C3M2: Peer Reviewed Assignment

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.

In [2]:
```
# Load the required packages
library(MASS)
```

Problem 1: Poisson Estimators

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

The Class Videos walked through the steps to get to the log-likelihood of a Poisson as follows:

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

The problem asks us to find the Null case (intercept only, focusing on β_0) we can assume all other predictors are moot (β_1, \beta_2, ...) . Now the slope of the intercept determines the result. Additionally, we know that $\lambda = e^{\beta_0}$ from the quiz.

$\eta_i = \beta_0$

We substitute in the β and begin the maximizing process:

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

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

$\frac{\partial}{\partial \beta_0} \ell(\beta_0) = \sum_{i=1}^{n} y_i - e^{\beta_0} =^{set} = 0$

----------> $\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} e^{\beta_0} = 0$

----------> $\sum_{i=1}^{n} y_i = n * e^{\beta_0}$

----------> $\frac{1}{n} \sum_{i=1}^{n} y_i = e^{\beta_0}$

----------> $\bar{Y} = e^{\beta_0}$

Since $\hat{\lambda} = e^{\eta_i}$ and we know $e^{\eta} = e^{\beta_0} = \bar{Y}$ we plug in the finding:

----------> $\hat{\lambda} = e^{\eta_i} = e^{\beta_0} = e^{log(\bar{Y})} = \bar{Y}$

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%).

In [3]:
```
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
```

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.

In [4]:
```
# Your Code Here
cat("Build Model and Show: \n")
shipglm = glm(incidents ~ type + period + year, data = train , family=
summary(shipglm)

cat("----------------------------------------\n Calculate MSPE \n")
preds = predict(shipglm, newdata = test, type = "response")
mspe = mean((test$incidents - preds)^2)
cat("MSPE:",round(mspe,4))
```

```
Build Model and Show:


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

----------------------------------------
 Calculate MSPE
MSPE: 131.0776
```

Analysis: The fitted GLM poisson model exhibits a MSPE of 131 and residual deviance of 109.21 on 18 degrees of freedom. This is a count model because the period of examination (or denominator of a rate in this instance) is not part of the requested analysis.

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 MSPE for this second model.

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

In [5]:
```
# Your Code Here
cat("Build Model and Show: \n")
ship2glm = glm(incidents ~ type + period , data = train , family=poiss
summary(ship2glm)

cat("----------------------------------------\n Calculate MSPE \n")
preds2 = predict( ship2glm, test, type="response")
mspe2 = mean((test$incidents - preds2)^2)
cat("MSPE:", round(mspe2,4))
```

```
Build Model and Show:


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

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 MSPE
MSPE: 275.1226
```

In [6]:
```
# Can compare nested poisson models with a chi-squared
df1 = shipglm$df.residual
df2 = ship2glm$df.residual
dev1 = shipglm$deviance
dev2 = ship2glm$deviance
cat("P-Value of Test:", pchisq(dev2 - dev1, df2-df1, lower.tail=FALSE)

# do it again... (more complex model goes first)
anova(shipglm, ship2glm, test = "Chisq")
```

```
P-Value of Test: 0.09292038
```

A anova: 2 × 5

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	18	109.2123	NA	NA	NA
2	21	115.6311	-3	-6.418801	0.09292038

We built and trained two count models with training data. We then tested by predicting results with test data (unseen). We can assume that the unseen data is independent as it was randomly selected...

The chisquared test to compare models resulted in a P-Value of .093 in both methods (ANOVA and Hand Calculated).

- This P-Value is less than an alpha at .1, but greater than an alpha of .05.
- We fail to reject the null hypothesis H_0 if alpha is .05, that the simple (less complex) model is better.
- We reject the null hypothesis H_0 if alpha is .1, that the more complex model is better.
- This relatively close call is probably p-hacking because we didn't have a set p-value prior to our testing.
- Let's move on to the other metrics to make a determination...

The MSPE for the simple model is over double that of the complex model (average error is over 2x more). Since we know these results were on testing data, and can incorporate the chisquared model comparison results, we can conclude that the more complex model is in fact more accurate on the test data. We would need to be transparent about how me made the decision not based purely on the alpha.

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 against the full model (the model fit to all predictors).

What conclusions should you draw from these tests?

In [23]:
```
# Your Code Here
# Test if the model is better than the null model
summary(shipglm)
# Test pearson chi_sq stat
pears = sum((train$incidents - shipglm$fitted)^2/shipglm$fitted)
cat("pearson test statistic:", pears,'\n \n')
#length(train$incidents) - length(shipglm$coef)

#pearson method:
cat("Pearson P-Value:", pchisq(pears,length(train$incidents) - length(
# chisq ((DNull - DResid) (DF Null - DF Proposed), lower.tail=FALSE)
cat("\n \n Deviance Method Pvalue:", pchisq(109.21,18,lower.tail=FALSE

# Test against the null model

cat("\n \n ~R2 of Model:", 1- 109.21/554.7)
```

```
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

pearson test statistic: 98.46892

Pearson P-Value: 4.221399e-13

 Deviance Method Pvalue: 4.415359e-15

 ~R2 of Model: 0.8031188
```

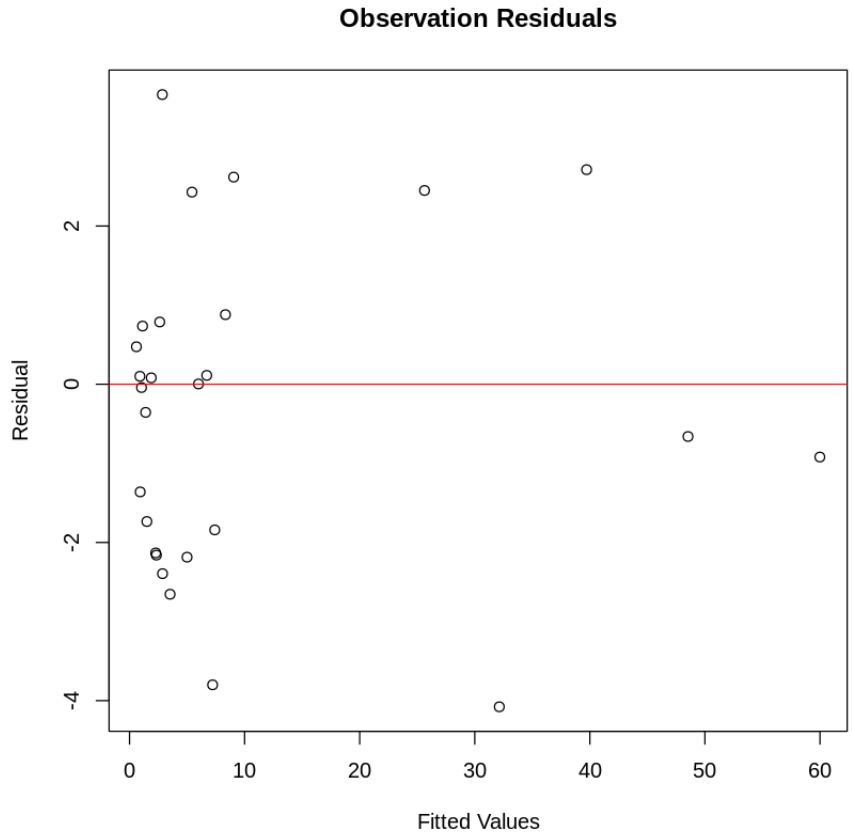
These calculations show that the model is indeed signficantly better than the null model, with an ability to model roughly 80% of the variance in the data. The p-values from both tests are well below the alpha at .05, giving us adequate reason to reject the null hypothesis, that the null model is better.

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.

In [51]:
```
# Your Code Here
plot(shipglm$fitted,residuals(shipglm, type='deviance'),xlab="Fitted V
abline(h=0,col="red")
```



The plot above shows the residuals on the y-axis and the fitted values on the x-axis. Because the residuals are evenly distributed and random about the centerline, I would have to conclude that there is no pattern here. There is not data sufficient to reject the null hypothesis, the homoscedacity of residuals assumption. The data is not uniformly distributed across the fitted values on the x-axis, but that does not violate the assumption.

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.

In [49]:
```
# Your Code Here
#method 1 compute dispersion parameter and alter original model
dpa <-  sum(residuals(shipglm, type="pearson")^2)/shipglm$df.residual
cat("dispersion parameter:",dpa)
summary(shipglm, dispersion = dpa)


summary(shipglm)

ship2glm = glm(incidents ~ type + period , data = train , family= quas
summary(ship2glm)

cat("\n Pvalue of quasipoisson:",pchisq(115.63, 21,lower.tail=FALSE))
```

```
dispersion parameter: 5.470496


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     0.5144   3.041 0.002356 **
typeB         1.6795     0.4419   3.801 0.000144 ***
typeC        -2.0789     1.0309  -2.017 0.043737 *
typeD        -1.1551     0.6852  -1.686 0.091849 .
typeE        -0.5113     0.6504  -0.786 0.431777
period75      0.4123     0.2998   1.375 0.169093
year65        0.4379     0.4408   0.993 0.320486
year70        0.2260     0.4481   0.504 0.614052
year75        0.1436     0.7362   0.195 0.845292
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 5.470496)

    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 + 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 = quasipoisson,
    data = train)

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7190     0.4083   4.210 0.000394 ***
```

```
typeB         1.7831     0.3957   4.506 0.000194 ***
typeC        -2.0573     0.9763  -2.107 0.047295 *
typeD        -1.1281     0.6484  -1.740 0.096540 .
typeE        -0.4831     0.6148  -0.786 0.440767
period75      0.4723     0.2716   1.739 0.096631 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 4.937925)

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

Number of Fisher Scoring iterations: 6


 Pvalue of quasipoisson: 4.493538e-15
```

There is a small amount of overdispersion, calculated between 4.9 and 5.4, in this model. The original model is calculated with a dispersion of 1, so the difference is between 3.9 and 4.4. However, this doesn't really affect the deviance much (the original deviance is 109.2 on 18df, and quasipoisson model has 115.63 on 21df). If we go ahead and compute the p-values because the overdispersion is not "large", they are also strikingly similar. To test the hypothesis, we are looking for a decrease in residual deviance in the original model summary (shipglm) computed with the altered dispersion, OR a lower p-value on a quasipoisson model. In this case, the pvalue of the quasipoisson (4.49e-15) is very similar to the p-value on the original poisson model (4.41e-15). Without much difference between the two models I would have to fail to reject the null hypothesis, the original model is adequate.

In [ ]: