HW7

Qiuying Li UNI ql2280

11/9/2017

# Problem 2 (a)

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)

##   
## 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:

**log (Failures) = −3.77 + 0.92*System2 + 1.02*System3 + 1.22*System4 + 0.33*System5 + 0.70*Operator2 − 1.19*Operator3 − 2.47\**Operator4 + 0.19\**Valve2 + 0.61\**Valve3 + 2.9\*6*Valve4 + 1.79*Valve5 + 1.01*Valve6 *− 0.01\*Size2 + 1.1\*Size3 − 0.21\*Mode2***

# (b). Interpret the estimated parameters.

There are total 16 parameters here in the model. There are four β for system variables, three β for operator variables, and five β for values, and two β for size variables, and one β for mode2

* β0 as the intercept, which means when System, Operator, Valve, Size and Mode all equal 1, the failure =e ^ (−3.77).

**Beta for variable “system”**

* Beta for system2: When other explanatory variables are fixed, then System 2 will cause e^ (0.92) = 2.509 times as many Failures as System 1
* Beta for system 3: System 3 will cause e^(1.02) = 2.773 times as many Failures as System 1
* Beta for system 4: System 4 will cause e ^(1.22) = 3.387 times as many Failures as System 1
* Beta for system 5: System 5 will cause e^(0.33) = 1.391 times as many Failures as System 1

**P -Values for variable “System”**

* when a = 5%, the p –values for System 2 and System 5 are not significant, which means that System 2 and System 5 are not different from System 1 in the sense of Large number of Failures

**Beta for variable “Operator”**

* Beta for Operator 2: When other explanatory variables are fixed, Operator 2 will cause e^(0.70) = 2.014 times as many Failures as Operator 1
* Beta for Operator 3: Operator 3 will cause e^( −1.19) = 0.304 times as many Failures as Operator 1
* Beta for Operator 4:Operator 4 will cause e^( −2.47) = 0.085 times as many Failures as Operator 1

**P-value for Operator**

* When a = 5%, p-value of Operator 2 > 0.05, which is not significant , we can not reject H0, and it means that Operator 2 is not different from System 1 in the sense of Large number of Failures.

**Beta for variable “Valve”**

* Beta for Valve 2: When other explanatory variables are fixed, then Valve 2 will cause e ^ (0.19) = 1.20 times as many Failures as Valve 1
* Beta for Valve 3: Valve 3 will cause e ^ (0.61) = 1.84 times as many Failures as Valve 1
* Beta for Valve 4: Valve 4 will cause e ^ (2.96) = 19.298 ties as many Failures as Valve 1
* Beta for Valve 5: Valve 5 will cause e ^（1.79 ）= 5.990 times as many Failures as Valve 1
* Beta for Valve 6: Valve 6 will cause e ^ (1.01) = 2.746 times as many Failures as Valve 1

**P-value for variables “Value”**

* Under a = 5, the p-value of Valve 2, Valve 3 and Valve 6 are greater than 0.05, which 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

**Beta for variable “Size”**

* Beta for Size 2: When other explanatory variables are fixed, then Size 2 will cause e ^ ( −0.01) = 0.990 times as many Failures as Size 1
* Beta for Size 3: Size 3 will cause e ^ (1.61) = 5.002 times as many Failures as Size 1

**P-value for variable “size”**

* When a =0.05, the p-value of size 2 is greater than 0.05, thus 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.

**Beta for variable “Mode”**

* When other explanatory variables are fixed, then Mode 2 will cause e^ (−0.20) = 0.819 times as many Failures as Mode 1.

**P-value for variable “Mode”**

* When a =0.05, the p-value for Mode 2 is greater than 0.05, which 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

library (glmnet)

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-5

with(glmpoisson, cbind(res.deviance = deviance, df = df.residual,p = pchisq(deviance, df.residual, lower.tail=FALSE)))

## res.deviance df p  
## [1,] 195.6781 74 6.198912e-13

***# According to Deviance Goodness of Fit Test, since the p-value is less than 0.05, we s hould reject the null hypothesis, and conclude that the model doesn't fit well.* anova(poisson.reg,test="Chisq")**

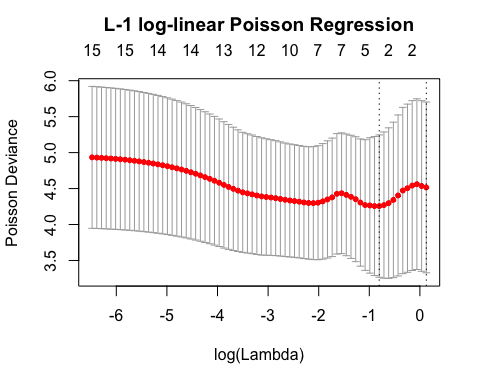
anova(glmpoisson,test="Chisq")

## 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

***# The Anova table shows that Operator and Mode are not significant, so they don't have strong correlation with Failures***

# Problem 3

plot(glmpoisson\_l1\_cv)  
title (main = "L-1 log-linear Poisson Regression", line = 2.5)



lambda = glmpoisson\_l1\_cv$lambda.min;lambda

## [1] 0.4477339

coeff = **coef**(model,lambda);coeff

## 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

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.