Lab 7 - Poisson Regression

Lightning Mining Data

The data set `lightning.csv` has data on lightning strikes in different regions in the North Central Plains in Texas. Mining is done for Copper Ore in the region, and miners in the area often tell the story of lighting around the mines. They say that the lightning is attracted to the copper which causes more lightning strikes around the mine.

A geologist found this claim interesting, as it relates to an electromagnetic method known as magnetotellurics for inferring the earth's subsurface of electrical materials. So, she decided to test this.

She obtained estimates of copper ore density in mines (`COD`), and the size of the top-down surface area that was susceptible to lightning strikes (`MineSize`). She also chose a few separate areas that were known to not have any magnetic material in the surrounding area, and these had `COD` $= 0$. Over 2 years, she recorded the number of lightning strikes over each area on nights where the was a thunderstorm in the area for the spring-summer storm season.

Variable	Description
`Strikes`	number of lightning strikes per storm
`COD`	copper density in corresponding mine
`MineID`	name of each mining site
`MineSize`	size of each mine

Since our data is a count of lightning strikes with no upper bound, we can model this with a poisson regresion.

Since our data is a count of lightning strikes with no upper bound, we can model this with a poisson regresion.

1. Fit a poisson regression, predicting the response, `Strikes`, with predictor copper density, `COD`.

```
Lightning = read.csv("Lightning.csv")

#fit poisson GLM
light.glm = glm(Strikes~COD, family = poisson, data = Lightning)
summary(light.glm)
```

```
##
## Call:
## glm(formula = Strikes ~ COD, family = poisson, data = Lightning)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.37377  -0.83755  -0.01938   0.66231   2.09193
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.03581    0.09777  10.594  < 2e-16 ***
## COD          0.65483    0.20538   3.188  0.00143 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 69.557  on 64  degrees of freedom
## Residual deviance: 59.601  on 63  degrees of freedom
## AIC: 256.11
##
## Number of Fisher Scoring iterations: 4
```

2. Run a goodness-of-fit test for this mode, then obtain the overdispersion parameter using the quasipoisson fit.

The residual deviance is 59.601, which under the null has a χ^2 distribution with 63 degres of freedom. The good-ness of fit test statistic is

```
1-pchisq(59.601,63)
```

```
## [1] 0.5982129
```

Therefore we do not have evidence of a poor fit.

Overdispersion parameter is obtained via

```
light.Qglm = glm(Strikes~COD, family = quasipoisson, data = Lightning)
summary(light.Qglm)
```

```
##
## Call:
## glm(formula = Strikes ~ COD, family = quasipoisson, data = Lightning)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.37377  -0.83755  -0.01938   0.66231   2.09193
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.03581    0.09337  11.093  < 2e-16 ***
## COD          0.65483    0.19614   3.339  0.00142 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.9120407)
##
##     Null deviance: 69.557  on 64  degrees of freedom
## Residual deviance: 59.601  on 63  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

The overdispersion parameter is 0.9120407, which is consistent with our goodness-of-fit test.

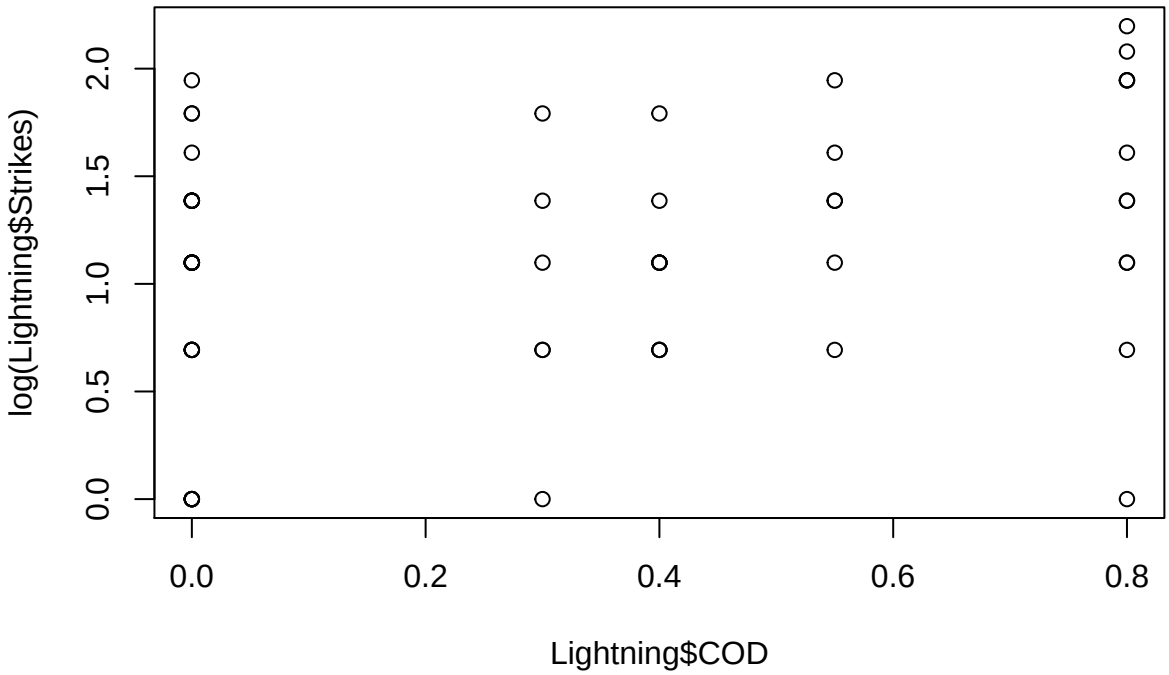
3. Do the assumptions hold for this model? What about the sampling scheme could have us worried about the assumptions holding?

Independence is potentially an issue due to the sampling scheme, however the overdispersion parameter not being large implies that there was not a significant correlation between poisson events.

Also, the deviance residuals are within an appropriate range in the original model.

We can also investigate the linearity assumption further, by plotting the count data.

```
plot(Lightning$COD, log(Lightning$Strikes))
```



Linearity appears to be a reasonable assumption.

4. Even if the assumptions don't hold, take a look at the Wald-test pvalue for the COD in the `family = poisson` model. Interpret this slope within the context of the current model and make a preliminary conclusion.

The Wald test gives a p-value = 0.00143. Therefore, in this model, we have significant evidence that copper mines with higher copper densities tend to attract some lightning strikes.

In this model, we predict that the rate of lightning strikes increases by a factor of $e^{0.65483} = 1.925$ for every unit increase in the density variable.

5. If a mine has twice as much available surface area to be struck with lightning, we would expect the twice as many lightning strikes. In order to account for this in our model, use the offset parameter to include `MineSize` in the glm fit.

```
light.glmW = glm(Strikes~COD, family = poisson, data = Lightning, offset = log(MineSize))
summary(light.glmW)
```

```
##
## Call:
## glm(formula = Strikes ~ COD, family = poisson, data = Lightning,
##     offset = log(MineSize))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.40283  -0.62558  -0.01134   0.61868   2.04310
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.06015    0.09716   10.91   <2e-16 ***
## COD         -0.12577    0.20299   -0.62    0.536
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 59.099  on 64  degrees of freedom
## Residual deviance: 58.714  on 63  degrees of freedom
## AIC: 255.22
##
## Number of Fisher Scoring iterations: 4
```

6. Write out the fitted model.

Our fitted model is

$log(E(Strikes)) = 1.06015 - 0.12577COD + log(MineSize)$

7. Make final conclusions for the offset model.

After including the mine sizes in our dataset, we don't have evidence of that the rate of lightning strikes is associated with Copper Ore density at the 0.05 significance level. Therefore, the increased rate of lightning strikes we were fitting in our model appear to be more a function of mine-size than a copper density relationship.

Note that the geologist may know of a true magnetic field effect here corresponding to the magnetotulleric methods, however she now has evidence that the effect that this magnetic field has on lightning strikes must be weak, and is probably not a concern for miners.

let-7c miRNA data

In this example, data was collected for both males and females with and without breast cancer. We are looking to test if the mutations in the micro RNA gene let-7c are associated with post-transcriptional regulation of breast cancer. Not having a significant increase in mutation rate would not imply regulation of breast cancer, however a significant increase in mutation rate against the population would at least imply that this gene is significantly associated with the regulation of gene expression in the presence of breast cancer. We wish to find not only if there is a significant difference in mutation rate when breast cancer is present, but also if this change in mutation rate is different for males and females.

A single-nucleotide polymorphism (SNP) is a relatively common mutation in a gene. In the dataset `br7cSNP.csv` 100 SNPs were sequenced and the SNP was labeled 0 if the SNP was the same as the reference gene, and 1 otherwise. Data was collected for 100 people. The data appears as follows

Variable	Description
`Sub`	Subject ID
`MF`	0 for male, 1 for female
`BCan`	0 if breast cancer was not present, 1 otherwise
`SNP"i"`	0 if mutation did not occur, 1 otherwise

1. Load in the data set

```
br7cSNP = read.csv("br7cSNP.csv")
#look at data set first 5 rows and 5 columns of data set
br7cSNP[1:5, 1:5]
```

```
##   MF BCan SNP1 SNP2 SNP3
## 1  1    1    1    0    0
## 2  1    1    1    0    0
```

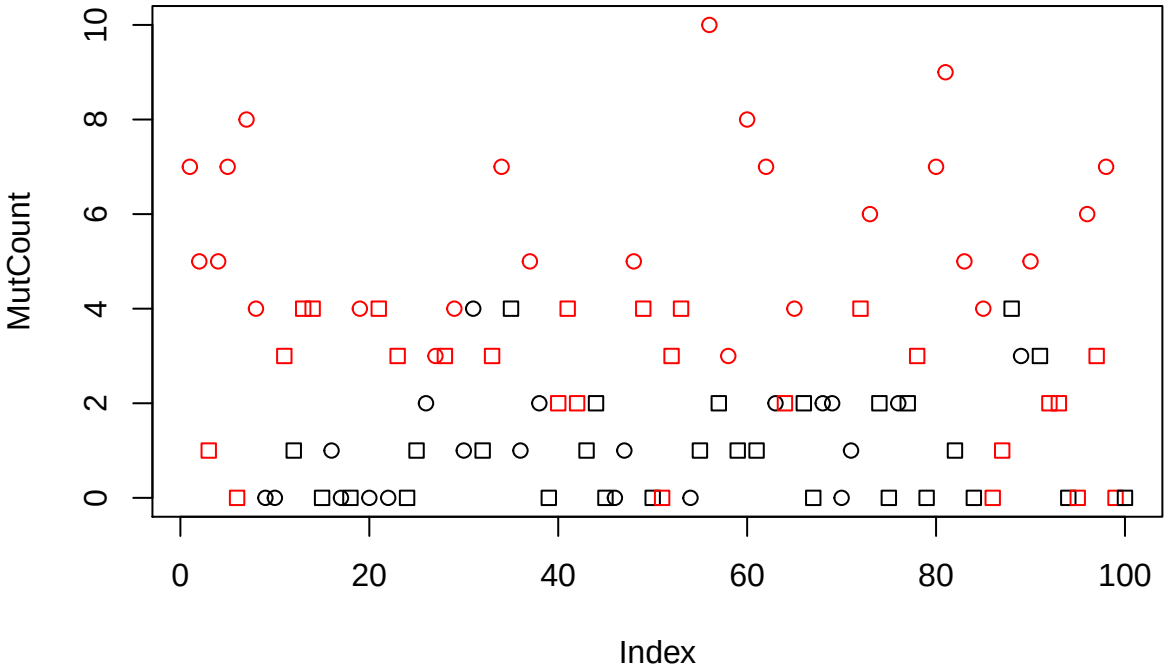
```
## 3  0    1    0    0    0
## 4  1    1    0    0    0
## 5  1    1    0    0    0
```

2. What is an appropriate distribution for SNP1 across subjects for a male without breast cancer?

Since `SNP1` is a either a mutation or not, and we can assume these mutations are independent across genes, we can conclude that $SNP1 \sim Binomial(1, p)$ for some mutation probability p.

3. Create a column in your dataset called `MutCount` that gives the total number of mutations across SNPs for each person.

```
br7cSNP$MutCount = rowSums(br7cSNP[,3:102])
plot(br7cSNP$MutCount, pch = br7cSNP$MF, col = br7cSNP$BCan+1, ylab = "MutCount")
```



4. Assuming SNP mutations are independnet, why is it appropriate to treat `MutCount` as either a Binomial or Poisson random variable? Why is it important to assume that SNP mutations are independent?

Since the number of mutations is very low for the binomial distribution, we are modelling rare events, therefore the poisson distribution provides a useful approximation. Therefore, you can work with the 3 variable dataset of `BCan`, `MF` and `MutCount` instead of the original 102 variable dataset.

A binomial distribution is a count of independent events so if the counts are not independent we could no longer conclude that distribution.

5. Fit Poisson GLM acounting for `MF` and `Bcan`. Do the assumptions hold? Was an interaction included in your model fit? Why or why not?

Fitting the GLM

```
miRNAglm = glm(MutCount ~ MF*BCan, family = 'poisson', data = br7cSNP)
summary(miRNAglm)
```

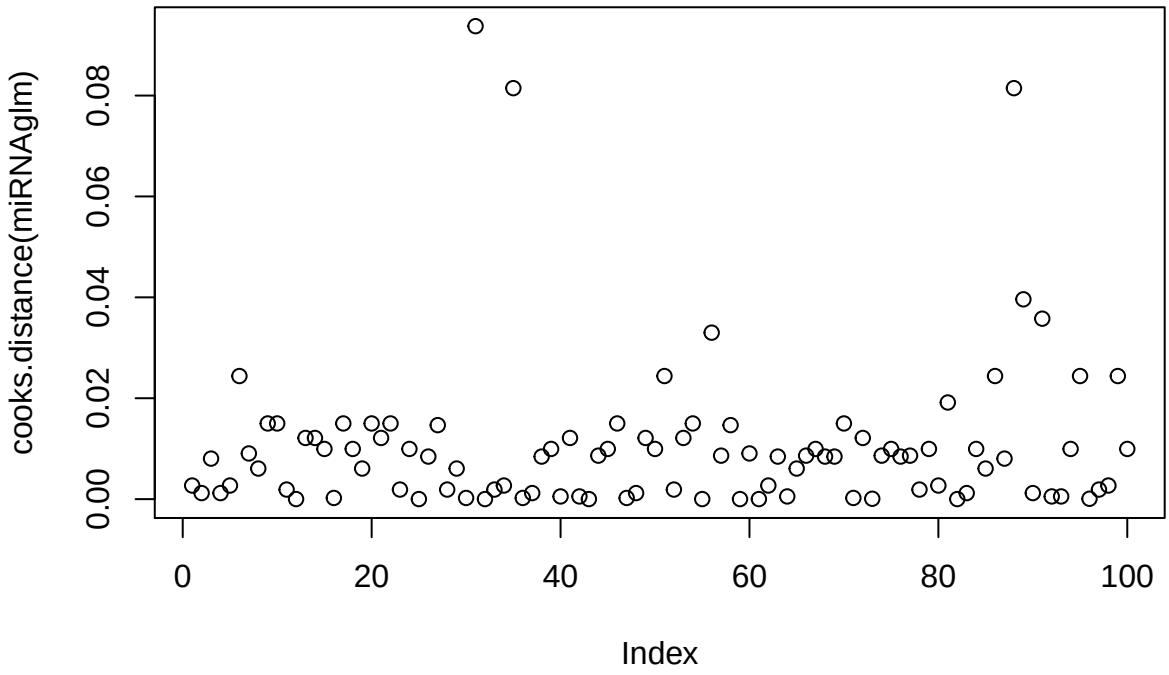
```
##
## Call:
## glm(formula = MutCount ~ MF * BCan, family = "poisson", data = br7cSNP)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -2.16617  -1.32160  -0.08593   0.72400   2.20931
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.03509    0.18569   0.189 0.850113
## MF           0.09844    0.27595   0.357 0.721293
## BCan         0.81769    0.22556   3.625 0.000289 ***
## MF:BCan      0.80664    0.31534   2.558 0.010527 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 245.44  on 99  degrees of freedom
## Residual deviance: 116.13  on 96  degrees of freedom
## AIC: 344.98
##
## Number of Fisher Scoring iterations: 5
```

```
#gof test
1-pchisq(116.13, 96)
```

```
## [1] 0.07935547
```

```
plot(cooks.distance(miRNAglm))
```



By looking at the cook's distance, no points are found to be influential, and the GOF test appears to imply an adequate fit.

The interaction is was included because one of our research questions is if the change in mutation rate is different for Males and Females, which cannot be ascertained from a non-interaction model.

6. Are either of `MF` or `Bcan` significant?

To test this we run an likelihood ratio test against the null model. First we fit the null model, and then test

if our full model significantly explains more of the noise in our observations.

```
miRNAglmSMALL = glm(MutCount ~ 1, family = 'poisson', data = br7cSNP)
anova(miRNAglmSMALL, miRNAglm, test = 'LRT')
```

```
## Analysis of Deviance Table
##
## Model 1: MutCount ~ 1
## Model 2: MutCount ~ MF * BCan
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        99     245.44
## 2        96     116.13  3    129.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have evidence that either `MF` or `BCan` is significant in our model.

7. Why is the hypothesis test in question 6 not exactly what we want to know from our data set? Run the appropriate hypothesis tests for our dataset, and make conclusions.

This hypothesis test could be significant simply if the mutation rate between males and females is significantly different, which would not be a significant finding given our research interests.

If we can show the following hypotheses, then we will have shown that the mutation rate is different for males and females, and the mutation rate increases for both males and females.

1) H_a : `MF:BCan` > 0, i.e. if the change in mutation rate is different in Males and Females
2) H_a : `BCan` > 0, i.e. if the mutation rate increases for Males

For testing 1), the pvalue can be obtained by hand for the one-sided test, although *since the estimate is positive* the pvalue can be obtained by halving the p-value from the summary table. This gives a pvalue for this test as $0.010527/2 = 0.0052635$.

For testing 2), similarly as for test 1), the pvalue is obtained as $0.000289/2 = 0.0001445$

Therefore we can conclude that the let-7c gene is associated with breast cancer, and the effect changes between males and females.

[Note that because we are doing multiple tests here, a multiple testing correction should be included.]