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.

1. Fit a poisson regression, predicting the response, `Strikes`, with predictor copper density, `COD`.
2. Run a goodness-of-fit test for this mode, then obtain the overdispersion parameter using the quasipoisson fit.
3. Do the assumptions hold for this model? What about the sampling scheme could have us worried about the assumptions holding?
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.
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.
6. Write out the fitted model.
7. Make final conclusions.

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
2. What is an appropriate distribution for SNP1 across subjects for a male without breast cancer?
3. Create a column in your dataset called `MutCount` that gives the total number of mutations across SNPs for each person.
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?
5. Fit Poisson GLM acounting for `MF` and `BFcan`. Do the assumptions hold? Was an interaction included in your model fit? Why or why not?
6. Are either of `MF` or `BRcan` significant?
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.