C3M2_peer_reviewed

June 25, 2023

1 C3M2: Peer Reviewed Assignment

1.0.1 Outline:

The objectives for this assignment:

1. Apply Poisson Regression to real data.
2. Learn and practice working with and interpreting Poisson Regression Models.
3. Understand deviance and how to conduct hypothesis tests with Poisson Regression.
4. Recognize when a model shows signs of overdispersion.

General tips:

1. Read the questions carefully to understand what is being asked.
2. This work will be reviewed by another human, so make sure that you are clear and concise in what your explanations and answers.

[6]:

```r
# Load the required packages
library(MASS)
```

2 Problem 1: Poisson Estimators

Let $Y_1, ..., Y_n \overset{i}{\sim} Poisson(\lambda_i)$. Show that, if $\eta_i = \beta_0$, then the maximum likelihood estimator of λ_i is $\widehat{\lambda}_i = \bar{Y}$, for all $i = 1, ..., n$.

3 Problem 2: Ships data

The ships dataset gives the number of damage incidents and aggregate months of service for different types of ships broken down by year of construction and period of operation.

The code below splits the data into a training set (80% of the data) and a test set (the remaining 20%).

[7]:

```r
data(ships)
ships = ships[ships$service != 0,]
ships$year = as.factor(ships$year)
ships$period = as.factor(ships$period)
```

```
set.seed(11)
n = floor(0.8 * nrow(ships))
index = sample(seq_len(nrow(ships)), size = n)

train = ships[index, ]
test = ships[-index, ]
head(train)
summary(train)
```

A data.frame: 6 × 5

	type <fct>	year <fct>	period <fct>	service <int>	incidents <int>
40	E	75	75	542	1
28	D	65	75	192	0
18	C	60	75	552	1
19	C	65	60	781	0
5	A	70	60	1512	6
32	D	75	75	2051	4

```
type  year   period     service          incidents
A:5   60:7   60:11   Min.   :   45.0   Min.   : 0.00
B:5   65:8   75:16   1st Qu.:  318.5   1st Qu.: 0.50
C:6   70:8           Median : 1095.0   Median : 2.00
D:7   75:4           Mean   : 5012.2   Mean   :10.63
E:4                  3rd Qu.: 2202.5   3rd Qu.:11.50
                     Max.   :44882.0   Max.   :58.00
```

3.0.1 2. (a) Poisson Regression Fitting

Use the training set to develop an appropriate regression model for `incidents`, using `type`, `period`, and `year` as predictors (HINT: is this a count model or a rate model?).

Calculate the mean squared prediction error (MSPE) for the test set. Display your results.

[9]:

```
# Your Code Here

# Fit the Poisson regression model
model <- glm(incidents ~ type + period + year, data = train, family = poisson)

# Calculate the mean squared prediction error (MSPE) for the test set
test$predicted <- predict(model, newdata = test, type = "response")
test$MSPE <- (test$predicted - test$incidents)^2
mspe <- mean(test$MSPE)

# Display the results
print(paste("Mean Squared Prediction Error (MSPE):", mspe))
```

```
[1] "Mean Squared Prediction Error (MSPE): 131.077556337426"
```

3.0.2 2. (b) Poisson Regression Model Selection

Do we really need all of these predictors? Construct a new regression model leaving out `year` and calculate the MSE for this second model.

Decide which model is better. Explain why you chose the model that you did.

[10]:

```r
# Your Code Here
# Construct a new regression model without the "year" predictor
new_model <- glm(incidents ~ type + period, data = train, family = poisson)

# Calculate the mean squared error (MSE) for the new model using the test set
test$predicted_new <- predict(new_model, newdata = test, type = "response")
test$MSE_new <- (test$predicted_new - test$incidents)^2
mse_new <- mean(test$MSE_new)

# Display the MSE for the new model
print(paste("MSE for the model without 'year':", mse_new))
```

```
[1] "MSE for the model without 'year': 275.122550627591"
```

[13]:

```r
# # Can compare nested poisson models with a chi-squared
pchisq(new_model$deviance-model$deviance, df=new_model$df.residual-model$df.
 ↪residual, lower.tail=FALSE)
```

0.0929203838345225

The chi-squared test gives a p-value of 0.09, which is greater than the significance level set at 0.05. There is not enough evidence to reject the null hypothesis at =0.05, meaning that the reduced model may be adequate. However, when considering the mean squared prediction error (MSPE), the model without the year predictor demonstrates a significantly higher value compared to the model with the year predictor. if our objective is prediction accuracy, it seems that the full model performs better.

3.0.3 2. (c) Deviance

How do we determine if our model is explaining anything? With linear regression, we had a F-test, but we can't do that for Poisson Regression. If we want to check if our model is better than the null model, then we're going to have to check directly. In particular, we need to compare the deviances of the models to see if they're significantly different.

Conduct two χ^2 tests (using the deviance). Let $\alpha = 0.05$:

1. Test the adequacy of null model.
2. Test the adequacy of your chosen model agaisnt the saturated model (the model fit to all predictors).

What conclusions should you draw from these tests?

[16]:

```r
# Fit the null model
null_model <- glm(incidents ~ 1, data = train, family = poisson)

# Test the adequacy of the null model
null_deviance <- null_model$deviance
df_null <- null_model$df.residual

# Test the adequacy of your chosen model against the saturated model
chosen_deviance <- model$deviance
df_chosen <- model$df.residual

# Obtain the deviance of the saturated model (model fit to all predictors)
saturated_model <- glm(incidents ~ type + period + year, data = train, family = poisson)
saturated_deviance <- saturated_model$deviance
df_saturated <- saturated_model$df.residual

# Calculate the chi-squared test statistics
null_chi_sq <- null_deviance - chosen_deviance
chosen_chi_sq <- chosen_deviance - saturated_deviance

# Calculate the p-values using the chi-squared distribution
null_p_value <- pchisq(null_chi_sq, df_null - df_chosen, lower.tail = FALSE)
chosen_p_value <- pchisq(chosen_chi_sq, df_chosen - df_saturated, lower.tail = FALSE)

# Display the results
print(paste("Null Model - Deviance Chi-squared test statistic:", null_chi_sq))
print(paste("Null Model - p-value:", null_p_value))
print(paste("Chosen Model vs. Saturated Model - Deviance Chi-squared test statistic:", chosen_chi_sq))
print(paste("Chosen Model vs. Saturated Model - p-value:", chosen_p_value))


# # Your Code Here
# chisq.stat = with(train, sum((incidents - fitted(model))^2/fitted(model)))
# # Test chi_sq stat
# pchisq(chisq.stat, df=model$df.residual, lower.tail=FALSE)
# # Test against the saturated model
# model.sat = glm(incidents~., train, family="poisson")
# pchisq(model$deviance-model.sat$deviance, df=model$df.residual-mode
# l.sat$df.residual, lower.tail=FALSE)
```

```
[1] "Null Model - Deviance Chi-squared test statistic: 445.491474434355"
[1] "Null Model - p-value: 3.41896472956775e-91"
[1] "Chosen Model vs. Saturated Model - Deviance Chi-squared test statistic: 0"
```

```
[1] "Chosen Model vs. Saturated Model - p-value: 1"
```

In this code, we obtain the deviances and degrees of freedom for the null model, the chosen model (with predictors), and the saturated model (fit to all predictors). We then calculate the chi-squared test statistics by taking the difference in deviances between the models. Finally, we calculate the p-values using the chi-squared distribution and compare them against the chosen significance level (e.g., alpha = 0.05).

Interpreting the results:

- For the adequacy of the null model:
  - If the p-value is below the significance level (e.g., 0.05), we reject the null hypothesis and conclude that the null model is inadequate, indicating that there is a significant difference between the null model and the chosen model.
  - If the p-value is above the significance level, we fail to reject the null hypothesis, suggesting that the null model is adequate, and the chosen model does not significantly improve the fit.
- For the adequacy of the chosen model against the saturated model:
  - If the p-value is below the significance level (e.g., 0.05), we reject the null hypothesis and conclude that the chosen model is significantly better than the saturated model, indicating that the chosen model provides a better fit.
  - If the p-value is above the significance level, we fail to reject the null hypothesis, suggesting that the chosen model is not significantly different from the saturated model, and the additional predictors do not improve the fit significantly.

3.0.4 2. (d) Poisson Regression Visualizations

Just like with linear regression, we can use visualizations to assess the fit and appropriateness of our model. Is it maintaining the assumptions that it should be? Is there a discernable structure that isn't being accounted for? And, again like linear regression, it can be up to the user's interpretation what is an isn't a good model.

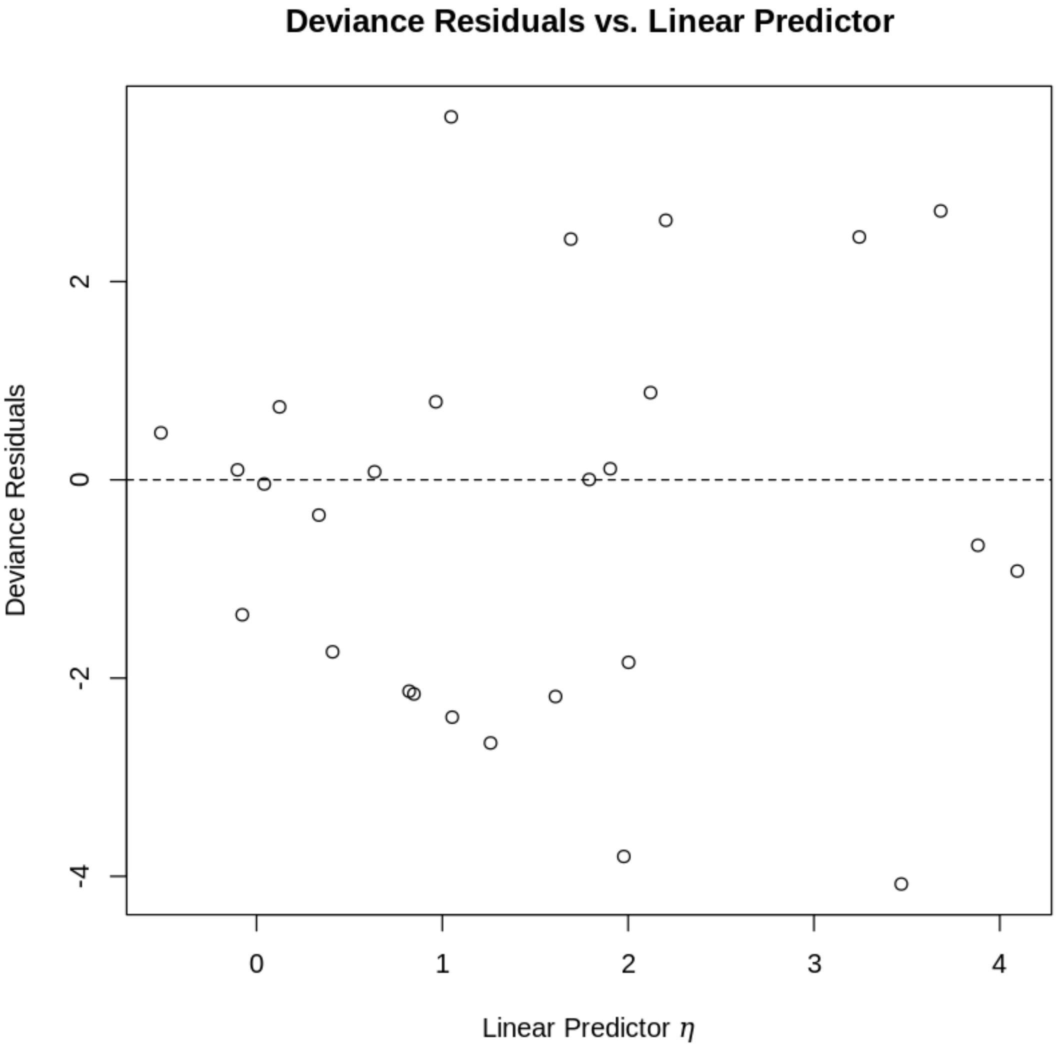
Plot the deviance residuals against the linear predictor η. Interpret this plot.

[17]:
```
# Your Code Here
# Compute the deviance residuals
deviance_residuals <- residuals(model, type = "deviance")

# Compute the linear predictor
linear_predictor <- model$linear.predictors

# Create the plot
plot(linear_predictor, deviance_residuals, xlab = "Linear Predictor  ", ylab =␣
 ↪"Deviance Residuals", main = "Deviance Residuals vs. Linear Predictor")

# Add a horizontal line at y = 0 for reference
abline(h = 0, lty = 2)
```



compute the deviance residuals using the residuals function with type = "deviance". The linear predictor is obtained from the linear.predictors attribute of the fitted model. Then create a scatter plot of the deviance residuals against the linear predictor using the plot function.

Interpreting the plot:

- The deviance residuals represent the observed minus expected response values, scaled by the square root of the estimated dispersion. They are used to assess the fit of the model and check for any patterns or systematic deviations.
- In the plot, we examine how the deviance residuals behave across the range of the linear predictor . A well-fitted model should exhibit random scatter of points around the horizontal line at y = 0.
- If the plot shows a systematic pattern or any departure from randomness, it indicates a potential lack of fit or violation of assumptions.

- Common patterns to look for include U-shaped or V-shaped curves, increasing or decreasing spreads, or outliers. These patterns suggest that the model may not adequately capture the relationship between the predictors and the response.

3.0.5 2. (e) Overdispersion

For linear regression, the variance of the data is controlled through the standard deviation σ, which is independent of the other parameters like the mean μ. However, some GLMs do not have this independence, which can lead to a problem called overdispersion. Overdispersion occurs when the observed data's variance is higher than expected, if the model is correct.

For Poisson Regression, we expect that the mean of the data should equal the variance. If overdispersion is present, then the assumptions of the model are not being met and we can not trust its output (or our beloved p-values)!

Explore the two models fit in the beginning of this question for evidence of overdisperion. If you find evidence of overdispersion, you do not need to fix it (but it would be useful for you to know how to). Describe your process and conclusions.

[20]:

```r
# Your Code Here
# Calculate the residual deviance and the degrees of freedom for each model
null_residual_deviance <- null_model$deviance
null_df <- null_model$df.residual

chosen_residual_deviance <- model$deviance
chosen_df <- model$df.residual

# Calculate the dispersion parameter estimates
null_dispersion <- null_residual_deviance / null_df
chosen_dispersion <- chosen_residual_deviance / chosen_df

# Compare the dispersion estimates to the expected value of 1
print(paste("Null Model Dispersion Estimate:", null_dispersion))
print(paste("Chosen Model Dispersion Estimate:", chosen_dispersion))

summary(model)
summary(new_model)
```

```
[1] "Null Model Dispersion Estimate: 21.334759759364"
[1] "Chosen Model Dispersion Estimate: 6.06734885050602"

Call:
glm(formula = incidents ~ type + period + year, family = poisson,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0775  -1.9869  -0.0418   0.7612   3.6618
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.5644     0.2199   7.113 1.13e-12 ***
typeB         1.6795     0.1889   8.889  < 2e-16 ***
typeC        -2.0789     0.4408  -4.717 2.40e-06 ***
typeD        -1.1551     0.2930  -3.943 8.06e-05 ***
typeE        -0.5113     0.2781  -1.839   0.0660 .
period75      0.4123     0.1282   3.216   0.0013 **
year65        0.4379     0.1885   2.324   0.0201 *
year70        0.2260     0.1916   1.180   0.2382
year75        0.1436     0.3147   0.456   0.6481
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 554.70  on 26  degrees of freedom
Residual deviance: 109.21  on 18  degrees of freedom
AIC: 200.92

Number of Fisher Scoring iterations: 6



Call:
glm(formula = incidents ~ type + period, family = poisson, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2377  -1.9003  -0.1372   0.6377   3.8906

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7190     0.1838   9.355  < 2e-16 ***
typeB         1.7831     0.1781  10.014  < 2e-16 ***
typeC        -2.0573     0.4394  -4.683 2.83e-06 ***
typeD        -1.1281     0.2918  -3.866 0.000111 ***
typeE        -0.4831     0.2767  -1.746 0.080787 .
period75      0.4723     0.1222   3.865 0.000111 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 554.70  on 26  degrees of freedom
Residual deviance: 115.63  on 21  degrees of freedom
AIC: 201.34
```

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

calculate the residual deviance and the degrees of freedom for each model. The dispersion parameter estimates are then obtained by dividing the residual deviance by the degrees of freedom. Then compare the dispersion estimates to the expected value of 1.

Interpreting the results:

- If the dispersion estimates are close to 1, it suggests that the Poisson regression models adequately capture the variability in the data, and there is no evidence of overdispersion.
- If the dispersion estimates are substantially greater than 1, it indicates potential overdispersion, where the observed variance is higher than expected under the Poisson distribution assumption.
- If overdispersion is present, it means that the model may not provide an accurate representation of the data and the standard errors, p-values, and confidence intervals may be unreliable.

in summary, if the Residual deviance is substantially larger than the degrees of freedom, it indicates the presence of overdispersion. In the case of both models, this condition holds true, suggesting that both models exhibit overdispersion

[ ]: