Stat 4201 HW8

Jiahong Hu jh3561

November 6, 2015

Problem 2.a Solution:

$$log(\frac{\mu}{t}) = \alpha + \beta_2 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

we set the system,operator,valve,size and mode as dummy variables and let μ be the expected mean of count of failures.

$log(\frac{\mu}{t}) = -3.76867 + 0.91556 System_2 + 1.01881 System_3 + 1.22309 System_4 + 0.33292 System_5 + 0.70437 Operator_2 - 1.19261 Operator_3 - 2.47233 Operator_4 + 0.18533 Valve_2 + 0.60674 Valve_3 + 2.95894 Valve_4 + 1.79318 Valve_5 + 1.00891 Valve_6 - 0.01219 Size_2 + 1.61457 Size_3 - 0.20934 Mode_2$

$log(\mu) = -3.76867 + 0.91556 System_2 + 1.01881 System_3 + 1.22309 System_4 + 0.33292 System_5 + 0.70437 Operator_2 - 1.19261 Operator_3 - 2.47233 Operator_4 + 0.18533 Valve_2 + 0.60674 Valve_3 + 2.95894 Valve_4 + 1.79318 Valve_5 + 1.00891 Valve_6 - 0.01219 Size_2 + 1.61457 Size_3 - 0.20934 Mode_2 + log(time)$

$\mu = time * exp(-3.76867 + 0.91556 System_2 + 1.01881 System_3 + 1.22309 System_4 + 0.33292 System_5 + 0.70437 Operator_2 - 1.19261 Operator_3 - 2.47233 Operator_4 + 0.18533 Valve_2 + 0.60674 Valve_3 + 2.95894 Valve_4 + 1.79318 Valve_5 + 1.00891 Valve_6 - 0.01219 Size_2 + 1.61457 Size_3 - 0.20934 Mode_2)$

Listing 1: Problem 2.a

```
> summary(fit1)

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

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

>
```

Porblem 2.b Solution:

Interpretation of coefficients

1) When all the variables in the function above are equal to zero, that is, system,operator,valve,size,mode are all in category 1, the expected mean of counts of failures during time interval t is t*exp(-3.76867)

2) The difference of the expected mean counts of failures during time interval t between observations in system category 2 and system category 1 is t*exp(0.91556), with all the other variables equal to 0, that is, all in category 1

3) The difference of the expected mean counts of failures during time interval t between observations in system category 3 and system category 1 is t*exp(1.01881), with all the other variables equal to 0, that is, all in category 1

4) The difference of the expected mean counts of failures during time interval t between observations in system category 4 and system category 1 is t*exp(1.22309), with all the other variables equal to 0, that is, all in category 1

5) The difference of the expected mean counts of failures during time interval t between observations in system category 5 and system category 1 is t*exp(0.33292), with all the other variables equal to 0, that is, all in category 1

6) The difference of the expected mean counts of failurese during time interval t between observations in operator 2 and operator 1 is t*exp(0.70437), with all the other variables equal to 0, that is, all in category 1

7) The difference of the expected mean counts of failures during time interval t between observations in operator 3 and operator 1 is t*exp(-1.19261), with all the other variables equal to 0, that is, all in category 1

8) The difference of the expected mean counts of failures during time interval t between observations in operator 4 and operator 1 is t*exp(-2.47233), with all the other variables equal to 0, that is, all in category 1

9) The difference of the expected mean counts of failures during time interval t between observations in valve 2 and valve 1 is t*exp(0.18533), with all the other variables equal to 0, that is, all in category 1

10) The difference of the expected mean counts of failures during time interval t between observations in valve 3 and valve 1 is t*exp(0.60674), with all the other variables equal to 0, that is, all in category 1

11) The difference of the expected mean counts of failures during time interval t between observations in valve 4 and valve 1 is t*exp(2.95894), with all the other variables equal to 0, that is, all in category 1

12) The difference of the expected mean counts of failures during time interval t between observations in valve 5 and valve 1 is t*exp(1.79318), with all the other variables equal to 0, that is, all in category 1

13) The difference of the expected mean counts of failures during time interval t between observations in valve 5 and valve 1 is t*exp(1.00891), with all the other variables equal to 0, that is, all in category 1

14) The difference of the expected mean counts of failures during time interval t between observations in valve 6 and valve 1 is t*exp(1.00891), with all the other variables equal to 0, that is, all in category 1

15) The difference of the expected mean counts of failures during time interval t between observations in Size 2 and Size 1 is t*exp(-0.01219), with all the other variables equal to 0, that is, all in category 1

16) The difference of the expected mean counts of failures during time interval t between observations in Size 3 and Size 1 is t*exp(1.61457), with all the other variables equal to 0, that is, all in category 1

17) The difference of the expected mean counts of failures during time interval t between observations in Mode 2 and Mode1 is t*exp(-0.20934), with all the other variables equal to 0, that is, all in category 1

Problem 2.c Solutions:

Check the fitness of model

H_0: The model is exactly correct H_1: The model is not exactly correct

1)Deviance godness of fitness test

Listing 2: Problem 2.b

```
> pchisq(fit1$deviance, df=fit1$df.residual, lower.tail=FALSE)
[1] 6.198912e-13
```

Under the full model, the p-value = $6.198912exp(-13)$, which is less than 0.05. Hence, we reject the Null Hypothesis and the model is not exactly fit.

2) anova test

we can see that the operator and model explantory variables are not significant, which p value larger than 0.05

Listing 3: Problem 2.c

```
> anova(fit1,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
>
```

3) Now, let's try to do the model selection using AIC backwards method

Listing 4: Problem 2.c

```
> step(fit1, direction="backward", trace=TRUE)
Start:  AIC=332.02
Failures ~ System + Operator + Valve + Size + Mode

           Df Deviance    AIC
- Mode      1   196.89 331.24
<none>          195.68 332.02
- System    4   207.76 336.11
- Size      2   246.17 378.51
- Operator  3   253.95 384.30
- Valve     5   299.19 425.54

Step:  AIC=331.24
Failures ~ System + Operator + Valve + Size

           Df Deviance    AIC
<none>          196.89 331.24
- System    4   209.13 335.47
- Size      2   247.63 377.98
- Operator  3   256.25 384.60
- Valve     5   300.23 424.57

Call:  glm(formula = Failures ~ System + Operator + Valve + Size, family = poisson(link
     = log),
    data = data, offset = ltime)

Coefficients:
(Intercept)      System2      System3      System4      System5    Operator2
  -3.765773     0.889192     0.971160     1.130913     0.231690     0.675207
  Operator3    Operator4       Valve2       Valve3       Valve4       Valve5
  -1.158327    -2.504533     0.271693     0.567502     2.915529     1.713994
     Valve6        Size2        Size3
   0.928622    -0.002418     1.522295

Degrees of Freedom: 89 Total (i.e. Null);  75 Residual
Null Deviance:      385.5
Residual Deviance: 196.9        AIC: 331.2
>
```

We can conclude the best model under the AIC backwards selection does not include the variale Mode. However,we notice that the change of coefficients is not very remarkable.

$log(\frac{\mu}{t}) = -3.765773 + 0.889192System_2 + 0.971160System_3 + 1.130913System_4 + 0.231690System_5 + 0.675207Operator_2 - 1.158327Operator_3 - 2.504533Operator_4 + 0.271693Valve_2 + 0.567502Valve_3 + 2.915529Valve_4 + 1.713994Valve_5 + 0.92862Valve_6 - 0.002418Size_2 + 1.5222957Size_3$

Listing 5: Problem 2.c

```
> summary(fit2)

Call:
glm(formula = Failures ~ System + Operator + Valve + Size, family = poisson(link = log)
    ,
    data = data, offset = ltime)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9962  -1.0518  -0.4519   0.4316   4.8857

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.765773   0.819535  -4.595 4.33e-06 ***
System2      0.889192   0.532845   1.669  0.09516 .
System3      0.971160   0.503572   1.929  0.05379 .
System4      1.130913   0.547659   2.065  0.03892 *
System5      0.231690   0.575447   0.403  0.68722
Operator2    0.675207   0.567052   1.191  0.23376
Operator3   -1.158327   0.244506  -4.737 2.16e-06 ***
Operator4   -2.504533   0.474995  -5.273 1.34e-07 ***
Valve2       0.271693   0.756012   0.359  0.71931
Valve3       0.567502   0.782345   0.725  0.46822
Valve4       2.915529   0.598550   4.871 1.11e-06 ***
Valve5       1.713994   0.606591   2.826  0.00472 **
Valve6       0.928622   0.929165   0.999  0.31759
Size2       -0.002418   0.282508  -0.009  0.99317
Size3        1.522295   0.307493   4.951 7.40e-07 ***
---
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: 196.89  on 75  degrees of freedom
AIC: 331.24

Number of Fisher Scoring iterations: 7
```

Problem 3 Solution:

The model under the glmnet method is

$log(\frac{\mu}{t}) = -1.424761 + 0.7307952Valve_4 + 0.5531644Size_3$

$\mu = time * exp(-1.424761 + 0.7307952Valve_4 + 0.5531644Size_3)$

comments:

the model has fewer variabes than the model produced under GLM

when valve has category other than 4 and size has category other than 3, the expected mean of counts of failure over time interval t is t*exp(-1.424761).

when the valve has category 4, the difference of the expected mean of failure over time interval t between valve having category 4 and categoty other than 4, with size having category other than 3, is t*exp(0.7307952)

when the size has category 3, the difference of the expected mean of failure over time interval t between size having category 3 and category other than 3, with valve having category other than 4, is t*exp(0.5531644)

Listing 6: Problem 3.a

```
> fit4$a0
       s0
-1.424761
> fit4$beta
15 x 1 sparse Matrix of class "dgCMatrix"
                      s0
data$System2   .
data$System3   .
data$System4   .
data$System5   .
data$Operator2 .
data$Operator3 .
data$Operator4 .
data$Valve2    .
data$Valve3    .
data$Valve4    0.7307952
data$Valve5    .
data$Valve6    .
data$Size2     .
data$Size3     0.5531644
data$Mode2     .
>
```

Code Solution:

Listing 7: Code

```
library("Sleuth3")
data<-ex2224
ltime=log(data$Time)
data$System=as.factor(data$System)
data$Operator=as.factor(data$Operator)
data$Valve=as.factor(data$Valve)
data$Size=as.factor(data$Size)
data$Mode=as.factor(data$Mode)

fit1<-glm(Failures~System+Operator+Valve+Size+Mode,offset=ltime,family=poisson(link=log
    ),data=data)
summary(fit1)
anova(fit1,test="Chisq")

pchisq(fit1$deviance, df=fit1$df.residual, lower.tail=FALSE)
null<-glm(Failures~1,offset=ltime,family=poisson(link=log),data=data)
step(fit1, direction="backward", trace=TRUE)

fit2<-glm(Failures~System+Operator+Valve+Size,offset=ltime,family=poisson(link=log),
    data=data)
fit3<-glm(Failures~System+Valve+Size,offset=ltime,family=poisson(link=log),data=data)

x=model.matrix(data$Failures~data$System+data$Operator+data$Valve+data$Size+data$Mode)
    [,-1]
y=as.vector(data$Failures)
set.seed=1
cv<-cv.glmnet(x,y,family="poisson",offset=ltime)
cv$lambda.min
fit4<-glmnet(x,y,family="poisson",offset=ltime,lambda=0.4477339)
fit4$a0
fit4$beta
```