C3M2_peer_reviewed

March 21, 2022

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.

[45]:
```
# 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 $i = 1, ..., n$.

Since all other predictors are zero, then we know all instances of η is equal to β_0. Therefore, the MLE for β_0 is actually :

$$\sum_{i=1}^{n}[y_i\beta_0 - e^{\beta_0} - log(y_i!)]$$

Now we maximize our MLE by taking the derivative, in respect of β_0, of our MLE stated before , and setting it to zero. Then, we solve for β_0

$$\frac{dMLE(\beta_0)}{d\hat{\beta}_0} = \sum_{i=1}^{n}[y_i - e^{\hat{\beta}_0}] = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} e^{\hat{\beta}_0} = 0 \rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} e^{\hat{\beta}_0} \rightarrow \sum_{i=1}^{n} y_i = ne^{\hat{\beta}_0} \rightarrow \frac{\sum_{i=1}^{n} y_i}{n} = e^{\hat{\beta}_0} \rightarrow \bar{y} = e^{\hat{\beta}_0}$$

We know, in Poisson Regression, $\hat{\lambda} = e^{\hat{\beta}_0}$. Thefore we can replace $e^{\hat{\beta}_0}$ with $\hat{\lambda}$ and get our result $\hat{\lambda} = \bar{Y}$

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%).

[46]:
```
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.

[47]:
```
# Your Code Here
lmod.ship = glm(incidents~ type + period + year, data = train,
family=poisson)
summary(lmod.ship)

mean((test$incidents - predict(lmod.ship, test, type="response"))^2)
```

```
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
```

131.077556337426

3.1 Answer

The MSPE is 131.0776

3.1.1 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.

[48]:
```
# Your Code Here
lmod.ship_dropyear = glm(incidents~ type + period, data = train,
    family=poisson)
mean((test$incidents - predict(lmod.ship_dropyear, test,
    type="response"))^2)
```

275.122550627591

[49]:
```
pchisq(lmod.ship$deviance, lmod.ship$df.resid, lower.tail=FALSE)
```

4.4110510959471e-15

[50]:
```
pchisq(lmod.ship_dropyear$deviance - lmod.ship$deviance,
    lmod.ship_dropyear$df.resid - lmod.ship$df.residual, lower.tail=FALSE)
```

0.0929203838345225

[51]:
```
# Can compare nested poisson models with a chi-squared
```

3.2 Answer

When dropping the year predictor, our MSPE increases a lot. However, when testing for goodness of fit using the nest Chi-square test, fail to reject the null hypothesis that the reduced model is sufficient. However, Tthe full model is much better at making predictions; also, the p-value of the chi-squared test is insignificant but not strongly. Thus, I would choose the full model.

3.2.1 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?

[52]:
```r
# Your Code Here
# Test if the model is better than the null model
chisq.stat = sum((train$incidents - fitted(lmod.ship))^2 / fitted(lmod.ship))
# Test chi_sq stat
pchisq(chisq.stat, lmod.ship$df.residual, lower.tail=FALSE)
# Test against the saturated model
lmod.sat = glm(incidents~., data=train, family="poisson")
pchisq(lmod.ship$deviance-lmod.sat$deviance, lmod.ship$df.residual-lmod.sat$df.
↪residual, lower.tail=FALSE)
```

4.22139949448423e-13

1.85320875968548e-19

3.3 Answer

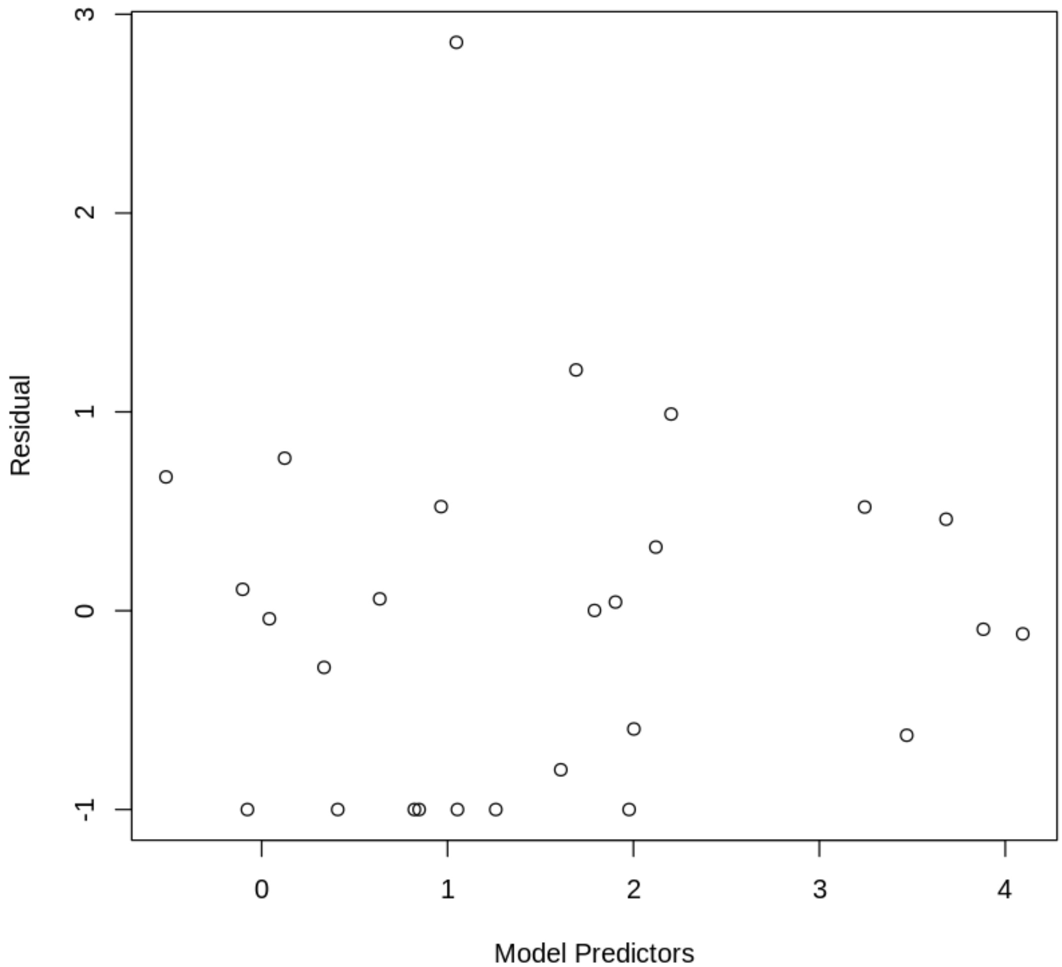
Both of our tests are strongly significant. Therefore, we can say our model is better than the null model; however, it is not better than the saturated model.

3.3.1 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.

[53]:
```r
# Your Code Here
plot(lmod.ship$linear.predictors, lmod.ship$residual, xlab="Model Predictors",
ylab="Residual")
```



3.4 Answer

The linear predictor is the value of η_i's before each one of them is transfomred by the link function. We interpret this plot similar to how we would a residul vs fitted plot in linear regression. Overall, the residuas show nonconstant variance, which is good. The only issue may seem to be the residual that is close to three. However, that would take more exploration.

3.4.1 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.

[54]:
```
# Your Code Here
summary(lmod.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
```

[55]:
```
summary(lmod.ship_dropyear)
```

```
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
```

3.5 Answer:

We can check if a model has overdispersion my deviding the residual deviance by the degrees of freedom. If the resulting quotient is greater than 1, then you have overdispersion. For both models the degrees of freddom do not equal the residual deviance; therefore, we would have a quotient greater than one in both cases. Thus, overdispersion is apparent in both models.