GR5291 Advanced Data Analysis Problem Set glm 2

Francis Zhang

November 8, 2024

Question

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

a)Using an appropriate Poisson model, determine if there is association between valve failure and operator

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

2.In each of the above, i.e., 1(a) and 1(b):

a)Interpret the estimated parameters

b)Assess the goodness of fit of the model

3.Repeat 1(b) using the glmnet package and comment on the results.

Solution

Question 1

```
library(Sleuth3)

# Load the valve data
valvedata <- ex2224
head(valvedata)
```

a)

```
##   System Operator Valve Size Mode Failures Time
## 1      1        3     4    3    1        2    4
## 2      1        3     4    3    2        2    4
## 3      1        3     5    1    1        1    2
## 4      2        1     2    2    2        0    2
## 5      2        1     3    2    1        0    2
## 6      2        1     3    2    2        0    1
```

```
# Poisson model for association between failures and operator
model_operator <- glm(Failures ~ Operator + offset(log(Time)),
                      family = poisson(link = "log"),
                      data = valvedata)

# Summary of the model
summary(model_operator)
```

```
##
## Call:
## glm(formula = Failures ~ Operator + offset(log(Time)), family = poisson(link = "log"),
##     data = valvedata)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -0.79841    0.16130  -4.950 7.43e-07 ***
## Operator    -0.10595    0.07832  -1.353    0.176    
## ---
## 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: 383.64  on 88  degrees of freedom
## AIC: 491.99
##
## Number of Fisher Scoring iterations: 6
```

In this Poisson model, the estimated coefficient for Operator is -0.10595, with a standard error of 0.07832 and a p-value of 0.176. Since the p-value is greater than the typical significance level of 0.05, we do not have strong evidence to conclude that the Operator variable is significantly associated with valve failures.

```r
# Full Poisson model with all factors
model_full <- glm(Failures ~ System + Operator + Valve + Size + Mode +
                    offset(log(Time)),
                  family = poisson(link = "log"),
                  data = valvedata)

# Summary of the model
summary(model_full)
```

b)

```
##
## Call:
## glm(formula = Failures ~ System + Operator + Valve + Size + Mode +
##     offset(log(Time)), family = poisson(link = "log"), data = valvedata)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -3.98837    0.67673  -5.894 3.78e-09 ***
## System      -0.12868    0.09128  -1.410  0.15860    
## Operator    -0.26470    0.08485  -3.120  0.00181 ** 
## Valve        0.35791    0.07747   4.620 3.83e-06 ***
## Size         1.22147    0.17492   6.983 2.89e-12 ***
## Mode        -0.09291    0.18195  -0.511  0.60959    
## ---
## 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: 309.69  on 84  degrees of freedom
## AIC: 426.03
##
```

```
## Number of Fisher Scoring iterations: 6
```

The Poisson regression results indicate that certain factors are significantly associated with valve failure rates. Specifically, the coefficient for operator is -0.26470 with a p-value of 0.00181, suggesting that certain operator types are associated with a lower failure rate. The coefficient for valve is 0.35791 with a p-value of 3.83e-06, indicating that specific valve types are significantly linked to a higher failure rate. Additionally, the coefficient for size is 1.22147 with a p-value of 2.89e-12, showing that larger valve sizes are significantly associated with an increased failure rate.

Question 2

a)

1(a) In the Poisson regression model for 1(a), where only operator is included as a predictor, the intercept is estimated at -0.79841 with a p-value of 7.43e-07, indicating a statistically significant baseline log failure rate for the reference operator level. This intercept corresponds to an expected failure rate of exp(-0.79841) ≈ 0.45 failures per unit of time. However, the coefficient for operator is -0.10595 with a p-value of 0.176, which is not statistically significant. This suggests that, without adjusting for other factors, the type of operator alone does not have a significant impact on the valve failure rate.

1(b) In the full Poisson regression model for 1(b), which includes system, operator, valve, size, and mode, several factors are significantly associated with valve failure rates. The coefficient for operator is -0.26470 with a p-value of 0.00181, indicating that certain operator types are associated with a lower failure rate when other variables are controlled for. Additionally, valve has a coefficient of 0.35791 with a p-value of 3.83e-06, suggesting that specific valve types are linked to higher failure rates. The size variable has a positive coefficient of 1.22147 with a p-value of 2.89e-12, showing that larger valve sizes significantly increase the failure rate. This analysis highlights the importance of including multiple factors to accurately identify independent associations with valve failures, as controlling for additional factors reveals the significant impact of operator type, valve type, and valve size on failure rates.

```r
# Check goodness of fit for model_operator
deviance(model_operator) / df.residual(model_operator)
```

b)

```
## [1] 4.359589
```

```r
# Check goodness of fit for model_full
deviance(model_full) / df.residual(model_full)
```

```
## [1] 3.686763
```

The goodness-of-fit checks show that both models exhibit overdispersion, as the deviance-to-degrees-of-freedom ratios are greater than 1. Specifically, the ratio for the model containing only the operator variable is 4.359589, and for the full model with all variables, it is 3.686763. Although the full model's ratio is slightly lower than that of the operator-only model, both values are significantly greater than 1, indicating that even with all variables included, the model fails to fully capture the data's variability. So we can try with negative binomial distribution.

```r
# Load the MASS package for negative binomial regression
library(MASS)

# Negative binomial model with only the operator variable
model_operator_nb <- glm.nb(Failures ~ Operator + offset(log(Time)),
                            data = valvedata)
summary(model_operator_nb)
```

```
## 
## Call:
## glm.nb(formula = Failures ~ Operator + offset(log(Time)), data = valvedata, 
##     init.theta = 0.332509562, link = log)
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.3804     0.4412  -0.862    0.389
## Operator     -0.1838     0.1834  -1.002    0.316
## 
## (Dispersion parameter for Negative Binomial(0.3325) family taken to be 1)
## 
##     Null deviance: 74.591  on 89  degrees of freedom
## Residual deviance: 73.696  on 88  degrees of freedom
## AIC: 277.22
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  0.3325 
##           Std. Err.:  0.0828 
## 
##  2 x log-likelihood:  -271.2150
```

```
# Negative binomial model with all variables
model_full_nb <- glm.nb(Failures ~ System + Operator + Valve + Size + Mode +
                          offset(log(Time)), data = valvedata)
summary(model_full_nb)
```

```
## 
## Call:
## glm.nb(formula = Failures ~ System + Operator + Valve + Size + 
##     Mode + offset(log(Time)), data = valvedata, init.theta = 0.4799938384, 
##     link = log)
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)   
## (Intercept)  -2.8734     1.3727  -2.093  0.03632 * 
## System       -0.1742     0.1718  -1.014  0.31072   
## Operator     -0.2987     0.1780  -1.678  0.09325 . 
## Valve         0.5024     0.1745   2.879  0.00399 **
## Size          1.0011     0.3205   3.124  0.00178 **
## Mode         -0.5805     0.4103  -1.415  0.15707   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(0.48) family taken to be 1)
## 
##     Null deviance: 92.486  on 89  degrees of freedom
## Residual deviance: 73.725  on 84  degrees of freedom
## AIC: 269.84
## 
## Number of Fisher Scoring iterations: 1
## 
## 
```

```
##                Theta:  0.480
##            Std. Err.:  0.130
##
##  2 x log-likelihood:  -255.843
```

The Negative Binomial models show a clear improvement in fit over the Poisson models, as evidenced by the substantial reduction in both deviance and AIC values. For the operator-only model, the deviance decreases from 385.53 (AIC = 491.99) in the Poisson model to 73.696 (AIC = 277.22) in the Negative Binomial model. In the full model, the deviance decreases from 309.69 (AIC = 426.03) to 73.725 (AIC = 269.84). This reduction indicates that the Negative Binomial model effectively addresses the overdispersion in the data, providing a more appropriate fit for identifying factors associated with valve failure rates.

Question 3

```r
library(glmnet)
```

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

```
## Loaded glmnet 4.1-8
```

```r
# Create model matrix excluding the intercept for glmnet
X <- model.matrix(Failures ~ System + Operator + Valve + Size + Mode +
                    offset(log(Time)), data = valvedata)[, -1]
y <- valvedata$Failures
offset_log_time <- log(valvedata$Time)

# Fit LASSO model with cross-validation to select lambda
set.seed(123)
cv_model <- cv.glmnet(X, y, family = "poisson", offset = offset_log_time,
                      alpha = 1, nfolds = 5)

# Best lambda
best_lambda <- cv_model$lambda.min
best_lambda
```

```
## [1] 0.1365504
```

```r
# Fit model at best lambda
lasso_model <- glmnet(X, y, family = "poisson", offset = offset_log_time,
                      alpha = 1, lambda = best_lambda)

# Coefficients at best lambda
coef(lasso_model)
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                      s0
## (Intercept) -3.40373584
## System      -0.06356216
## Operator    -0.12716110
## Valve        0.24883956
## Size         0.89369616
## Mode         .
```

The LASSO regression results indicate that, with the optimal lambda of 0.1365504, the model has selected key predictors for valve failure rates while eliminating the non-significant Mode variable. The variables Valve and Size have positive coefficients (0.24883956 and 0.89369616, respectively), suggesting that higher values of these factors are associated with increased failure rates. In contrast, System and Operator have negative coefficients (-0.06356216 and -0.12716110, respectively), indicating a potential reduction in failure rates associated with these factors. This model highlights Valve and Size as primary contributors to higher failure rates.