Homework 8

Qianbo Wang
uni: qw2180

Problem 1

Reading Assignment: Ramsey and Schafer (2nd Ed), Chapter 22 (Log-Linear Regression for Poisson Counts)

Problem 2

Consider the Valve characteristics data (Display 22.16, Ramsey and Schafer, 2nd Ed).

Valve Failure in Nuclear Reactors. Display 22.16 shows characteristics and numbers of failures observed in valve types from one pressurized water reactor. There are five explanatory factors:

- system (1 = containment, 2 = nuclear, 3 = power conversion, 4 = safety, 5 = process auxiliary);
- operator type (1 = air, 2 = solenoid, 3 = motor-driven, 4 = manual);
- valve type (1 = ball, 2 = butterfly, 3 = diaphragm, 4 = gate, 5 = globe, 6 = directional control);
- head size (1 = less than 2 inches, 2 = 210 inches, 3 = 1030 inches);
- operation mode (1 = normally closed, 2 = normally open).

The lengths of observation periods are quite different, as indicated in the last column, time. Using an offset for log of observation time, identify the factors associated with large numbers of valve failures.

(a). Do Problem Number 24, Page 667, using the R function glm.

First, use all of the variables to fit a log-linear poisson regression model with log(time) as offset. And the result is as follows:

Log-linear Poisson Regression

```
Call:  glm(formula = Failures ~ System + Operator + Valve + Size + Mode,
    family = "poisson", data = Nuclear, offset = log(Time))

Coefficients:
(Intercept)      System2      System3      System4      System5    Operator2    Operator3    Operator4
   -3.76867      0.91556      1.01881      1.22309      0.33292      0.70437     -1.19261     -2.47233
     Valve2       Valve3       Valve4       Valve5       Valve6        Size2        Size3        Mode2
    0.18533      0.60674      2.95894      1.79318      1.00891     -0.01219      1.61457     -0.20934

Degrees of Freedom: 89 Total (i.e. Null);  74 Residual
Null Deviance:            385.5
Residual Deviance: 195.7         AIC: 332

Call:
glm(formula = Failures ~ System + Operator + Valve + Size + Mode,
    family = "poisson", data = Nuclear, offset = log(Time))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1892  -1.0074  -0.4357   0.3361   5.3138
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.76867    0.81935  -4.600 4.23e-06 ***
System2      0.91556    0.53184   1.721  0.08516 .
System3      1.01881    0.50548   2.016  0.04385 *
System4      1.22309    0.55518   2.203  0.02759 *
System5      0.33292    0.58408   0.570  0.56869
Operator2    0.70437    0.56669   1.243  0.21389
Operator3   -1.19261    0.24851  -4.799 1.59e-06 ***
Operator4   -2.47233    0.47660  -5.187 2.13e-07 ***
Valve2       0.18533    0.76105   0.244  0.80761
Valve3       0.60674    0.78107   0.777  0.43727
Valve4       2.95894    0.60010   4.931 8.19e-07 ***
Valve5       1.79318    0.61040   2.938  0.00331 **
Valve6       1.00891    0.93009   1.085  0.27803
Size2       -0.01219    0.28340  -0.043  0.96568
Size3        1.61457    0.32104   5.029 4.93e-07 ***
Mode2       -0.20934    0.19033  -1.100  0.27138
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 385.53  on 89  degrees of freedom
Residual deviance: 195.68  on 74  degrees of freedom
AIC: 332.02

Number of Fisher Scoring iterations: 7
```

So the model is:

$$\begin{aligned} log(Failures) = -3.77 &+ 0.92System2 + 1.02System3 + 1.22System4 + 0.33System5 \\ &+ 0.70Operator2 - 1.19Operator3 - 2.47Operator4 \\ &+ 0.19Valve2 + 0.61Valve3 + 2.96Valve4 + 1.79Valve5 + 1.01Valve6 \\ &- 0.01Size2 + 1.1Size3 \\ &- 0.21Mode2 \end{aligned}$$

And since there are several coefficients not significant, this means that these corresponding variables are not associated with Failures.

(b). Interpret the estimated parameters.

There are total 16 parameters here in the model. β_0 as the intercept, i.e., when System, Operator, Valve, Size and Mode all equal 1, the $= e^{-3.77}$. $\beta_1 - \beta_4$, as the coefficients of 4 different System levels 2-5, respectively. $\beta_5 - \beta_7$, as the coefficients of 3 different Operator levels 2-4, respectively. $\beta_8 - \beta_{12}$, as the coefficients of 5 different Valve levels 2-6, respectively. $\beta_{13} - \beta_{14}$, as the coefficients of 5 different levels Size 2-3, respectively and β_{15}, as the coefficient of Mode level 2.

- System

When other explanatory variables are fixed, then System 2 will cause $e^{0.92} = 2.509$ times as many Failures as System 1, and similarly System 3 will cause $e^{1.02} = 2.773$ times as many Failures as System 1, System 4 will cause $e^{1.22} = 3.387$ times as many Failures as System 1. System 5 will cause $e^{0.33} = 1.391$ times as many Failures as System 1. And since under 0.05 significance level, the coefficients of System 2 and System 5 are not significant, then this means that System 2 and System 5 are not different from System 1 in the sense of Large number of Failures.

- Operator

When other explanatory variables are fixed, then Operator 2 will cause $e^{0.70} = 2.014$ times as many Failures as Operator 1, and similarly Operator 3 will cause $e^{-1.19} = 0.304$ times as many Failures as System 1, Operator

4 will cause $e^{-2.47} = 0.085$ times as many Failures as Operator 1. And since under 0.05 significance level, the coefficient of Operator 2 is not significant, then this means that Operator 2 is not different from System 1 in the sense of Large number of Failures.

- Valve

When other explanatory variables are fixed, then Valve 2 will cause $e^{0.19} = 1.20$ times as many Failures as Valve 1, and similarly Valve 3 will cause $e^{0.61} = 1.84$ times as many Failures as Valve 1, Valve 4 will cause $e^{2.96} = 19.298$ times as many Failures as Valve 1, Valve 5 will cause $e^{1.79} = 5.990$ times as many Failures as Valve 1, Valve 6 will cause $e^{1.01} = 2.746$ times as many Failures as Valve 1. And since under 0.05 significance level, the coefficients of Valve 2, Valve 3 and Valve 6 are not significant, then this means that Valve 2, Valve 3 and Valve 6 are not different from System 1 in the sense of Large number of Failures.

- Size

When other explanatory variables are fixed, then Size 2 will cause $e^{-0.01} = 0.990$ times as many Failures as Size 1, and similarly Size 3 will cause $e^{1.61} = 5.002$ times as many Failures as Size 1. And since under 0.05 significance level, the coefficient of Size 2 is not significant, then this means that Size 2 is not different from Size 1 in the sense of Large number of Failures.

- Mode

When other explanatory variables are fixed, then Mode 2 will cause $e^{-0.20} = 0.819$ times as many Failures as Mode 1. And since under 0.05 significance level, the coefficient of Mode 2 is not significant, then this means that Mode 2 is not different from Mode 1 in the sense of Large number of Failures.

(c). Assess the goodness of fit of the model

Goodness of fit test Result

```
Goodness of Fit Test

res.deviance   df   p.value
195.67809951   74   0.00000000
```

From result of Deviance Goodness of Fit Test, though most of the parameters' coefficients are significant, however, the p-value is approximate 0, which means this model doesn't fit well, and is not adequate. The lack of fit maybe due to missing data, covariates or overdispersion.

And the Chi-square Anova Table of log-linear regression result is as follows:

Log-Linear ANOVA Table

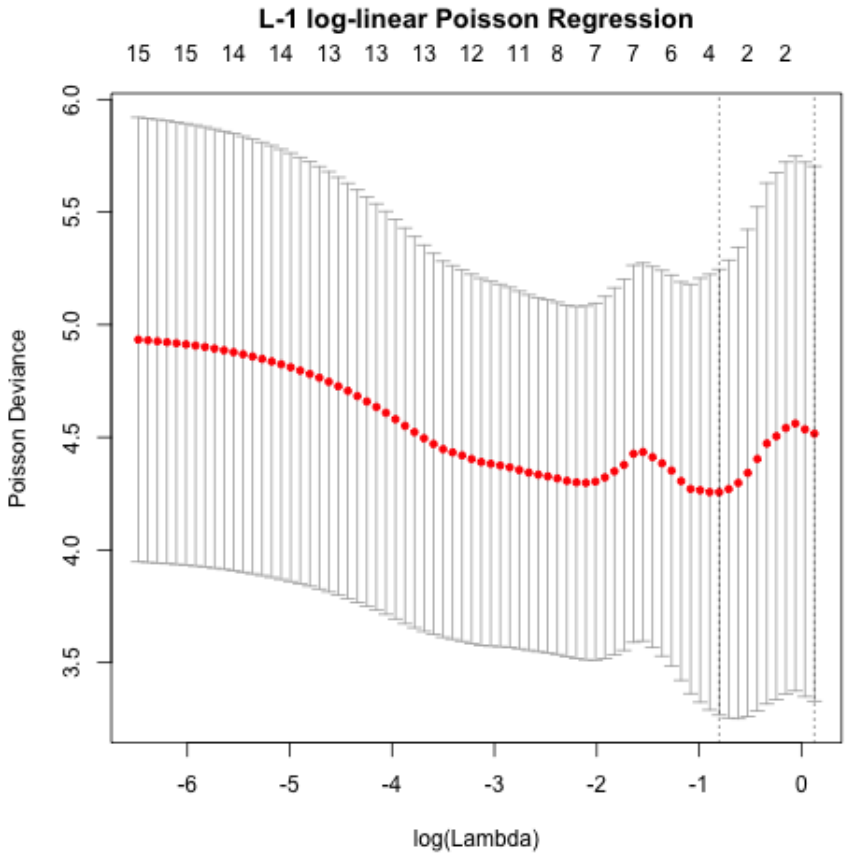
```
Analysis of Deviance Table

Model: poisson, link: log

Response: Failures

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      89     385.53
System    4   22.704        85     362.83 0.0001451 ***
Operator  3    5.335        82     357.49 0.1488176
Valve     5  109.857        77     247.63 < 2.2e-16 ***
Size      2   50.742        75     196.89 9.584e-12 ***
Mode      1    1.213        74     195.68 0.2708352
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Since under 0.05 significance level, the Operator and Mode of explanatory variables are not significant. Then we can conclude that they are not strongly associated with Failures.

Problem 3

Repeat 2(a) using the glmnet package and comment on the results.

Use cross validation and log linear poisson regression on the data.
The following plot shows the best lambda:



So the best lambda is 0.1644754 and the corresponding coefficients are:

Coefficients of log-linear poisson regression with lasso

```
16 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -1.4248815
System2      .
System3      .
System4      .
System5      .
Operator2    .
Operator3    .
Operator4    .
Valve2       .
Valve3       .
Valve4       0.7310400
Valve5       .
Valve6       .
Size2        .
Size3        0.5531311
Mode2        .
```

And the model is:

$$log(Failures) = -1.4249 + 0.7310Valve4 + 0.5531Size3$$

Compare the results:

Since from the two models we can find that, only Valve 4 and Size 3 are included in the model. Then this means only Valve 4 and Size 3 have significant effect on Failures. And compare the two models, we can find that these two models are quite different. And using Lasso with cross validation really depends on the seed because it depends on the training dataset. And lasso will reduce many variables, whereas, regular log-linear regression won't.

R Code:

```
rm(list=ls())
library(Sleuth2)
data(ex2224)
Nuclear<-ex2224
levels(Nuclear$System)<-c("containment","nuclear","power conversion","safety","process auxiliary")
levels(Nuclear$Operator)<-c("air","solenoid","motor-driven","manual")
levels(Nuclear$Valve)<-c("ball","butterfly","diaphragm","gate","globe","directional control")
levels(Nuclear$Size)<-c("<2","2-10","10-30")
levels(Nuclear$Mode)<-c("closed","open")
for (names in names(Nuclear)){
  if (class(Nuclear[[names]])!="numeric"){
    Nuclear[[names]]<-as.factor(as.numeric(Nuclear[[names]]))
  }
}

#loglinear poisson regression
glmpoisson<-glm(Failures~System+Operator+Valve+Size+Mode,offset=log(Time),data=Nuclear,family="poisson")
summary(glmpoisson)

sink('/Users/raymond/Drive/STAT W4201/HW8/glmpoisson.txt')
glmpoisson
summary(glmpoisson)
sink()

# goodness of fit test
Goodness_of_Fit<-cbind(res.deviance = sprintf("%.8f",glmpoisson$deviance),df = glmpoisson$df.residual, p.value = sprintf(
        "%.8f",(1-pchisq(glmpoisson$deviance,glmpoisson$df.residual))))

sink("/Users/raymond/Drive/STAT W4201/HW8/glmgoodness.txt")
cat("Goodness of Fit Test \n \n")
write.table(Goodness_of_Fit,row.names = FALSE,quote = FALSE,sep="    ")
sink()

sink("/Users/raymond/Drive/STAT W4201/HW8/glmchisq.txt")
anova(glmpoisson,test="Chisq")
sink()

#lasso log linear poisson regression
#data matrix transformation of dummy variables
library(glmnet)

NuclearData<-subset(Nuclear,select=c("Failures","Time"))
for (names in names(Nuclear)[1:5]){
  for (factor in levels(Nuclear[[names]])[-1]){
    NuclearData[[paste(names,factor,sep="")]]<-as.numeric(Nuclear[[names]]==factor)
  }
}
NuclearMat<-as.matrix(NuclearData[,-c(1,2)])
set.seed(123)
glmpoisson_l1<-glmnet(NuclearMat,NuclearData$Failures,offset=log(NuclearData$Time),family="poisson")
glmpoisson_l1_cv<-cv.glmnet(NuclearMat,NuclearData$Failures,offset=log(NuclearData$Time),family="poisson")

#plot the best lambda
png(filename = "/Users/raymond/Drive/STAT W4201/HW8/glmpoisson_l1_cv.png")
plot(glmpoisson_l1_cv)
title(main = "L-1 log-linear Poisson Regression", line = 2.5)
dev.off()

#get model coefficients
lambda<-glmpoisson_l1_cv$lambda.min
model<-glmpoisson_l1_cv$glmnet.fit
coeff<-coef(model,lambda)
sink('/Users/raymond/Drive/STAT W4201/HW8/glmpoisson_l1_cv.txt')
coeff
sink()
```