso regression with ED.visits as the Poisson response variable and all variables in X as the predicting variables. Use the glmnet() R commands and set the type of measure to be the deviance and the number of folds to be 10 for obtaining the penalty constant lambda (*lambda*) using cross-validation. For this problem, do not force the main effects to be in the model.

Which of the following are true? (Select all that are true)

Note: Please run set.seed(100) before your code for this question.

- ☐ The penalty lambda (λ) value resulting in the lowest cross-validation error is between 0 and 2
- ☒ The first two variables entering the model are *sNumHospitals* and *DistPHS*
- ☒ When log lambda is 0, 15 predicting variables are selected
- ☐ glmnet has estimated a lower coefficient value for *A5.9* than for *A10.14* for all log lambda below 0

```
set.seed(100)
library(glmnet)
```

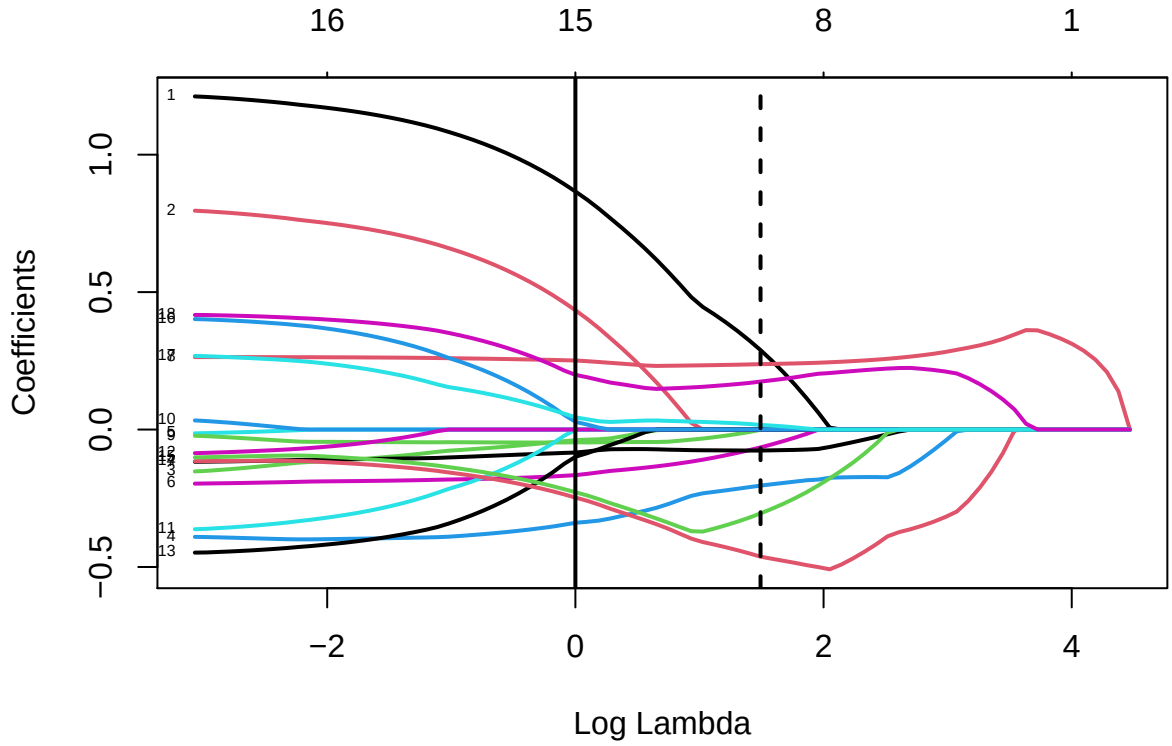
```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1
```

```
mod4.cv =cv.glmnet(as.matrix(X), data$ED.visits,
                   family="poisson",
                   alpha = 1,
                   type.measure ="deviance",
                   nfolds=10)
mod4 = glmnet(as.matrix(X), data$ED.visits,
```

```
              family="poisson",
              alpha=1,
              nlambda = 100)

plot(mod4, xvar="lambda", label=TRUE, lwd=2)
abline(v=log(mod4.cv$lambda.min), col='black', lty=2,lwd=2)
abline(v=0, col='black', lty = 1, lwd=2)
```




```
cat("Lambda with lowest CV error:",
    mod4.cv$lambda.min)
```

```
## Lambda with lowest CV error: 4.446217
```

```
cat("First two variables entering the model: sNumHospitals and DistPHS")
```

```
## First two variables entering the model: sNumHospitals and DistPHS
```

```
coef(mod4, s=mod4.cv$lambda[10])
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                            1
## (Intercept)    3.45517947
## A5.9           .
```

```
## A10.14         .
## sAvgDistS      .
## sAvgDistP      .
## sMedianIncome  .
## sPercentLessHS .
## sPercentHS     .
## sNumHospitals  0.36177400
## DistA5.9       .
## DistA10.14     .
## DistIncome     .
## DistLessHS     .
## DistHS         .
## DistPA5.9      .
## DistPA10.14    .
## DistPIncome    .
## DistPLessHS    .
## DistPHS        0.02109759
```

```
cat("Total variables selected when log-lambda = 0: ",
    sum(coef(mod4, s=exp(0))[,1] != 0) - 1, "\n")
```

```
## Total variables selected when log-lambda = 0:  15
```

```
coef(mod4, s=exp(0))
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                              1
## (Intercept)     2.3078305569
## A5.9            0.8661722627
## A10.14          0.4339963713
## sAvgDistS      -0.0388173214
## sAvgDistP      -0.3393602507
## sMedianIncome   .
## sPercentLessHS -0.1658208217
## sPercentHS     -0.0836761303
## sNumHospitals   0.2512044246
## DistA5.9       -0.0464059826
## DistA10.14      .
## DistIncome     -0.0008863881
## DistLessHS      .
## DistHS         -0.0995677746
## DistPA5.9      -0.2473383110
## DistPA10.14    -0.2278661577
## DistPIncome     0.0286646131
## DistPLessHS     0.0447055714
## DistPHS         0.1992465544
```

Question 25

Which of the following is true regarding the optimal model selected in the previous question?

☒ Neither all main effects nor all interaction terms are selected

☐ All main effects are selected
☐ All interaction terms are selected

```
coef(mod4, s=mod4.cv$lambda.min)
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                              1
## (Intercept)     2.7810757879
## A5.9            0.2891460747
## A10.14          .
## sAvgDistS       .
## sAvgDistP      -0.2039280766
## sMedianIncome   .
## sPercentLessHS -0.0648938468
## sPercentHS     -0.0763077459
## sNumHospitals   0.2379130840
## DistA5.9       -0.0008748252
## DistA10.14      .
## DistIncome      .
## DistLessHS      .
## DistHS          .
## DistPA5.9      -0.4624424045
## DistPA10.14    -0.3050941866
## DistPIncome     .
## DistPLessHS     0.0168001543
## DistPHS         0.1745401458
```

Question 26

Similar to Question 24, perform elastic net regression. Use alpha = 0.5, and do not force the main effects to be in the model.

Note: Please run set.seed(100) before your codes for this question.

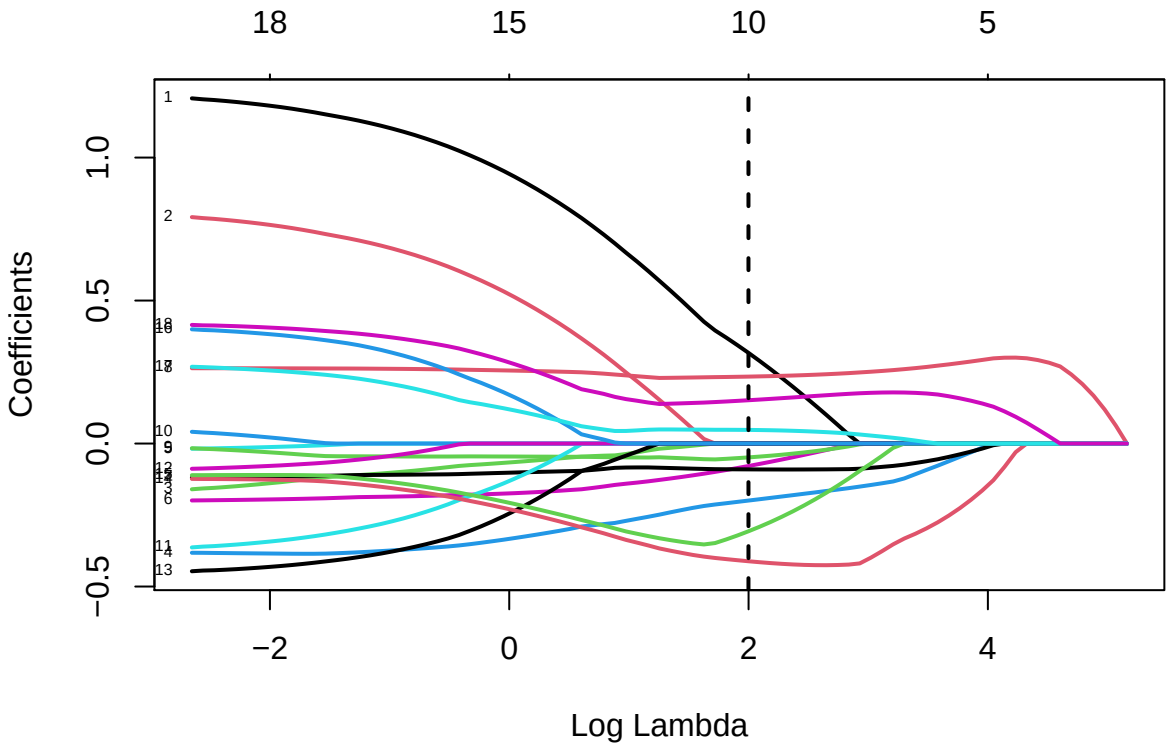
Which of the following is true regarding the optimal model?

☐ All main effects are selected
☐ All interaction terms are selected
☒ Neither all main effects nor all interaction terms are selected

```
set.seed(100)

mod6.cv = cv.glmnet(as.matrix(X), data$ED.visits,
                    family="poisson",
                    alpha=0.5,
                    type.measure ="deviance",
                    nfolds=10)
mod6 = glmnet(as.matrix(X), data$ED.visits,
              family="poisson",
              alpha =0.5,
              nlambda = 100)

plot(mod6, xvar="lambda", label=TRUE, lwd=2)
abline(v=log(mod6.cv$lambda.min), col='black', lty=2, lwd=2)
```




```
cat("Variables selected at optimal lambda value: ")
```

```
## Variables selected at optimal lambda value:
```

```
coef(mod6, s=mod6.cv$lambda.min)
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                            1
## (Intercept)     2.76589120
## A5.9            0.31713232
## A10.14          .
## sAvgDistS       .
## sAvgDistP      -0.19936469
## sMedianIncome   .
## sPercentLessHS -0.07860798
## sPercentHS     -0.09037706
## sNumHospitals   0.23408277
## DistA5.9       -0.04847212
## DistA10.14      .
## DistIncome      .
## DistLessHS      .
## DistHS          .
## DistPA5.9      -0.41156530
## DistPA10.14    -0.30680297
## DistPIncome     .
```

```
## DistPLessHS     0.04768910
## DistPHS         0.15143753
```