HW8

Hanao Li hl3202

10/29/2019

Problem 1

a)

```
if(!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
p_load(Sleuth2, glmnet)

data <- ex2224
data$System <- as.factor(as.numeric(data$System))
data$Operator <- as.factor(as.numeric(data$Operator))
data$Valve <- as.factor(as.numeric(data$Valve))
data$Size <- as.factor(as.numeric(data$Size))
data$Mode <- as.factor(as.numeric(data$Mode))
Poisson_fit <- glm(Failures~Operator, data = data, family = "poisson")
summary(Poisson_fit)
```

```
##
## Call:
## glm(formula = Failures ~ Operator, family = "poisson", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0953  -1.9954  -0.9043   0.4427   8.1521
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.7862     0.1054   7.459 8.72e-14 ***
## Operator2    -0.7862     0.5110  -1.539  0.12389
## Operator3    -0.4233     0.1812  -2.336  0.01951 *
## Operator4    -1.7417     0.4595  -3.791  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 413.76  on 89  degrees of freedom
## Residual deviance: 386.71  on 86  degrees of freedom
## AIC: 499.05
##
## Number of Fisher Scoring iterations: 6
```

```
anova(Poisson_fit, 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     413.76
## Operator  3   27.047        86     386.71 5.755e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the result, we could see that our p-value is less than 0.05, we reject the null and conclude that there is an association between valve failure and operator.

b)

```
Log_fit <- glm(Failures~System+Operator+Valve+Size+Mode, data = data, family = "poisson", offset
= log(Time))
summary(Log_fit)
```

```
##
## Call:
## glm(formula = Failures ~ System + Operator + Valve + Size + Mode,
##     family = "poisson", data = data, 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
```

Our model is log(failures) = -3.76867+0.91556System2+1.01881System3+1.22309System4+0.33292System5+0.70437Operator2-1.19261Operator3-2.47233Operator4+0.18533Valve2+0.60674Valve3+2.95894Valve4+1.79318Valve5+1.00891Valve6-0.01219Size2+1.61457Size3-0.20934Mode2

```
anova(Log_fit,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
```

From the result, we could see that our p value for opeartor is larger than 0.05, we do not reject the null and we concluded that Operator and Mode does not have association with Failures. Instead, we can conclude that System, Valve and Size have association with Failures.

Problem 2

1a)

Operator 2 will reduce $0.54 * (1 - e^{-0.7862})$ failures than Operator 1. Operator 3 will reduce $0.34 * (1 - e^{-0.4233})$ failures than Operator 1. Operator 4 will increase $0.82 * (1 - e^{-1.7417})$ failures than Operator 1. From the result above, only Operator 3 and Operator 4 have p-values less than 0.5 and it means Operator 3 and Operator 4 are significant and they can affect the failures relative to Operator 1.

1b)

For system variable, When other variables are fixed, System 2 will increase $1.5 * (e^{0.91556} - 1)$ times of failures than System 1. System 3 will increase $1.76 * (e^{1.01881} - 1)$ times of failures than System 1. System 4 will increase $2.40 * (e^{1.22309} - 1)$ times of failures than System 1. System 5 will increase $0.4 * (e^{0.33292} - 1)$ times of failures than System 1. Based on the result above, only System 3 and System 4 are significant variables which means we can conclude that these two variables can effect the failures relative to System 1. For operator variable, When other variables are fixed, Operator 2 will increase $1.02 * (e^{0.70437} - 1)$ times of failures than Operator 1. Operator 3 will reduce $0.7 * (1 - e^{1.19261})$ failures than Oeprator 1. Operator 4 will increase $0.92 * (1 - e^{-2.47233})$ failures than Operator 1.Based on the result above, only Operator 3 and Operator 4 are significant which means we can conclude that these two variables can effect the failures relative to Operator 1. For value variable, When other variables are fixed, Valve 2 will increase $0.203 * (e^{0.1853} - 1)$ times of failures than Valve 1. Valve 3 will increase $0.834 * (e^{0.60674} - 1)$ times of failures than Valve 1. Valve 4 will increase $18 * (e^{2.95894} - 1)$ failures than Valve 1. Valve 5 will increase $5 * (e^{1.79318} - 1)$ times of failures than Valve 1. Valve 6 will increase $1.7426 * (e^{1.00891} - 1)$ times of failures than Valve 1. Based on the result above, only Valve 4 and Valve 5 are significant which means we can conclude that these two variables can effect the failures relative to Valve 1. For size variable, When other variables are fixed, Size 2 will reduce $0.0121 * (1 - e^{-0.01219})$ failures than Size 1. Size 3 will increase $e^{1.61457} - 1$ failures than Size 1. Based on the result above, only Size 3 is significant which means we can conclude that these two variables can effect the failures relative to Size 1. For mode variable, When other variables are fixed, Mode 2 will reduce $0.1889 * (1 - e^{-0.209})$ failures than Size 1. Size 3 will increase $e^{-0.209} - 1$ failures than Mode 1. Based on the result above, there is no significant variables which means mode does not effect failures.

b)

1a)

```
chisq.test(data$Failures, p = Poisson_fit$fitted.values, rescale.p = TRUE)
```

```
## Warning in chisq.test(data$Failures, p = Poisson_fit$fitted.values,
## rescale.p = TRUE): Chi-squared approximation may be incorrect
```

```
## 
##  Chi-squared test for given probabilities
## 
## data:  data$Failures
## X-squared = 647.44, df = 89, p-value < 2.2e-16
```

From the result above, the p-value of goodness of fit test is very small. Therefore, we reject the null and it can be concluded that we reject that the data comes from a specified distribution.

1b)

```
chisq.test(data$Failures, p = Log_fit$fitted.values, rescale.p = T)
```

```
## Warning in chisq.test(data$Failures, p = Log_fit$fitted.values, rescale.p =
## T): Chi-squared approximation may be incorrect
```

```
## 
##  Chi-squared test for given probabilities
## 
## data:  data$Failures
## X-squared = 334.86, df = 89, p-value < 2.2e-16
```

From the result above, the p-value of goodness of fit test is very small. Therefore, we reject the null and it can be concluded that we reject that the data comes from a specified distribution.

Problem 3

```
dat <- subset(data, select=c("Failures", "Time"))
for (names in names(data)[1:5]){
  for (factor in levels(data[[names]])[-1]){
    dat[[paste(names, factor, sep="")]] <- as.numeric(data[[names]] == factor)
  }
}
X <- as.matrix(dat[,-c(1,2)])
cv <- cv.glmnet(X, dat$Failures,family = "poisson", offset = log(dat$Time))
glm_fit <- cv$glmnet.fit
coef(glm_fit,cv$lambda.min)
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept) -2.03051879
## System2      .
## System3      0.17256222
## System4      0.09319092
## System5     -0.32860148
## Operator2    .
## Operator3   -0.83800481
## Operator4   -1.61238681
## Valve2       .
## Valve3       .
## Valve4       1.84969030
## Valve5       0.65481692
## Valve6       .
## Size2        .
## Size3        1.17225666
## Mode2        .
```

Based on the result above, the model would be: log(failure)=-1.46452212-1.46452212System3-0.03407912System5+0.77050393Valve4+0.58879067Size3

This model means only System3, System 5, Valve4 and Size3 are significantly effect failures. Using penalty-based lasso methods can help us reduced the number of variables significant while it might be differ from each other due to the seed.