SS4864 Final Project Part I

Angela Zheng[1] and Oluwatitomi Adebajo[2]

[1]University of Western Ontario
[2]University of Western Ontario

December 2, 2022

1 Introduction

Generalized linear mixed models are an efficient way to analyze data sets with repeated measures, as they take into account random effects. For this project, we will be analyzing the **epilepsy.csv** data set with a model that has the underlying assumptions:

$$Y_{ij}|\mu_i \sim \text{Poisson}(\exp(\beta_0 + x_i\beta + Z_i))$$
$$Z_i \sim \text{N}(0, \sigma^2)$$

where Z_i represents the random effect that takes into account correlation between observations within a subject, and Y_{ij} represents the j-th observation for the i-th subject. Each entry Y_{ij} represents the number of seizures that the individual i had during the j-th measurement period. As the data set displays information about each subject's seizure count which is dependent on a multitude of factors about the individual, this project will mainly focus on the effectiveness of the drug Progabide to treat epileptic seizures.

2 Model

We will fit the model to a set of linear predictors that is assumed to have an effect on the number of seizures a patient experiences. The age of each subject (denoted by β_{age}) is a continuous variable that can take on any positive value greater than 0. The treatment group that the individual belongs to (denoted by β_{treat}) is a categorical variable with two levels, that can either take on the value 0 (placebo group) or 1 (drug group). Whether the seizure count was recorded during the baseline period or not (denoted β_{expind}) is also a categorical variable with two levels, that can either take on the value 0 (baseline measurement) or 1 (trial measurement).

When writing the expression for the linear predictors, the model takes on the following form:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} x_{3i} + \beta_3 x_3 i + Z_i) \tag{1}$$

where $\beta_1 = \beta_{age}$, $\beta_2 = \beta_{treat*expind}$, $\beta_3 = \beta_{expind}$, and Z_i represents the random effect ($Z_i \sim \text{N}(0, \sigma^2)$). To illustrate how the categorical data affects the model, let us focus on β_2 and β_3. When we only look at the baseline measurements where $x_{3i} = 0$, the form of linear predictors is:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{1i})$$

where μ_i does not depend on whether the individual was given the placebo or drug, as the number of seizures should be the same for both treatment groups during the baseline period. On the other hand, when we only look at the trial measurements where $x_{3i} = 1$, the form of linear predictors is either:

$$\mu_i = \begin{cases} \exp(\beta_0 + \beta_1 x_{1i} + \beta_3), & \text{if } x_{2i} = 0 \text{ (placebo treatment)} \\ \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 + \beta_3), & \text{if } x_{2i} = 1 \text{ (drug treatment)} \end{cases}$$

where both cases account for age. However, the main difference between the two cases is the presence of β_2 when an individual is given the drug treatment. β_2 represents the effect of the drug that is administered *during* the clinical trial. As such, we are interested in testing whether there is significant interaction between the drug and the measurement period.

3 Research Hypothesis

For our research question, we hypothesize that using the drug significantly reduces the number of seizures a patient experiences after accounting for age. Mathematically, the hypothesis can be written as such:

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 < 0$$

where $\beta_2 = \beta_{treat}$. A left-tailed hypothesis test is used in order to determine whether or not β_2 is a negative value. If the null hypothesis is true and β_2 is equal to 0, whether the drug is taken or not would not contribute to the number of seizures. In that case, we could conclude that the drug is not effective in reducing the number of seizures. Otherwise, if the alternative hypothesis is true and β_2 is negative, we could conclude that the drug is indeed effective in reducing the number of seizures. This is because a negative β_2 will decrease the value of μ_i in equation (1).

Traditionally, we would conduct the following t-test:

$$t' = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)}$$

in which the test statistic t' will be used to determine a p-value that can be compared to $\alpha = 0.05$. However, since the standard error will be calculated using bootstrapping, it means that conducting a t-test will lead to large computational times. Instead, multiple data sets will be generated via a parametric bootstrap (described in Section 5) and estimates of $\hat{\beta}_2$ will be found. It can be denoted as a vector of estimates: $\hat{\beta}_2^* = [\hat{\beta_2}^{1*}, \hat{\beta_2}^{2*}, \dots, \hat{\beta_2}^{n*}]$ where n is the number of bootstrapped data sets we choose to simulate. If the null hypothesis is true, then the large proportion of values in $\hat{\beta_2}^*$ should be very close to 0. We can estimate the p-value by counting the number of entries in $\hat{\beta}_2^*$ that are less than 0. Although not as efficient as going through the process of finding $SE(\hat{B_2})$, it allows for a good approximation.

If the p-value is less than α, we can reject the null hypothesis, indicating that the drug significantly reduces the number of seizures. That means we must find the estimate of $\hat{\beta_2}$, and this topic will be covered in Section 5. The remainder of the text will detail our proposed methodology to test this particular research question.

4 Fitting the Model: EM Algorithm

The EM algorithm will be used to fit the Poisson model to the data. The complete data likelihood can be described as:

$$L_C(\beta_0, \beta, \sigma^2 | Y, X, Z) = \prod_{i=1}^{n} \frac{\mu_i \exp(-\sum_{j=1}^{m} y_{ij}/\mu_i)}{(\sum_{j=1}^{m} y_{ij})!} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{Z_i^2}{2\sigma^2}\right)$$

We can then compute the complete data log-likelihood as follows:

$$l_C(\beta_0, \beta, \sigma^2 | Y, X, Z) = log(L_C(\beta_0, \beta, \sigma^2 | Y, X, Z))$$

$$\begin{aligned} l_C(\beta_0,\beta,\sigma^2|Y,X,Z) &= \sum_{i=1}^{n}\Bigg(log(\mu_i) - log(\exp(\sum_{j=1}^{m} y_{ij}/\mu_i)) - log((\sum_{j=1}^{m} y_{ij})!) - log(\exp(z_i^2/2\sigma^2)) \\ &- \frac{1}{2}log(2\pi\sigma^2)\Bigg) \\ &= \sum_{i=1}^{n}\Bigg(\beta_0 + x_i\beta + Z_i - \sum_{j=1}^{m} y_{ij}\cdot\exp(-(\beta_0 + x_i\beta + Z_i)) - log((\sum_{j=1}^{m} y_{ij})!) - \frac{Z_i^2}{2\sigma^2} \\ &- \frac{1}{2}log(2\pi\sigma^2)\Bigg) \end{aligned}$$

Therefore, the simplified form of the complete data log-likelihood is:

$$\begin{aligned} l_C(\beta_0,\beta,\sigma^2|Y,X,Z) &= n\beta_0 + \sum_{i=1}^{n}(x_i\beta + Z_i) - e^{-\beta_0}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}e^{-x_i\beta}e^{-Z_i}) - \sum_{i=1}^{n} log((\sum_{j=1}^{m} y_{ij})!) \\ &- \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Z_i^2) - \frac{n}{2}log(2\pi\sigma^2) \end{aligned}$$

4.1 The E-Step

To obtain $Q(\beta_0,\beta,\sigma^2|\beta_0^{(t)},\beta^{(t)},\sigma^{2^{(t)}})$, we must find the expected value of the complete data log-likelihood:

$$Q(\beta_0,\beta,\sigma^2|\beta_0^{(t)},\beta^{(t)},\sigma^{2^{(t)}}) = E[l_C(\beta_0,\beta,\sigma^2|Y,X,Z)|\beta_0^{(t)},\beta^{(t)},\sigma^{2^{(t)}}]$$

After taking the expected value of each term in $l_C(\beta_0,\beta,\sigma^2|Y,X,Z)$, we get the following:

$$\begin{aligned} Q(\beta_0,\beta,\sigma^2|\beta_0^{(t)},\beta^{(t)},\sigma^{2^{(t)}}) &= n\beta_0 + \sum_{i=1}^{n} x_i\beta + E(Z_i|y_i) - e^{-\beta_0}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}e^{-x_i\beta}e^{-E(Z_i|y_i)}) \\ &- \sum_{i=1}^{n} log((\sum_{j=1}^{m} y_{ij})!) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} E(Z_i^2|y_i) - \frac{n}{2}log(2\pi\sigma^2) \end{aligned}$$

From the equation above, the expected value of Z_i must be calculated in order to be able to complete the E-step. The following expression is used to determine $E(Z_i|y_i)$:

$$E(Z_i|y_i) = \int_{-\infty}^{\infty} zf(z_i|y_i)dz$$

To find the density function $f(z_i|y_i)$ (i.e. the conditional probability of Z_i given y_i), we can use Bayes' theorem in the form:

$$f(z_i|y_i,x_i) = \frac{f(y_i|x_i,z_i)f(z_i|\sigma^2)}{\int_{-\infty}^{\infty} f(y_i|x_i,z_i)f(z_i|\sigma^2)} \tag{2}$$

It is quite challenging to compute the denominator in equation (2), as the integral does not have a closed form. However, we can approximate $E(Z_i|y_i)$ using a simulation technique known as rejection sampling. Rejection sampling is useful since the inverse of $f_{Z_i|y_i}(z_i|y_i)$ cannot be found. If this method of sampling is used, the integral portion in the denominator of equation (2) can also be ignored, as its value will be approximated in the constant used to perform rejection sampling. The steps are as follows:

1. Sample $Y_i \sim g(y)$ where $f_{Z_i|y_i}(z_i|y_i) \leq cg(z_i)$ at every point

2. Sample $U_i \sim U(0,1)$
3. Reject Y_i and return to step 1 if $U_i > \frac{f(Y_i)}{e(Y_i)}$ where $e(Y_i) = cg(Y_i)$
4. Otherwise set $Z_i = Y_i$

After running the rejection sampling algorithm, we will obtain a simulated set of z_i's in which we can calculate its mean value in order to compute $E(Z_i|y_i)$. Additionally, the second moment $E(Z_i^2|y_i)$ can also be determined from the simulated values. Although the integral lies between $-\infty$ to ∞, a reasonable range that encompasses most values will be used upon further analysis of the structure of the function. For easier notation, $E(Z_i|y_i)$ and $E(Z_i^2|y_i)$ will be denoted $Z_i^{(t)}$ in the following section.

4.2 The M-Step

Upon completing the E-step, the expected value of the complete data log-likelihood is now:

$$Q(\beta_0, \beta, \sigma^2|\beta_0^{(t)}, \beta^{(t)}, \sigma^{2^{(t)}}) = n\beta_0 + \sum_{i=1}^{n}(x_i\beta + Z_i^{(t)}) - e^{-\beta_0}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}e^{-x_i\beta}e^{-Z_i^{(t)}})$$

$$-\sum_{i=1}^{n} log((\sum_{j=1}^{m} y_{ij})!) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} Z_i^{(t)} - \frac{n}{2}log(2\pi\sigma^2)$$

$\frac{\partial Q}{\partial \beta_0}, \frac{\partial Q}{\partial \beta}$, and $\frac{\partial Q}{\partial \sigma^2}$ are found and set to 0 in order to find the value of $\beta_0^{(t+1)}$, $\beta^{(t+1)}$, and $\sigma^{2^{(t+1)}}$ that maximizes $Q(\beta_0, \beta, \sigma^2|\beta_0^{(t)}, \beta^{(t)}, \sigma^{2^{(t)}})$.

5 Maximum Likelihood Inference

5.1 Parameter Estimates

Once the expressions for the parameter estimates are obtained, we propose to run the EM algorithm in *R*. First, we must start with an initial set of values for $\beta_0^{(t)}$, $\beta^{(t)}$, and $\sigma^{2^{(t)}}$. Then, the EM algorithm will continue to iterate through the E-step and M-step, and recalculate the point estimates until the algorithm converges. Once it does, the last iteration will result in the final parameter estimates which will be closer to their true parameter values. After we have values for $\hat{\beta}_0$, $\hat{\beta}$, and $\hat{\sigma}^2$, we can compute the estimated rate of seizures μ_i using equation (1). The MLE estimates of $\beta_0^{(t)}$, and $\sigma^{2^{(t)}}$ are shown below:

$$\hat{\beta_0} = -log\left(\frac{-n}{\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij}e^{-x_i\beta})e^{-Z_i^{(t)}}}\right)$$

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n} Z_i^{(t)}}{n}$$

However maximizing $Q(\beta_0, \beta, \sigma^2|\beta_0^{(t)}, \beta^{(t)}, \sigma^{2^{(t)}})$ with respect to β cannot be done analytically. As such, the Newton-Raphson algorithm will be applied to $Q(\beta_0, \beta, \sigma^2|\beta_0^{(t)}, \beta^{(t)}, \sigma^{2^{(t)}})$ in order to calculate the value(s) of β that maximize the function.

5.2 Standard Errors

We know that Y_{ij} follows a Poisson distribution with the rate parameter μ_i. Traditionally, this would mean that the variance is also equal to μ_i, which is based on equation (1) in our case. However, due to the random effect Z_i, this is going to influence how the standard errors are calculated. Additionally, one of the drawbacks when using the EM algorithm is that it does not directly give us estimates for the standard errors of the sampling distribution. A way to work around this is to use parametric bootstrapping to approximate the standard errors of the point estimates. The steps are as follows:

1. Randomly simulate a set of $\hat{z}_i$ values using our computed MLE estimate $\hat{\sigma}^2$, in which we know that $Z_i \sim \mathrm{N}(0, \widehat{\sigma^2})$
2. Compute μ_i using equation (1) now that we have all the estimates $\hat{\beta}_0$, $\hat{\beta}$, and $\hat{z}_i$
3. Simulate an entirely new set of Y_{ij} values since we know that $Y_{ij} \sim \mathrm{Poisson}(\hat{\mu_i})$
4. Input the simulated data of Y_{ij} values into the EM algorithm where the point estimates can be returned and stored

The above process is replicated many times so that we can obtain a large sample of $\hat{\beta}_0$, $\hat{\beta}$, and $\hat{\sigma}^2$ values. The standard error for each set of point estimates can then be calculated accordingly.

5.3 Confidence Intervals

Once we find the standard errors via bootstrapping, they will be used to generate a 95% confidence interval. Since $\hat{\beta}_0$ and $\hat{\beta}$ follow a multivariate normal distribution, we may use the following formulas:

$$\text{C.I. for } \beta_0 = \hat{\beta}_0 \pm 1.96(SE(\hat{\beta}_0))$$
$$\text{C.I. for } \beta = \hat{\beta} \pm 1.96(SE(\hat{\beta}))$$

This is not the case for $\hat{\sigma}^2$, as it follows a Chi-squared distribution and we must compute it differently. We can use the formula below:

$$\text{C.I. for } \sigma^2 = \left[\frac{(n-1)\hat{\sigma}^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n-1}}\right]$$