Michael_Ghattas_WP8

Michael Ghattas

2024-11-01

START:

```r
# Load data
data(NMES1988)
NMES1988 <- NMES1988 %>% mutate(emergency = visits)

# Split the data into training and test sets
set.seed(1)
n <- nrow(NMES1988)
train_index <- sample(1:n, size = 0.7 * n)
train <- NMES1988[train_index, ]
test <- NMES1988[-train_index, ]
```

```r
# Fit a Negative Binomial model for emergency room visits
negbin_model <- glm.nb(emergency ~ ., data = train)
```

```
## Warning in glm.nb(emergency ~ ., data = train): alternation limit reached
```

```r
summary(negbin_model)
```

```
## 
## Call:
## glm.nb(formula = emergency ~ ., data = train, init.theta = 7.139002744, 
##     link = log)
## 
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)    
## (Intercept)       0.153336   0.155980   0.983 0.325582    
## visits            0.118713   0.001455  81.593  < 2e-16 ***
## nvisits          -0.007267   0.001819  -3.996 6.45e-05 ***
## ovisits          -0.017048   0.003643  -4.680 2.87e-06 ***
## novisits         -0.003540   0.003575  -0.990 0.322010    
## hospital          0.026100   0.013565   1.924 0.054335 .  
## healthpoor        0.013888   0.034552   0.402 0.687715    
## healthexcellent -0.077769   0.046865  -1.659 0.097028 .  
## chronic           0.068549   0.008696   7.883 3.21e-15 ***
## adllimited       -0.040592   0.030425  -1.334 0.182152    
## regionnortheast -0.004513   0.031900  -0.141 0.887507    
## regionmidwest    0.015236   0.029693   0.513 0.607880
## regionwest       0.037078   0.032401   1.144 0.252478
## age              0.042532   0.019469   2.185 0.028914 *
## afamyes         -0.071371   0.039415  -1.811 0.070175 .
## gendermale      -0.071940   0.025651  -2.805 0.005039 **
## marriedyes       0.046620   0.026665   1.748 0.080403 .
## school           0.007357   0.003400   2.164 0.030473 *
## income           0.002136   0.004013   0.532 0.594469
## employedyes     -0.060236   0.039821  -1.513 0.130363
## insuranceyes     0.122110   0.034866   3.502 0.000461 ***
## medicaidyes      0.112392   0.046096   2.438 0.014761 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for Negative Binomial(7.139) family taken to be 1)
##
##     Null deviance: 10578  on 3083  degrees of freedom
## Residual deviance:  3152  on 3062  degrees of freedom
## AIC: 13640
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  7.139
##           Std. Err.:  0.328
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -13594.034
```

```
# Display the Negative Binomial model summary
summary(negbin_model)
```

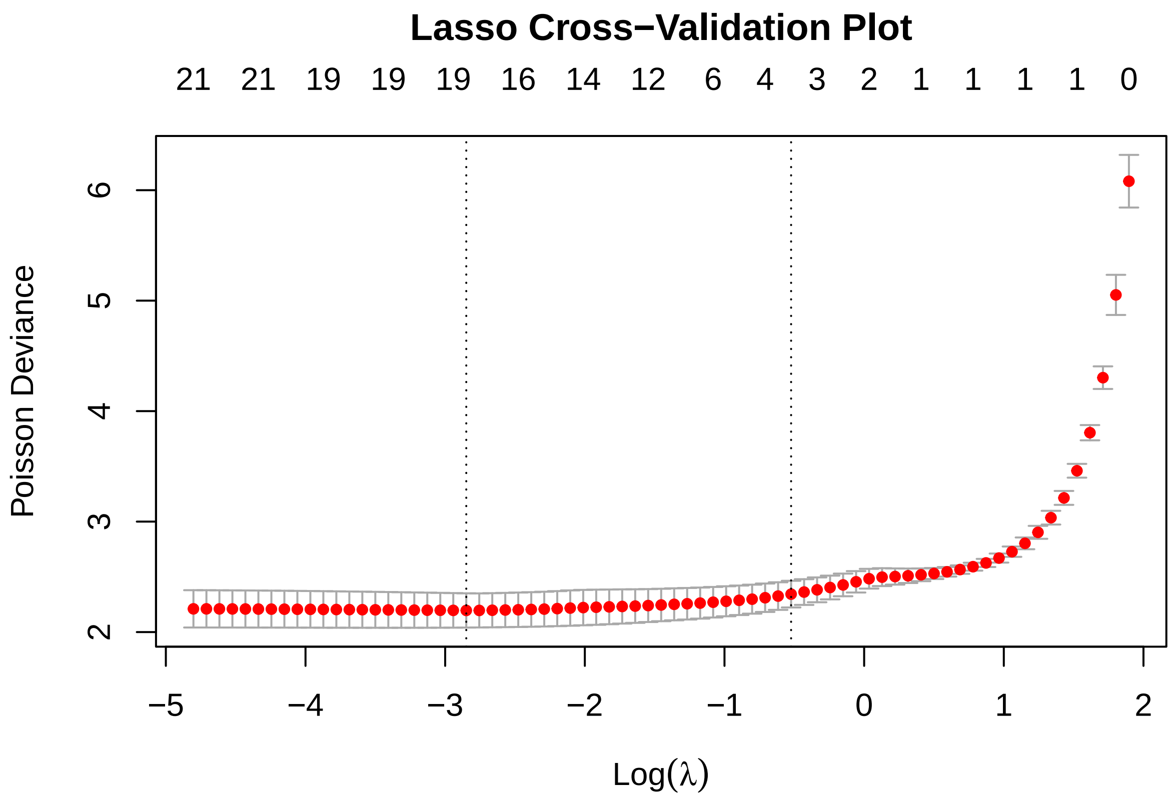
```
##
## Call:
## glm.nb(formula = emergency ~ ., data = train, init.theta = 7.139002744,
##     link = log)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.153336   0.155980   0.983 0.325582
## visits           0.118713   0.001455  81.593  < 2e-16 ***
## nvisits         -0.007267   0.001819  -3.996 6.45e-05 ***
## ovisits         -0.017048   0.003643  -4.680 2.87e-06 ***
## novisits        -0.003540   0.003575  -0.990 0.322010
## hospital         0.026100   0.013565   1.924 0.054335 .
## healthpoor       0.013888   0.034552   0.402 0.687715
## healthexcellent -0.077769   0.046865  -1.659 0.097028 .
## chronic          0.068549   0.008696   7.883 3.21e-15 ***
## adllimited      -0.040592   0.030425  -1.334 0.182152
## regionnortheast -0.004513   0.031900  -0.141 0.887507
## regionmidwest    0.015236   0.029693   0.513 0.607880
## regionwest       0.037078   0.032401   1.144 0.252478
## age              0.042532   0.019469   2.185 0.028914 *
## afamyes         -0.071371   0.039415  -1.811 0.070175 .
## gendermale      -0.071940   0.025651  -2.805 0.005039 **
## marriedyes       0.046620   0.026665   1.748 0.080403 .
## school           0.007357   0.003400   2.164 0.030473 *
## income           0.002136   0.004013   0.532 0.594469
## employedyes     -0.060236   0.039821  -1.513 0.130363
## insuranceyes     0.122110   0.034866   3.502 0.000461 ***
## medicaidyes      0.112392   0.046096   2.438 0.014761 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for Negative Binomial(7.139) family taken to be 1)
##
##     Null deviance: 10578  on 3083  degrees of freedom
## Residual deviance:  3152  on 3062  degrees of freedom
## AIC: 13640
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  7.139
##           Std. Err.:  0.328
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -13594.034
```

```
# Variable selection using Lasso
X_train <- model.matrix(emergency ~ . - 1, data = train)
Y_train <- train$emergency

# Fit Lasso using cross-validation
set.seed(3456787)
lasso_model <- cv.glmnet(X_train, Y_train, alpha = 1, family = "poisson")
lasso_lambda_min <- lasso_model$lambda.min
coef(lasso_model, s = lasso_lambda_min)
```

```
## 23 x 1 sparse Matrix of class "dgCMatrix"
##                           s1
## (Intercept)      0.7166888819
## visits           0.0685240172
## nvisits         -0.0119247488
## ovisits         -0.0076753445
## novisits        -0.0028727488
## hospital         0.0632125890
## healthpoor       0.0716157924
## healthaverage    .
## healthexcellent -0.1181178506
## chronic          0.0856593417
## adllimited      -0.0195439602
## regionnortheast  .
## regionmidwest    0.0036067607
## regionwest       0.0688309764
## age              0.0264302227
## afamyes         -0.0648208189
## gendermale      -0.0938143764
## marriedyes         .
## school             0.0007595192
## income             0.0037325904
## employedyes       -0.0401495383
## insuranceyes       0.1455939432
## medicaidyes        0.1204311011
```

```
# Cross-validation plot for Lasso
plot(lasso_model)
title("Lasso Cross-Validation Plot", line = 2.5)
```

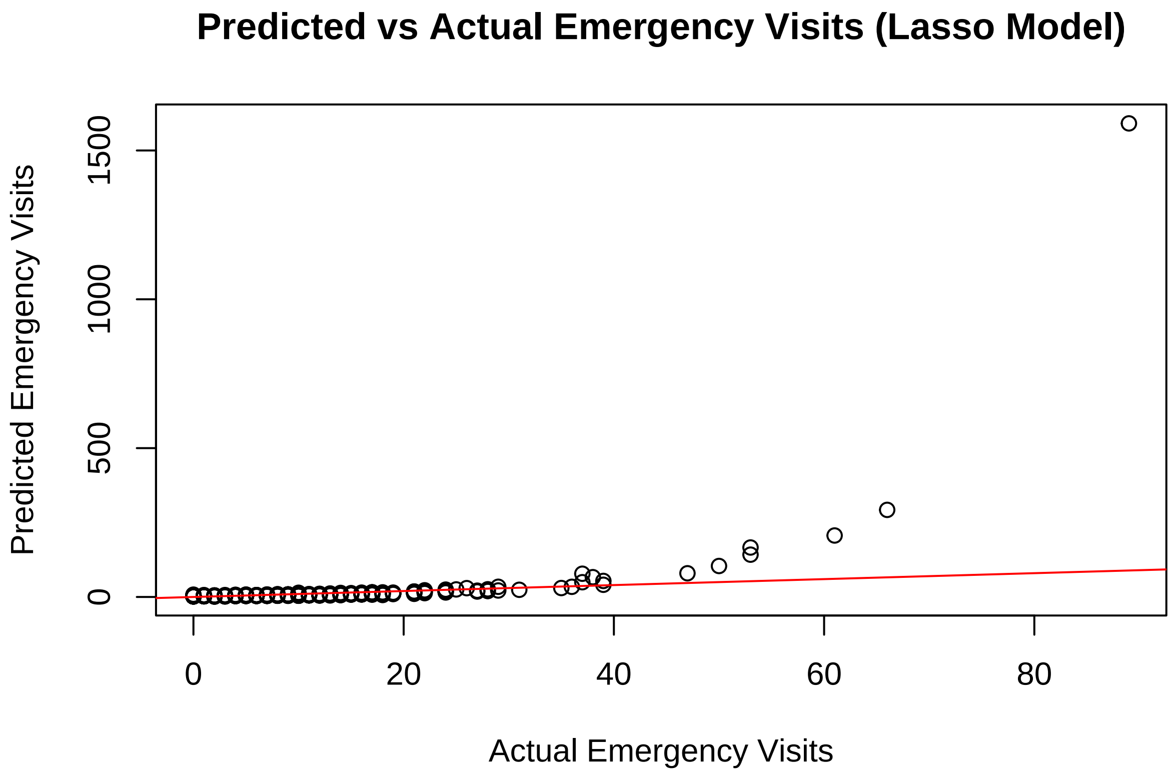


```
# Display selected coefficients for the lambda.min model in Lasso
coef(lasso_model, s = lasso_lambda_min)
```

```
## 23 x 1 sparse Matrix of class "dgCMatrix"
##                             s1
## (Intercept)       0.7166888819
## visits            0.0685240172
## nvisits          -0.0119247488
## ovisits          -0.0076753445
## novisits         -0.0028727488
## hospital          0.0632125890
## healthpoor        0.0716157924
## healthaverage     .
## healthexcellent -0.1181178506
## chronic           0.0856593417
## adllimited       -0.0195439602
## regionnortheast   .
## regionmidwest     0.0036067607
## regionwest        0.0688309764
## age               0.0264302227
## afamyes          -0.0648208189
## gendermale       -0.0938143764
## marriedyes        .
## school            0.0007595192
## income            0.0037325904
## employedyes      -0.0401495383
## insuranceyes      0.1455939432
## medicaidyes       0.1204311011
```

```r
# Evaluation on test data for the final Lasso model
X_test <- model.matrix(emergency ~ . - 1, data = test)
Y_test <- test$emergency
pred_final <- predict(lasso_model, X_test, s = lasso_lambda_min, type = "response")

# Predicted vs Actual Plot for Lasso model
plot(Y_test, pred_final, xlab = "Actual Emergency Visits", ylab = "Predicted Emergency Visits",
     main = "Predicted vs Actual Emergency Visits (Lasso Model)")
abline(0, 1, col = "red")
```



```r
# Calculate log-likelihood on the test data for the selected model
log_likelihood <- sum(Y_test * log(pred_final) - pred_final - log(factorial(Y_test)))
log_likelihood
```

```
## [1] -4725.585
```

Here’s a summary of the key results:

Model Selection and Fitting:

- A Negative Binomial model was fit to the training data to model emergency room visits, as instructed in the exercise.
- Additionally, a Lasso regression with cross-validation was performed as the variable selection method, using a Poisson distribution.

Model Summaries and Selected Coefficients:

- The Negative Binomial model summary indicated significant variables and was displayed in the output.
- The Lasso model identified non-zero coefficients based on the lambda.min value.

Evaluation on Test Data:

- The log-likelihood for the Lasso model on the test data was calculated and found to be approximately -4725.585, which assesses the fit of the selected model.
- A Predicted vs. Actual plot for emergency room visits was generated, showcasing the performance of the Lasso model on test data.

END.