C3M2_peer_reviewed

June 13, 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.

[1]:
```
# 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 $i = 1, ..., n$.

If we dont consider any predictor to calculate the MLE of the joint poisson PMF, then by default only based on the observed counts we know that the MLE is the sample average.

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%).

[2]:
```
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	year	period	service	incidents
	<fct>	<fct>	<fct>	<int>	<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.

[3]:
```
# Your Code Here

glm_ship = glm(incidents ~ type + period + year, data = train, family = poisson)
```

[4]:
```
summary(glm_ship)
```

```
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
```

[24]:
```
y_hat = predict(glm_ship, test, type="response")
```

[25]:
```
MSPE = mean((y_hat - test$incident)^2)
```

[26]:
```
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.

[8]:
```
# Your Code Here

glm_ship_r = glm(incidents ~ type + period, data = train, family = poisson)
```

[9]:
```
summary(glm_ship_r)
```

```
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
```

[27]:
```
y_hat_r = predict(glm_ship_r, test, type="response")
```

[28]:
```
MSPE_r = mean((y_hat_r - test$incident)^2)
```

[29]:
```
MSPE_r
```

275.122550627591

[35]:
```
# Can compare nested poisson models with a chi-square

pchisq(glm_ship_r$deviance-glm_ship$deviance, df=glm_ship_r$df.residual -␣
↪glm_ship$df.residual, lower.tail=FALSE)
```

0.0929203838345225

Looks like the reduced model has higher MSPE, but the chi-squared test has a p-value of 0.09> , so we fail to reject the null at =0.05 and might conclude that the reduced model is sufficient.

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?

[36]:
```
sum(residuals(glm_ship_r, type = "pearson")^2)
```

103.696416445113

[41]:
```
# Your Code Here
# Test if the model is better than the null model
chisq.stat = with(train, sum((incidents - fitted(glm_ship))^2/fitted(glm_ship)))
# Test chi_sq stat
pchisq(chisq.stat, df=glm_ship$df.residual, lower.tail=FALSE)
# Test against the saturated model
glm.sat = glm(incidents~., train, family="poisson")
pchisq(glm_ship$deviance-glm.sat$deviance, df=glm_ship$df.residual-glm.sat$df.
↪residual, lower.tail=FALSE)
```

4.22139949448423e-13

1.85320875968548e-19

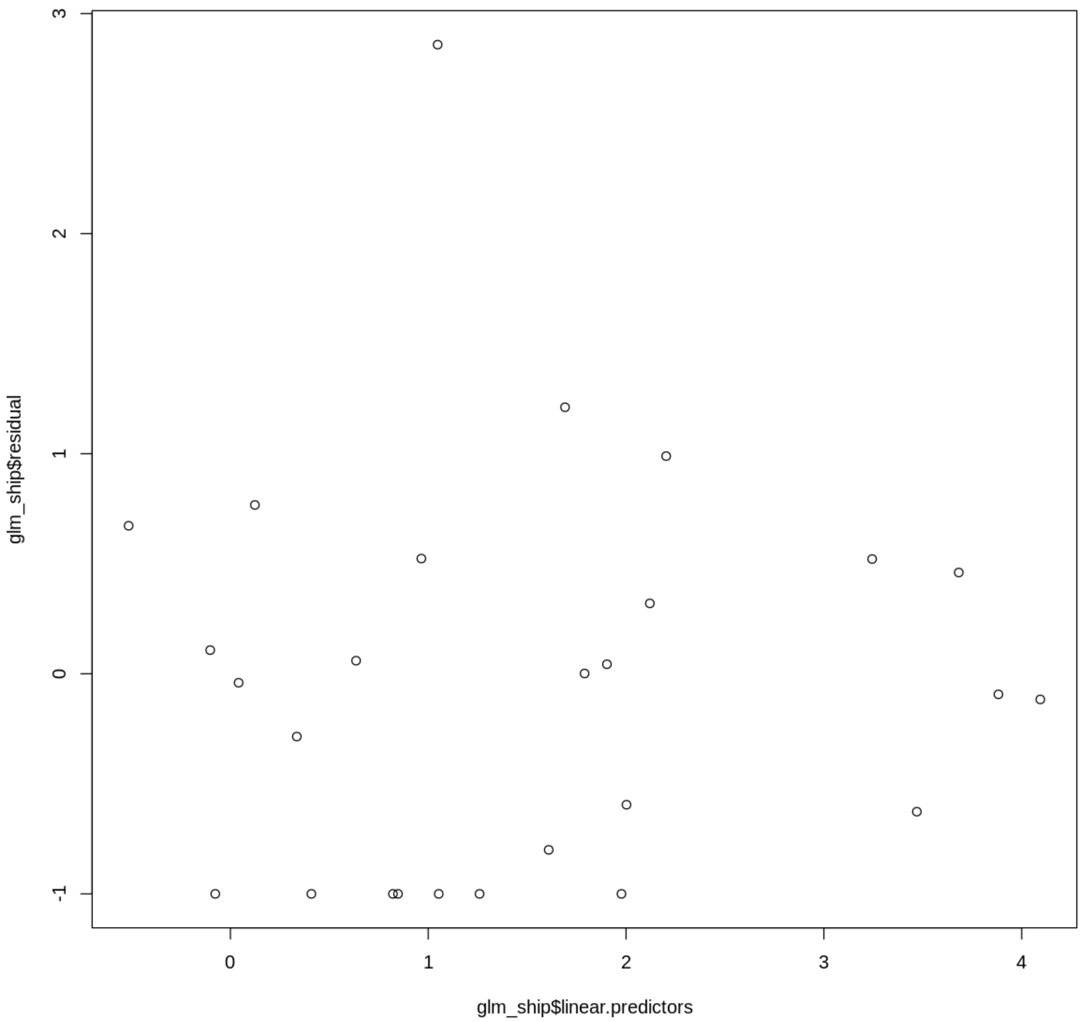
The model performs better than the null model based on the chisquare-test, but compared to the saturated model it is not explaining well enough the data. Obviously this way the null model is not adequate either. We need to look further to see if there are unnecessary predictors in our model or on the contrary we are missing a couple of more powerful ones. Overdispersation can be present as well.

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.

[42]:
```
# Your Code Here
plot(x=glm_ship$linear.predictors, y=glm_ship$residual)
```



Nothing extremely strange, 1 outlier

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.

```
[43]: dp = sum(residuals(glm_ship, type = "pearson")^2)/glm_ship$df.res
```

```
[44]: dp
```

5.47049561223492

```
[45]: # Your Code Here
      dp_r = sum(residuals(glm_ship_r, type = "pearson")^2)/glm_ship_r$df.res
```

```
[46]: dp_r
```

4.93792459262441

```
[21]: glm_ship_q = glm(incidents ~ type + period, data = train, family = quasipoisson)
```

```
[22]: drop1(glm_ship_q, test = "F")
```

A anova: 3 × 4

	Df	Deviance	F value	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>
<none>	NA	115.6311	NA	NA
type	4	554.4650	19.924386	6.615564e-07
period	1	130.7545	2.746601	1.123269e-01

Based on the results some overdispersion is present. For the full model it is above 5 slightly (Residual deviance is significantly greater than the degrees of freedom in both cases). If we use a quasi Poisson model that adjusts the standard errors, we see that the adjustment changes the significance of some of the levels of the period factor. We can test whether we should leave the period factor in the model. Accroding to F-test we can even drop the period factor as well.

```
[ ]:
```