KU LEUVEN

FACULTY OF SCIENCE

Assignment

Generalized Linear Models

Daniel Gerardo GIL SANCHEZ
daniel.gilsanchez@student.kuleuven.be

Prof. Emmanuel Lesaffre

Academic year 2017-2018

Renal dataset

Graft rejection is the mayor concern when a transplantation is conducted in any patient. It is believed that some specific characteristics of the patients may increase the likelihood of rejecting a new organ and so, it is of interest to understand which factors influence this rejection. For this reason, just before the transplant was conducted, some variables were collected from a sample of 1158 patients such as *age*, *sex*, whether the patient had *vascular problems* and some other variables were collected three months after the transplant, such as the *haematocrit level* and whether the patient presented *symptoms of graft failure*. The goal of this study is to reveal which are the key factors that determine the risk of graft rejection in renal transplantation. In the following, a descriptive analysis is performed to identify patterns in the data. Then, different binary regression models are conducted to establish the factors that impact the probability of present symptoms of graft failure.

From Table 1, it can be seen that the average *Haematocrit level (HC)* in the patients before the transplant is approximately 32%. The minimum and maximum *HC* level found are 14% and 60%, respectively. The patients are in *age* range from 15 to 76 years old, where the mean is about 46 years. The proportion of males is about 57%, whereas the proportion of females is around 43%. Almost 68% of them presented *symptoms of graft failure*. Finally, 82% of them presented *cardio-vascular problems* before the transplant.

Variable	**Description**	**Mean Frequency in %**	**Std. Deviation**	**Min/Max**
HC	Haematocrit level (%)	31.86	6.07	14/60
Age	Age in years	46.43	13.31	15/76
Reject	Symptoms of graft failure			
	0 = No	68.4		
	1 = Yes	31.6		
Male	Gender			
	0 = Female	42.7		
	1 = Male	57.3		
Cardio	Cardio-vascular problems			
	0 = No	82.1		
	1 = Yes	17.9		

Table 1: Descriptive statistics

From a bivariate perspective, the correlation coefficient between the continuous variables, *HC* and *Age*, is 0.20, indicating that the *Haematocrit level* is not greatly influenced by the *age* of the patient. Regarding the response variable, it seems that the patients that presented *symptoms of graft failure* are younger than the ones who do not, although it seems that this difference is not significant. On the other hand, it can be seen that there are not differences in *HC* levels between the patients that presented *symptoms of graft failure* and the ones who do not (see Figure 1).

Similar behavior can be found in the categorical variables, where the proportion of patients that presented *symptoms of graft failure* is practically the same for both men

and women, 34% and 29% respectively, as well as for the patients that had *cardiovascular problems* or not, 32% and 28% respectively. Therefore, these results may suggest that these factors do not influence the rejection of the organ.

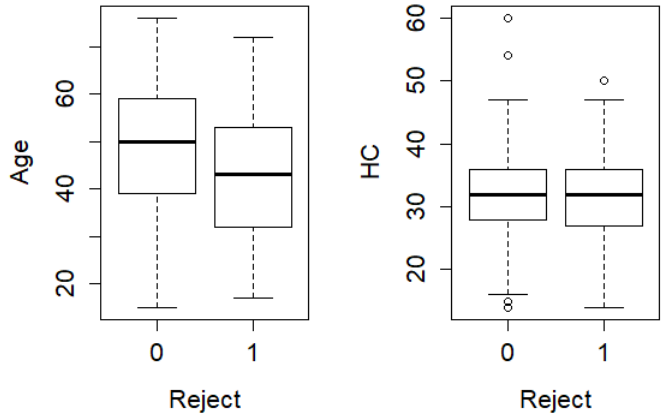

Figure 1: Boxplot of *Age* and *HC*, by *symptoms of graft failure*.

So, in order to evaluate the significance of each factor and to be sure about the findings in the descriptive section, different binary regression models are conducted. These models are performed in the statistical software R using the stats package to fit the frequentist models and the MCMCpack package to fit the Bayesian models.

It is important to mention that before doing the models, the continuous variables, *HC* and *Age*, are centered around the mean for two reasons. First, to be able to interpret the intercept (β_0) when all the covariates are equal to zero. And second, to be able to add polynomial terms in the model without introducing multicollinearity.

As it was mentioned before, different binary models using different link functions are used to fit the models, namely logit, probit, cauchit and complementary log-log (cloglog). Hence, the process done to choose the final model consists in select the best model for each link function via Likelihood Ratio Test (LRT), because all of these models are nested. Then, the choice of the best link function is done via AIC measure, because these models are not nested anymore.

Thus, the process to select the best model in each link function begins with the null model. In this case and from now on, the logit function is used. This model is useful because it can be used to compare complex models giving the overall significance of the model[1]. Then, a main effects model is fitted using all covariates, i.e., *age*, *male*, *cardio* and *HC*. As a result, the comparison between the null model and the main effects model is significant using LRT[2]. This means that the added terms improve the fit of the model compared with a model that only has the intercept. However, the Wald tests, used to evaluate whether the estimated coefficient for each variable is significantly different from zero, indicate that only *Age* is significant. This is an expected result as it was shown in the descriptive section, where no large differences were shown in each covariate.

Nevertheless, variable selection procedures are performed to be completely sure about these findings. To do so, two functions are used in R, namely anova and drop1. The first function begins with the null model and in each step adds only one variable, then the significance of this new model is evaluated comparing it with the model in the previous step via LRT. The second function starts with the full model and evaluate the significance of each variable by removing it from the model and comparing it with the

[1]This procedure can be thought as equivalent to the overall significance of a model in Ordinary Least Squares, when the null hypothesis is that all parameters are equal to zero simultaneously.

[2]In this study the significance level is set at 5%.

full model also using LRT[3]. The result of both functions is that only *age* seems to be significant.

Given these results, different attempts are conducted to improve the fit of the model. So it is considered to add interactions and polynomial terms to the model. To do so, add1 is a function that adds an extra term to the model and makes a comparison of the simpler model with the new one via LRT, just like the previous functions mentioned. The first attempt is to add second-order interactions of each pair of covariates in the model, resulting in no significant differences. The second attempt is to add squared terms of *age* and *HC*, one at a time, resulting also in no significant differences.

Now, this add1 function does not allow to add two variables simultaneously, so a new model is performed adding both polynomial terms in the model. As a result, the deviance decreases less than 1 unit indicating that no improvement has been achieved. Therefore, a new model with second-order terms is fitted, i.e., including all second-order interactions and squared polynomials. Again, there is no significant improvement with all these terms. After all these attempts, it is very clear that only *age* is significant in the model, which is an expected result given the findings in the descriptive part. So it can be concluded that the final model for the logit link function is the one that only has *age* as a covariate.

This process is repeated again using the remain link functions: probit, cauchit and cloglog. First the null model is fitted, then a main effect model is fitted and several attempts are performed to test interactions and polynomials one at a time. Next, a model with squared polynomials added simultaneously is fitted. Then, a model with second-order terms added simultaneously is fitted. Finally, a model with only *age* as covariate is fitted.

As a consequence, in all link functions the final model is the one that only has the covariate *age*. Thus, the choice of the final model is influenced by the AIC measure[4]. In Table 2 the AIC for each model fitted is presented.

Model	Logit	Probit	Cauchit	Cloglog
Null	1446.87	1446.87	1446.87	1446.87
Main effects	1407.89	1407.76	1409.14	1408.1
Polynomial terms	1410.55	1410.64	1410.24	1410.37
Second-order terms	1417.37	1417.31	1417.63	1417.41
Only age effect	**1405.7**	**1405.52**	1407.28	1406.17

Table 2: AIC of each model fitted

In the previous table can be seen that the AIC of the last model is smaller in all link functions. In detail, it can be seen that the probit and cauchit models have the smaller and large AIC, respectively. However, according to the rule of thumb that a

[3]It is important to mention that both functions are sensitive to the order in which the covariates are written in the glm function. So a model with *age* and *male* could lead to different results than a model with *male* and *age*.

[4]BIC can also be used to compare non-nested models. According to Fahrmeir et al., 2013, the main difference between AIC and BIC is that the later penalizes complex models much more than the AIC. Usually, the best model resulting from BIC is more parsimonious than using AIC. In this case, because the model to compare only has one covariate, there is no difference between both criteria.

model is considered to be better if the difference in AIC is larger than five units, it can be concluded that all link functions are very similar in terms of fit. For this reason and because of interpretation of the estimates is straightforward[5], the logit link function is chosen as the final model.

Therefore the final model is, as follows[6]:

$$P(reject = 1) = \frac{exp(\beta_0 + \beta_1 Agec)}{1 + exp(\beta_0 + \beta_1 Agec)} \Leftrightarrow \frac{P(reject)}{P(no\ reject)} = exp(\beta_0) exp(\beta_1 Agec)$$

The estimation of these parameters is $\beta_0 = -0.80$ and $\beta_1 = -0.03$, indicating that the odds of presenting *symptoms of graft failure* when a patient has the average age, which in this case is 46 years, is $exp(-0.8) = 0.45$. And an increment of one year in the *age* of a patient is associated with a change in the odds of presenting *symptoms of graft failure* by a factor of $exp(-0.03) = 0.97$. This actually indicates that the older the patient, the less likely it is to reject the new organ (see Figure 2).

In this perspective, it is necessary to assess the fit and the predictive quality of the model, via goodness of fit measures, to be completely sure about the findings.

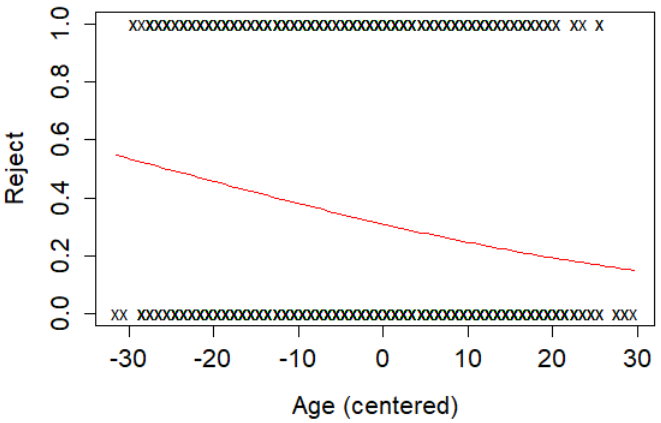

Figure 2: Fit of logit model

To assess the quality of the model, Hosmer-Lemeshow statistic is used for a simple reason. According to Lesaffre, 2017/2018, all goodness of fit test for a logistic regression involves comparing observed with estimated frequencies. This is easily performed when the covariates are categorical and, Pearson and Deviance statistics are built to deal with this kind of covariates. In this case, the covariate *age* is continuous, so a different approach is needed, as Hosmer-Lemeshow does[7]. Here, the null hypothesis is that the observed and expected proportions are the same across all groups formed. In this sense, ten groups are formed and the test is performed, indicating that the null hypothesis is not rejected (P-value = 0.77). So it can be concluded that the model fits the data well.

On the other hand, the prediction quality can be assessed via Nagelkerke's R^2 and Concordance measures. According to ibid. for a perfect binary regression model, if all predicted values are close to 1 (0) for a success (failure), then the Nagelkerke's R^2 is closer to 1. Regarding the Concordance measure, the higher the percent of concordance pair the better the model is[8]. In this case, Nagelkerke's R^2 is 0.05 and the

[5]It is important to notice that the estimates from a logit and probit models are essentially the same, up to a constant (Fahrmeir et al., 2013).

[6]*Agec* stands for the covariate *age* after being centered around its mean.

[7]Hosmer-Lemeshow test is similar to Pearson or Deviance test in the sense that compares observed with estimated frequencies. The difference is how it deals with continuous covariates, it divides the data into groups based on having similar predicted probabilities Lesaffre, 2017/2018.

[8]A concordance measure close to 50% indicates that a model is predicting as if it were just by chance.

Concordance measure is 61%, so it can be concluded even though the model fits the data well is not good to predict.

In addition to quality assessment of the model, it is also imperative to identify possible outliers or influential observations that could lead to a miss-specification of the model. To identify outliers in the dataset Pearson and Deviance diagnostics are considered. Thus, each residual is calculated from the model and then these are plotted against the covariate *age* to see what pattern appears, in Figure 3 can be seen the deviance residual plot. It is important to mention that a smoothing technique is necessary because these residuals have a bimodal distribution (represented as blue line). Now, given that the trend is just a line, it can be concluded that there are not outliers in the dataset[9].

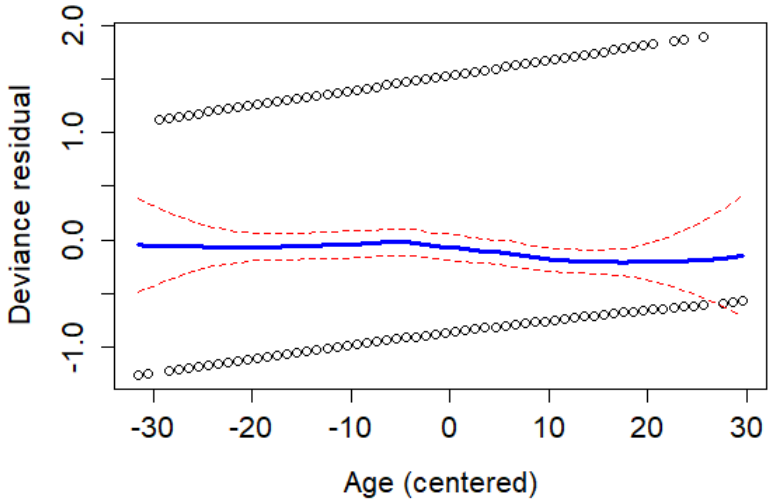

Figure 3: Outliers: *Age* vs Deviance residuals

On the other hand, the influential observations can be identified by using the standardized residuals (Pearson or Deviance), hat values, Dffits, Cook's distance, among others. As a result, two observations are identified by Cook's distance and other two observations are identified by the hat values. Hence, in order to evaluate whether these observations are indeed influential, four different models are fitted excluding each of the observations. Next, the estimates of the coefficients are compared to see whether there are large changes. As a result, none of these observations are considered to be influential, since the change in the estimates is really small.

It is also of interest to fit the final model from a Bayesian perspective, thereby prior distributions are needed for each coefficient. According to Lesaffre, 2017/2018 a common choice for the prior is to take a normal distribution with mean 0 and large variance. However, this kind of priors may lead to unrealistic results. Therefore, a reasonable prior for the coefficients is to have a precision 100 times smaller than the precision from the frequentist approach. In Table 3 the comparison between both approaches is shown[10].

[9]The same analysis is conducted with Pearson residuals but the result is not shown because it is similar and leads to the same conclusion

[10]To obtain these results, 11 thousand samples are taken from the posterior distribution of each coefficient, from which the first thousand belong to the burn-in period and so, they are not taken into account in the posterior measures.

	Frequentist		Bayesian	
	Estimate	**Confidence Int.**	**Estimate**	**Credible Int.**
Intercept	-0.8028	[-0.9315,-0.6763]	-0.7937	[-0.9203,-0.6698]
Age	-0.0315	[-0.0411,-0.0220]	-0.0311	[-0.0408,-0.02129]

Table 3: Estimation comparison

In the comparison can be seen that the results from both approaches are pretty similar, actually the length of the intervals differ until the fourth decimal. This outcome is expected since the frequentist results are used to construct the prior for the Bayesian approach. Regarding the interpretation of the estimates, it changes a little bit because in the Bayesian approach the results reflect the posterior distribution of the estimates.

To sum up, several binary regression models are fitted to reveal the influence of the factors in the presence of *symptoms of graft rejection* after the transplantation. Four different link functions are used: logit, probit, cauchit and complementary log-log. In all of them the only covariate that resulted to be significant is *age*. Hosmer-Lemeshow test indicates that the model fits well the data, but the predictive quality is not good enough, according the Nagelkerke's R^2 and the Concordance measure. Using Pearson and Deviance residuals, hat values, Cook's distance and other diagnostics measures, none of the observations are found to be outliers or influential. The final model is also fitted from a Bayesian approach leading to similar results. So the final model leads to conclude that neither the *Haematocrit level*, nor the *gender* of the patient nor whether the patient presented *cardio-vascular problems* are factors that may increase the probability of rejecting the new organ after transplantation. *Age* is the only factor that influences the outcome, where the older the patient is, the less likely it is to present *symptoms of graft failure*.

Epilepsy dataset

Epilepsy is a chronic disorder of the brain that affects people worldwide. It is characterized by recurrent seizures, which are brief episodes of involuntary movement that may involve a part of the body or the entire body (WHO, 2018). A new drug, called *Progabide*, has been developed to treat this disorder. So in order to evaluate whether the new treatment is more effective than the actual, a randomized clinical trial was conducted. The experiment involved 59 patients suffering from epilepsy, where they were randomized to receive either the new drug or a placebo in addition to the standard chemotherapy. The measure to compare the treatments is the *number of seizures* in the fourth treatment period, i.e., between weeks 7 and 8. The *age* of the patient was also collected to be able to control the possible differences between the groups. The goal of this study is to model the expected *number of seizures* as a function of the covariates *treatment* and *age* to reveal which treatment is more effective. In the following, a descriptive analysis is performed to identify patterns in the data. Then, different Poisson regression models are conducted to establish the possible differences between treatments.

From Table 4, it can be seen that the average *number of seizures* in the patients in the fourth treatment period is approximately 7. The minimum and maximum *number of*

seizures found are 0 and 63, respectively. The patients are in age range from 18 to 42 years old, where the mean is about 28 years. The proportion of patients in the placebo group is about 47%, whereas the proportion of patients in the new drug group is around 53%. It is important to notice the difference between the average *number of seizures* and its variance. This large difference may indicate overdispersion in the data, since it is hypothesized that a count follows a Poisson distribution[11].

Variable	**Description**	**Mean Frequency in %**	**Std. Deviation**	**Min/Max**
Seizure.rate	Number of seizures	7.31	9.65	0/63
Age	Age in years	28.34	6.3	18/42
Treatment	Treatment			
	0 = Placebo	47.5		
	1 = Progabide	52.5		

Table 4: Descriptive statistics

From a bivariate perspective, the correlation coefficient between the response variable and the covariate *age* is -0.08, indicating that the *number of seizures* is not greatly influenced by the *age* of the patient. Regarding the *treatment* variable, it seems that the patients that belong to the group of Progabide are younger than the ones who belong to the Placebo group, although it seems that this difference is not significant. On the other hand, it can be seen that the distribution of the *number of seizures* is skewed to the right, as it is expected in a variable that represents counts of an event (see Figure 4).

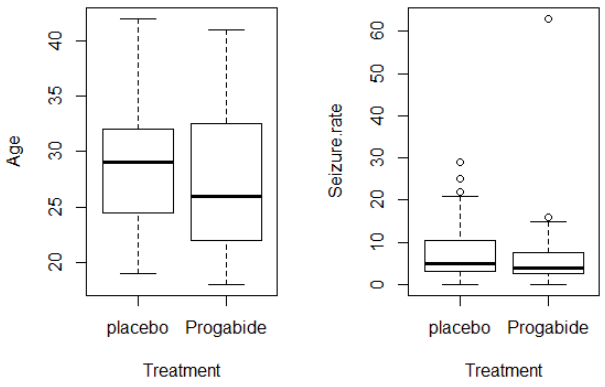

Figure 4: Boxplot of *Age* and *Number of seizures*, by *treatment*.

So, in order to evaluate the significance of each covariate and given that the response variable is a count as well as each subject has the same length of observation time, a Poisson regression model is conducted using the logarithmic function as a link between the expected value of the *number of seizures* and the linear predictor. This model is conducted in the statistical software R using the stats package to fit the frequentist models and the MCMCpack package to fit the Bayesian model. It is important to mention that before doing the model, the variable *age* is centered around its mean, to be able to interpret the intercept (β_0) when all the covariates are equal to zero, and the reference category for *treatment* is placebo.

Therefore, the process done to select the final model consists in select the best model via Likelihood Ratio Test (LRT), because all of these models are nested. Ac-

[11]It is important to mention that in a Poisson distribution, the expected value and the variance are the same. In this case, the fact that they are not close can lead to indicate overdispersion, even though this is just a rule of thumb because these measures do not take into account the covariates.

cording to Fahrmeir et al., 2013, the LRT is useful when the hypotheses are linear, like this: $H_0 : C\beta = d$ versus $H_1 : C\beta \neq d$ with C having full rank row $r \leq p$. In the special case of comparison between nested models, Lesaffre, 2017/2018 defines in the null hypothesis the model that has a subset of parameters of the model that belongs to the alternative hypothesis ($H_0 \subset H_1$), and consequently defines the maximized log-likelihoods under each hypothesis as $l^{(0)}(\beta|X)$ and $l^{(1)}(\beta|X)$. The test-statistic is then defined as: $lr = -2\left\{l^{(0)}(\beta|X) - l^{(1)}(\beta|X)\right\}$, which follows a χ^2 distribution with r degrees of freedom[12]. The implication of a significant difference is that the complex model fits the data in a better way than the simpler one.

Thus, the process to select the best model begins with the null model. This model is useful because it can be used to compare complex models giving the overall significance of the model[13]. Then, a main effects model is fitted using all covariates, i.e., *age* and *treatment*. As a result, the comparison between the null model and the main effects model is significant using LRT[14]. This means that the added terms improve the fit of the model compared with a model that only has the intercept. In addition, the Wald tests, used to evaluate whether the estimated coefficient for each variable is significantly different from zero, indicate that both variables are significant.

Given these results and the context of the study, it is natural to think in an interaction between the covariates, because the efficacy of Progabide could be influenced by the age of the patient. Consequently, a new model is fitted adding the interaction term and it is compared with the main effects model, see Table 5.

Model	**Df**	**Deviance**	**LRT**	**Pr(>Chi)**
Main effects	56	467.77		
Main effects + Interaction	55	447.34	20.431	¡0.0001

Table 5: Comparison between nested models

In the previous table, the statistical main effects model is[15]

$$log(Seizure.rate) = \beta_0 + \beta_1 Agec + \beta_2 Treatment$$

and the main effect + Interaction model is

$$log(Seizure.rate) = \beta_0 + \beta_1 Agec + \beta_2 Treatment + \beta_{12} Agec * Treatment$$

It can be easily seen that the difference in the number of parameters is just one, so the test statistic is compared to a χ^2 distribution with one degree of freedom. The result is significant, so the final model is the one that has the interaction.

The estimation of these parameters is $\beta_0 = 2.06$, $\beta_1 = 0.02$, $\beta_2 = -0.24$ and $\beta_{12} = -0.07$, indicating that the expected *number of seizures* when a patient has the average age, which in this case is 28 years, and belongs to the placebo group is $exp(2.06) = 7.82$. Furthermore, the expected *number of seizures* when a patient has the average age and belongs to the Progabide group is $exp(2.06 - 0.24) = 6.13$. Now, for a patient

[12]This r is actually the difference between the number of parameters between both models.

[13]This procedure can be thought as equivalent to the overall significance of a model in Ordinary Least Squares, when the null hypothesis is that all parameters are equal to zero simultaneously.

[14]In this study the significance level is set at 5%.

[15]*Agec* stands for the covariate *age* after being centered around its mean.

that belongs to the placebo group, an increment of one year in its *age* is associated with an increment in its expected *number of seizures* by a factor of $exp(0.02) = 1.02$. And for a patient that belongs to the Progabide group, an increment of one year in *age* is associated with a change in its expected *number of seizures* by a factor of $exp(0.02 - 0.07) = 0.95$. This actually indicates that the expected *number of seizures* in a patient in the placebo group increases as it gets older, whereas the expected *number of seizures* in a patient in the new treatment decreases. For example, the expected *number of seizures* of a 30 years old patient that is in the placebo group is 8.06, while if the same patient is in the new treatment group its expected *number of seizures* is 5.60.

In this perspective, it is necessary to assess the fit of the model. In fact, one of the issues that often happens in Poisson regression is overdispersion. An easy way to detect overdispersion in the model is by looking the residual deviance and degrees of freedom, the closer they are the better. However, in this case the residual deviance is 447.34 and the degrees of freedom are 55, so the assumption that the mean and the variance are equal does not hold in this case.

One possible solution is to use the Negative Binomial regression because it has the same structure as Poisson regression and it has an extra parameter to model overdispersion. A graphical comparison between the Poisson and Negative Binomial distributions can be done fitting a null model and plotting each result over a histogram of the response variable. In Figure 5 can be seen that the Poisson distribution is not fitting the data well specially in the low frequencies. Actually the observed maximum is near 4 *number of seizures*, whereas the maximum of the distribution is near 8. On the other hand, the Negative Binomial distribution seems to improve the fit specially in the tail of the distribution.

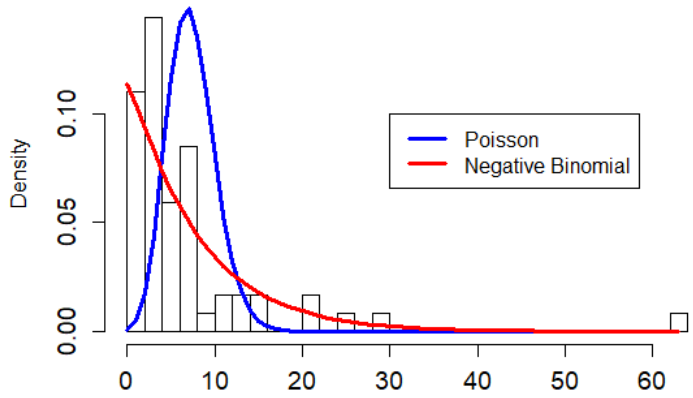

Figure 5: Comparison between Poisson fit and Negative Binomial fit

Consequently, the main effect + interaction model is fitted using the Negative Binomial distribution. As a result, the residual deviance and the degrees of freedom are much closer compared to the previous model, 67.02 and 55 respectively. The AIC of the Poisson model is 643.11, whereas the AIC of the Negative Binomial is 365.71, indicating that the later model is much better than the previous one.

In addition to this result, the statistical significance of the covariates has changed a lot, now only the intercept is significant. This can be explained by the distribution part of the model: usually when the distribution part of the model is miss-specified, the

systematic part compensates the lack of fit of the model. In this case, given that the distribution part fits the data in a better way, the covariates are not useful anymore. The implication of these results is that it seems that the new drug does not make any difference in the treatment of epilepsy.

For this reason and because we are only interested in the effect of the covariates on the expected *number of seizures* using a correct variance function, a quasi-poisson model is fitted. To do so, the logarithmic function is again used as a link between the response variable and the linear predictor. In addition, to obtain robust standard errors, the sandwich estimator is used for the covariance matrix. As a consequence, the final estimates are not significant, indicating once again that Progabide is not more effective than the standard therapy.

Moreover, it is of interest to fit the final model from a Bayesian perspective. However, given that the final model consists in a quasi-likelihood and according to Lesaffre, 2017/2018, it does not exist a standard Bayesian analogue because this approach requires a likelihood to obtain any kind of estimate, is not possible to do it.

In conclusion, different models are fitted to reveal which drug is better in the treatment of epilepsy. Poisson regression is used to fit a main effects model. Then, using Likelihood Ratio Test the interaction between *age* and *treatment*, lead to conclude that the efficacy of the treatment also depends on how old the patient is. However, when the fit of the model is assessed, overdispersion is found in the model, making the conclusion unreliable. To deal with this problem a Negative Binomial model is fitted because it has an extra parameter to model the overdispersion. The conclusion of this model is that none of the variables is significant, indicating that the covariates do not explain the variability of the *number of seizures*. As a final step, a quasi-Poisson regression model is fitted using the sandwich estimator for the covariance matrix. The final results are similar to the Negative Binomial. In addition, it is not possible to fit the final model in a Bayesian approach because an appropriate likelihood is not found in the final model. So the final result leads to conclude that neither *age* nor *treatment* are factors that influence the expected *number of seizures*, and so the new drug, Progabide, is not more effective than the standard treatment.

References

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, methods and applications*. Springer Berlin Heidelberg.

Lesaffre, E. (2017/2018). Generalized linear models - course notes. KU Leuven.

WHO. (2018). Epilepsy fact sheet. Retrieved from `http://www.who.int/en/news-room/fact-sheets/detail/epilepsy`

Renal dataset - R Code

```r
##   Clear workspace
rm(list=ls())

##   Set directory
setwd("...")

##   Load packages
library(readxl) #   To import excel files
library(ggplot2)  #   plots
library(generalhoslem)  #   To Hosmer-Lemeshow GoF statistic (logitgof)
library(fmsb)   #   To calculate Nagelkerke R^2
library(MCMCpack)   #   To conduct bayesian logistic regression

##   Read data
renal = read_excel("renal.xlsx")
head(renal)
class(renal)
str(renal)

#   Setting dummies as factors
renal$male = as.factor(renal$male)
renal$cardio = as.factor(renal$cardio)
renal$reject = as.factor(renal$reject)
# str(renal)

#   Checking reference category
contrasts(renal$male)
contrasts(renal$cardio)
contrasts(renal$reject)

#   Centered and squared variables for polynomial
renal$age_c = renal$age - mean(renal$age)
renal$HC_c = renal$HC - mean(renal$HC)
renal$age_c_sq = renal$age_c^2
renal$HC_c_sq = renal$HC_c^2

##   Descriptive statistics
#   General
summary(renal)

#   Amount of NA in each column
sapply(renal,function(x) sum(is.na(x)))

#   Amount of unique values in each column
sapply(renal, function(x) length(unique(x)))

#   Univariate analysis
#   Frequency of each dummy
prop.table((table(renal$reject)))
prop.table((table(renal$male)))
prop.table((table(renal$cardio)))

#   Continous
```

```
summary(renal[,c(2,6)])
sd(renal$age);sd(renal$HC)
hist(renal$age);plot(density(renal$age))
hist(renal$HC);plot(density(renal$HC))
boxplot(renal$age)
boxplot(renal$HC)

#   Bivariate analysis
cor(renal[,c(2,6)])
cor(renal[,c(7,9,8,10)])
#   Reject(rows) vs Male(Cols)
cbind(table(renal$reject,renal$male),
      prop.table(table(renal$reject,renal$male),1),
      prop.table(table(renal$reject,renal$male),2))
#   Reject(rows) vs Cardio(Cols)
cbind(table(renal$reject,renal$cardio),
      prop.table(table(renal$reject,renal$cardio),1),
      prop.table(table(renal$reject,renal$cardio),2))
#   3 way contigency table
tab = xtabs(~reject+male+cardio,data=renal);tab

#   Plots
par(mfrow=c(1,2))
boxplot(renal$age~renal$reject, xlab="Reject",
 ↪  ylab="Age",cex.lab=1.3,cex.axis=1.3)
boxplot(renal$HC~renal$reject, xlab="Reject", ylab="HC",cex.lab=1.3,cex.axis=1.3)
par(mfrow=c(1,1))
boxplot(renal$age_c~renal$reject)
boxplot(renal$HC_c~renal$reject)
boxplot(renal$age_c_sq~renal$reject)
boxplot(renal$HC_c_sq~renal$reject)
plot(renal$age,renal$reject)
plot(renal$HC,renal$reject)

# Show graphically dependence of p(reject) on Age
# qplot(age,reject,data=renal,geom=c("point","smooth"),span=0)
# qplot(age,reject,data=renal,geom=c("point","smooth"),
      # method="lm")

##  Models
##  Null model
logit_null = glm(reject~1,data=renal,family = binomial(link='logit'))

##  Logit: main effects model
head(renal)
logit = glm(reject~age_c+male+cardio+HC_c,data=renal,
            family=binomial(link='logit'));logit
logit_sum = summary(logit);logit_sum

#   LR test
logit$null.deviance-logit$deviance
1 - pchisq(logit$null.deviance-logit$deviance,logit$df.null-logit$df.residual)

#   This give me the same result:
anova(logit_null,logit,test="Chisq")
```

```r
#   Variable selection
anova(logit, test="Chisq")  #   Drawback: order matters
drop1(logit, test="Chisq")

#   Interaction terms
add1(logit,~.+male*cardio+age_c*male+age_c*cardio+HC_c*male+HC_c*cardio+
     age_c*HC_c,test='LRT')

#   No improvement adding each term squared
add1(logit,~.+age_c_sq+HC_c_sq,test='LRT')

#   logit2: With both polynomial terms simultaneously
logit2 = glm(reject~age_c+age_c_sq+male+cardio+HC_c+HC_c_sq,data=renal,
             family=binomial(link='logit'));logit2
logit2_sum = summary(logit2);logit2_sum

#   There is not significant improvement of the model
anova(logit,logit2,test='Chisq')

#   logit3: second order terms
logit3 = glm(reject~age_c+age_c_sq+male*cardio+HC_c+HC_c_sq+age_c*male+
             age_c*cardio+HC_c*male+HC_c*cardio,data=renal,
             family=binomial(link='logit'));logit3
logit3_sum = summary(logit3);logit3_sum

#   There is not significant improvement of the model
anova(logit,logit3,test='Chisq')

#   Logit4: with only age
logit4 = glm(reject~age_c,data=renal,
               family=binomial(link='logit'));logit4
logit4_sum = summary(logit4);logit4_sum
anova(logit,logit4,test="Chisq")

#   LR test
logit4$null.deviance-logit4$deviance
1 - pchisq(logit4$null.deviance-logit4$deviance,logit4$df.null-logit4$df.residual)
anova(logit_null,logit4,test="Chisq")

#   Model with only age centered, adding the squared term.
add1(logit4,~.+age_c_sq,test='LRT')
logit5_aic = add1(logit4,~.+age_c_sq,test='LRT')$AIC[2];logit5_aic

#   So far, logit4, model with only age, is significant and is the final model
#   using logit link


##  Probit
##  Null model
probit_null = glm(reject~1,data=renal,family = binomial(link='probit'))

##  probit: main effects model
head(renal)
probit = glm(reject~age_c+male+cardio+HC_c,data=renal,
```

```
             family=binomial(link='probit'));probit
probit_sum = summary(probit);probit_sum

#   LR test
probit$null.deviance-probit$deviance
1 - pchisq(probit$null.deviance-probit$deviance,probit$df.null-probit$df.residual)

#   This give me the same result:
anova(probit_null,probit,test="Chisq")

#   Variable selection
anova(probit, test="Chisq")  #   Drawback: order matters
drop1(probit, test="Chisq")

#   Interaction terms
add1(probit,~.+male*cardio+age_c*male+age_c*cardio+HC_c*male+HC_c*cardio+
         age_c*HC_c,test='LRT')

#   No improvement adding each term squared
add1(probit,~.+age_c_sq+HC_c_sq,test='LRT')

#   probit2: With both polynomial terms simultaneously
probit2 = glm(reject~age_c+age_c_sq+male+cardio+HC_c+HC_c_sq,data=renal,
              family=binomial(link='probit'));probit2
probit2_sum = summary(probit2);probit2_sum

#   There is not significant improvement of the model
anova(probit,probit2,test='Chisq')

#   probit3: second order terms
probit3 = glm(reject~age_c+age_c_sq+male*cardio+HC_c+HC_c_sq+age_c*male+
                  age_c*cardio+HC_c*male+HC_c*cardio,data=renal,
              family=binomial(link='probit'));probit3
probit3_sum = summary(probit3);probit3_sum

#   There is not significant improvement of the model
anova(probit,probit3,test='Chisq')

#   probit4: with only age
probit4 = glm(reject~age_c,data=renal,
              family=binomial(link='probit'));probit4
probit4_sum = summary(probit4);probit4_sum
anova(probit,probit4,test="Chisq")

#   LR test
probit4$null.deviance-probit4$deviance
1 -
 ↪  pchisq(probit4$null.deviance-probit4$deviance,probit4$df.null-probit4$df.residual)
anova(probit_null,probit4,test="Chisq")

#   Model with only age centered, adding the squared term.
add1(probit4,~.+age_c_sq,test='LRT')
probit5_aic = add1(probit4,~.+age_c_sq,test='LRT')$AIC[2];probit5_aic

#   So far, probit4, model with only age, is significant and is the final model
```

```
#   using probit link


##  Cauchy
##  Null model
cauchit_null = glm(reject~1,data=renal,family = binomial(link='cauchit'))

##  cauchit: main effects model
head(renal)
cauchit = glm(reject~age_c+male+cardio+HC_c,data=renal,
              family=binomial(link='cauchit'));cauchit
cauchit_sum = summary(cauchit);cauchit_sum

#   LR test
cauchit$null.deviance-cauchit$deviance
1 -
 ↪  pchisq(cauchit$null.deviance-cauchit$deviance,cauchit$df.null-cauchit$df.residual)

#   This give me the same result:
anova(cauchit_null,cauchit,test="Chisq")

#   Variable selection
anova(cauchit, test="Chisq")  #   Drawback: order matters
drop1(cauchit, test="Chisq")

#   Interaction terms
add1(cauchit,~.+male*cardio+age_c*male+age_c*cardio+HC_c*male+HC_c*cardio+
         age_c*HC_c,test='LRT')

#   No improvement adding each term squared
add1(cauchit,~.+age_c_sq+HC_c_sq,test='LRT')

#   cauchit2: With both polynomial terms simultaneously
cauchit2 = glm(reject~age_c+age_c_sq+male+cardio+HC_c+HC_c_sq,data=renal,
               family=binomial(link='cauchit'));cauchit2
cauchit2_sum = summary(cauchit2);cauchit2_sum

#   There is not significant improvement of the model
anova(cauchit,cauchit2,test='Chisq')

#   cauchit3: second order terms
cauchit3 = glm(reject~age_c+age_c_sq+male*cardio+HC_c+HC_c_sq+age_c*male+
                   age_c*cardio+HC_c*male+HC_c*cardio,data=renal,
               family=binomial(link='cauchit'));cauchit3
cauchit3_sum = summary(cauchit3);cauchit3_sum

#   There is not significant improvement of the model
anova(cauchit,cauchit3,test='Chisq')

#   cauchit4: with only age
cauchit4 = glm(reject~age_c,data=renal,
               family=binomial(link='cauchit'));cauchit4
cauchit4_sum = summary(cauchit4);cauchit4_sum
anova(cauchit,cauchit4,test="Chisq")
```

```
#   LR test
cauchit4$null.deviance-cauchit4$deviance
1 -
 ↪  pchisq(cauchit4$null.deviance-cauchit4$deviance,cauchit4$df.null-cauchit4$df.residual)
anova(cauchit_null,cauchit4,test="Chisq")

#   Model with only age centered, adding the squared term.
add1(cauchit4,~.+age_c_sq,test='LRT')
cauchit5_aic = add1(cauchit4,~.+age_c_sq,test='LRT')$AIC[2];cauchit5_aic

#   So far, cauchit4, model with only age, is significant and is the final model
#   using cauchit link


##  complementary Log-Log
##  Null model
cloglog_null = glm(reject~1,data=renal,family = binomial(link='cloglog'))

##  cloglog: main effects model
head(renal)
cloglog = glm(reject~age_c+male+cardio+HC_c,data=renal,
              family=binomial(link='cloglog'));cloglog
cloglog_sum = summary(cloglog);cloglog_sum

#   LR test
cloglog$null.deviance-cloglog$deviance
1 -
 ↪  pchisq(cloglog$null.deviance-cloglog$deviance,cloglog$df.null-cloglog$df.residual)

#   This give me the same result:
anova(cloglog_null,cloglog,test="Chisq")

#   Variable selection
anova(cloglog, test="Chisq")  #   Drawback: order matters
drop1(cloglog, test="Chisq")

#   Interaction terms
add1(cloglog,~.+male*cardio+age_c*male+age_c*cardio+HC_c*male+HC_c*cardio+
         age_c*HC_c,test='LRT')

#   No improvement adding each term squared
add1(cloglog,~.+age_c_sq+HC_c_sq,test='LRT')

#   cloglog2: With both polynomial terms simultaneously
cloglog2 = glm(reject~age_c+age_c_sq+male+cardio+HC_c+HC_c_sq,data=renal,
               family=binomial(link='cloglog'));cloglog2
cloglog2_sum = summary(cloglog2);cloglog2_sum

#   There is not significant improvement of the model
anova(cloglog,cloglog2,test='Chisq')

#   cloglog3: second order terms
cloglog3 = glm(reject~age_c+age_c_sq+male*cardio+HC_c+HC_c_sq+age_c*male+
                   age_c*cardio+HC_c*male+HC_c*cardio,data=renal,
               family=binomial(link='cloglog'));cloglog3
```

```r
cloglog3_sum = summary(cloglog3);cloglog3_sum

#   There is not significant improvement of the model
anova(cloglog,cloglog3,test='Chisq')

#   cloglog4: with only age
cloglog4 = glm(reject~age_c,data=renal,
               family=binomial(link='cloglog'));cloglog4
cloglog4_sum = summary(cloglog4);cloglog4_sum
anova(cloglog,cloglog4,test="Chisq")

#   LR test
cloglog4$null.deviance-cloglog4$deviance
1 -
 ↪  pchisq(cloglog4$null.deviance-cloglog4$deviance,cloglog4$df.null-cloglog4$df.residual)
anova(cloglog_null,cloglog4,test="Chisq")

#   Model with only age centered, adding the squared term.
add1(cloglog4,~.+age_c_sq,test='LRT')
cloglog5_aic = add1(cloglog4,~.+age_c_sq,test='LRT')$AIC[2];cloglog5_aic

#   So far, cloglog4, model with only age, is significant and is the final model
#   using cloglog link


#   AIC comparison:
aic_comp = cbind(logit = c(logit_null$aic,logit_sum$aic,logit2_sum$aic,
                           logit3_sum$aic,logit4_sum$aic),
      probit = c(probit_null$aic,probit_sum$aic,probit2_sum$aic,probit3_sum$aic,
                 probit4_sum$aic),
      cauchit = c(cauchit_null$aic,cauchit_sum$aic,cauchit2_sum$aic,
                  cauchit3_sum$aic,cauchit4_sum$aic),
      cloglog = c(cloglog_null$aic,cloglog_sum$aic,cloglog2_sum$aic,
                  cloglog3_sum$aic,cloglog4_sum$aic))
rownames(aic_comp) = c("Null model","Main effects","Polynomial terms",
                       "Second order terms","Only age effect");aic_comp

#   Fitted curve
#   Create fhe function for each link
myexpit <- function(x,b0,b1){
    expit <- exp(b0+b1*x)/( 1+exp(b0+b1*x) )
    expit
}
myclllit <- function(x,b0,b1){cllit <- 1-exp(-exp(b0+b1*x))}

plot(renal$age_c,as.numeric(levels(renal$reject)[renal$reject]),pch="x",xlab="Age
 ↪  (centered)",ylab="Reject",cex.axis=1.3,cex.lab=1.3)
x <- seq(min(renal$age_c),max(renal$age_c),by=0.01)
# predict(logit4,data.frame(age_c=renal£age_c),type="resp")
curve(predict(logit4,data.frame(age_c=x),type="resp"),add=TRUE,col="red")

##  Goodness of Fit
logitgof(renal$reject, fitted(logit4))
logitgof(renal$reject, fitted(probit4))
logitgof(renal$reject, fitted(cauchit4))
```

```r
logitgof(renal$reject, fitted(cloglog4))

## Quality of prediction
# Nagelkerke's R^2
NagelkerkeR2(logit4)
NagelkerkeR2(probit4)
NagelkerkeR2(cauchit4)
NagelkerkeR2(cloglog4)

#   Concordance
OptimisedConc=function(model)
{
    Data = cbind(model$y, model$fitted.values)
    ones = Data[Data[,1] == 1,]
    zeros = Data[Data[,1] == 0,]
    conc=matrix(0, dim(zeros)[1], dim(ones)[1])
    disc=matrix(0, dim(zeros)[1], dim(ones)[1])
    ties=matrix(0, dim(zeros)[1], dim(ones)[1])
    for (j in 1:dim(zeros)[1])
    {
        for (i in 1:dim(ones)[1])
        {
            if (ones[i,2]>zeros[j,2])
            {conc[j,i]=1}
            else if (ones[i,2]<zeros[j,2])
            {disc[j,i]=1}
            else if (ones[i,2]==zeros[j,2])
            {ties[j,i]=1}
        }
    }
    Pairs=dim(zeros)[1]*dim(ones)[1]
    PercentConcordance=(sum(conc)/Pairs)*100
    PercentDiscordance=(sum(disc)/Pairs)*100
    PercentTied=(sum(ties)/Pairs)*100
    return(list("Percent Concordance"=PercentConcordance,"Percent
     ↪  Discordance"=PercentDiscordance,"Percent Tied"=PercentTied,"Pairs"=Pairs))
}

OptimisedConc(logit4)
OptimisedConc(probit4)
OptimisedConc(cauchit4)
OptimisedConc(cloglog4)

#   Outlying and influential observations
#   Outlier test
outlierTest(logit4) #It can be seen that there are not outliers

#   Leverage
par(mfrow=c(1,1))
lev = influenceIndexPlot(logit4)    #   Still not sure how to analyze it

#   Influence
inf = influencePlot(logit4);inf
logit4_1 = update(logit4,subset=c(-494))
logit4_2 = update(logit4,subset=c(-331))
```

```
logit4_3 = update(logit4,subset=c(-527))
logit4_4 = update(logit4,subset=c(-1096))

compareCoefs(logit4,logit4_1)
compareCoefs(logit4,logit4_2)
compareCoefs(logit4,logit4_3)
compareCoefs(logit4,logit4_4)
#    it doesn't seem to change a lot the estimates

# Residual diagnostics: Taken from logistiscSTM.R
# Deviance residuals
par(mfrow=c(1,2))
par(pty="s")
par(mar = c(5,6,4,2)+0.1)

logit4_rdev = residuals(logit4, type = "deviance")
summary(logit4_rdev)
#hist(logit5_rdev)

plot(renal$age_c,logit4_rdev,xlab="Age (centered)",ylab="Deviance
 ↪  residual",cex.axis=1.3,cex.lab=1.3)
loess.dev = loess(logit4_rdev~renal$age_c)
lo.pred = predict(loess.dev, se=T)

orderage <- order(renal$age_c)
lines(renal$age_c[orderage],lo.pred$fit[orderage],col="blue",lwd=3)
lines(renal$age_c[orderage],lo.pred$fit[orderage]+2*lo.pred$s[orderage],
 ↪  lty=2,col="red") #rough & ready CI
lines(renal$age_c[orderage],lo.pred$fit[orderage]-2*lo.pred$s[orderage],
 ↪  lty=2,col="red")

# Pearson residuals
logit4_rpear = residuals(logit4, type = "pearson")
summary(logit4_rpear)
# hist(logit5_rpear)

plot(renal$age_c,logit4_rpear,xlab="Age (Centered)",ylab="Pearson
 ↪  residual",cex.axis=1.3,cex.lab=1.3)
#   I modified this part here
loess.dev = loess(logit4_rpear~renal$age_c)
lo.pred = predict(loess.dev, se=T)

lines(renal$age_c[orderage],lo.pred$fit[orderage],col="blue",lwd=3)
lines(renal$age_c[orderage],lo.pred$fit[orderage]+2*lo.pred$s[orderage],
 ↪  lty=2,col="red") #rough & ready CI
lines(renal$age_c[orderage],lo.pred$fit[orderage]-2*lo.pred$s[orderage],
 ↪  lty=2,col="red")

#   Influence plots: Taken from broncho_freq.R
par(mfrow=c(1,2))
par(pty="s")
par(mar = c(5,6,4,2)+0.1)

N = length(renal$reject)
id = 1:N
```

```
#   According to agresti, absolute standard residuals (either pearson or
#   deviance) larger than 2 or 3 provide evidence of lack of fit
logit4_rstandard = rstandard(logit4,type = "deviance")
logit4_rstudent = rstudent(logit4,type = "deviance")

plot(id,logit4_rstandard,type="p",
     xlab="Identification",ylab="Standardized residual",
     cex.lab=1.5,cex.axis=1.3,col="red")
plot(id,logit4_rstudent,type="p",
     xlab="Identification",ylab="Studentized residual",
     cex.lab=1.5,cex.axis=1.3,col="red")

# Global influence plots
par(mfrow=c(1,2))
par(pty="s")
par(mar = c(5,6,4,2)+0.1)
logit4_hat = hatvalues(logit4)
plot(logit4_hat,logit4_rstudent,
     xlab="Hat value",ylab="Studentized residual",
     cex.lab=1.5,cex.axis=1.3,col="red")

logit4_dffits = dffits(logit4)
plot(id,logit4_dffits,type="l",
     xlab="Identification",ylab="Dffits",
     cex.lab=1.5,cex.axis=1.3,col="red")

logit4_cov = covratio(logit4)
plot(id,logit4_cov,type="l",
     xlab="Identification",ylab="Covariance ratio",
     cex.lab=1.5,cex.axis=1.3,col="red")

logit4_cook = cooks.distance(logit4)
plot(id,logit4_cook,type="l",
     xlab="Identification",ylab="Cook's distance",
     cex.lab=1.5,cex.axis=1.3,col="red")

##  Bayesian approach
#   Default par settings
par(mfrow=c(1,1))
par(pty="m")
par(mar =c(5.1, 4.1, 4.1, 2.1))

#   Non informative prior
logit_bayes = MCMClogit(reject~age_c,family=binomial,data=renal)
logit_bayes_sum = summary(logit_bayes);logit_bayes_sum
plot(logit_bayes)

# Compare with ML solution
plot(renal$age_c,as.numeric(levels(renal$reject)[renal$reject]),
     pch="|",xlab="Age (years)",
     cex.lab=1.5,cex.axis=1.3)
x = seq(min(renal$age_c),max(renal$age_c))
lines(x,myexpit(x,b0=logit4$coeff[1],b1=logit4$coeff[2]),
      lty=2,col="red",lwd=3)
```

```r
int_bayes = summary(logit_bayes)$statistics[1,1]
slo_bayes = summary(logit_bayes)$statistics[2,1]
lines(x,myexpit(x,b0=int_bayes,b1=slo_bayes),
      lty=3,col="steelblue",lwd=3)
legend("topright",legend=c("Frequentist","Bayesian"),lty=2:3,
       bty="n",col=c("red","steel blue"),cex=1.5)

# Use of informative priors
sebeta <- logit4_sum$coefficients[,2]
precbeta <- 1/(100*sebeta^2)

logit_bayes2 <- MCMClogit(reject~age_c,family=binomial,data=renal,B0=precbeta)
logit_bayes2_sum = summary(logit_bayes2);logit_bayes2_sum
plot(logit_bayes2)

# Compare with ML solution
plot(renal$age_c,as.numeric(levels(renal$reject)[renal$reject]),
     pch="|",xlab="Age (Centered)",
     cex.lab=1.5,cex.axis=1.3)
x = seq(min(renal$age_c),max(renal$age_c))
lines(x,myexpit(x,b0=logit4$coeff[1],b1=logit4$coeff[2]),
      lty=2,col="red",lwd=3)
int_bayes2 = summary(logit_bayes2)$statistics[1,1]
slo_bayes2 = summary(logit_bayes2)$statistics[2,1]
lines(x,myexpit(x,b0=int_bayes2,b1=slo_bayes2),
      lty=3,col="steelblue",lwd=3)
legend("topright",legend=c("Frequentist","Bayesian"),lty=2:3,
       bty="n",col=c("red","steel blue"),cex=1.5)
```

Epilepsy dataset - R Code

```r
##   Clear workspace
rm(list=ls())

##   Set directory
setwd("...")

##   Load packages
library(dplyr)  #   Manage dataset
library(lmtest)   #   Sandwich estimator
library(sandwich)
library(MCMCpack)   #   Fit bayesian models

##   Read data
epil = read.table("epilepsy.txt",header=T)
head(epil)
class(epil)
str(epil)

#   Checking reference category
contrasts(epil$treatment)

#   Variable age is centered to improve interpretation
```

```r
epil$age_c = epil$age - mean(epil$age)

##   Descriptive statistics
#   General
summary(epil)

#   Amount of NA in each column
sapply(epil,function(x) sum(is.na(x)))

#   Amount of unique values in each column
sapply(epil, function(x) length(unique(x)))

#   Univariate analysis
#   Frequency of each variable
prop.table((table(epil$treatment)))
cbind(freq=table(epil$age),prop=prop.table(table(epil$age)))
cbind(freq=table(epil$seizure.rate),prop=prop.table(table(epil$seizure.rate)))

#   Continous
summary(epil[,c(2,3)])
sd(epil$age);sd(epil$seizure.rate) # Overdispersion in seizure, because of extreme
 ↪  values
hist(epil$age)
plot(density(epil$age))
hist(epil$seizure.rate)
hist(epil$seizure.rate,breaks = 25) #   Not zero inflated
plot(density(epil$seizure.rate))
boxplot(epil$age)
boxplot(epil$seizure.rate)  #   Skewed distribution

#   Bivariate analysis
cor(epil[,2:3])
#   Mean and sd for each group in each variable
summarise(group_by(epil,treatment),
          mean_age=mean(age),
          sd_age=sd(age),
          mean_seizure = mean(seizure.rate),
          sd_seizure = sd(seizure.rate))
#   It doesn't seem to be differences

#   Plots
par(mfrow=c(1,2))
boxplot(epil$age~epil$treatment,xlab="Treatment",ylab="Age")
boxplot(epil$seizure.rate~epil$treatment,xlab="Treatment",ylab="Seizure.rate")
par(mfrow=c(1,1))
plot(epil$age,epil$seizure.rate,col=epil$treatment)
legend(35, 60, legend=c("Placebo", "Progabide"),
       col=c("black", "red"), pch=1, cex=0.8)

#   Models
#   Null model
poisson_null = glm(seizure.rate~1,data=epil,family=poisson(link="log"))

#   Model 1: main effects
poisson1 = glm(seizure.rate~age_c+treatment,data=epil,
```

```
              family=poisson(link="log"));poisson1
poisson1_sum = summary(poisson1);poisson1_sum
#   Testing interaction
add1(poisson1,~.+age_c*treatment,test='LRT')

#   LRT
anova(poisson_null,poisson1,test="Chisq")
#   Deviance goodness of fit
pchisq(poisson1$deviance, df=poisson1$df.residual, lower.tail=FALSE)

#   Testing each term
drop1(poisson1,test="Chisq")
anova(poisson1,test="Chisq")

#   Model 2: main effects and interaciont
poisson2 = glm(seizure.rate~age_c*treatment,data=epil,
               family=poisson(link="log"));poisson2
poisson2_sum = summary(poisson2);poisson2_sum
#   It seems there is a problem with overdispersion

#   Testing each term
drop1(poisson2,test="Chisq")
anova(poisson2,test="Chisq")
#   Deviance goodness of fit
pchisq(poisson2$deviance, df=poisson2$df.residual, lower.tail=FALSE)
plot(poisson2)

#   Estimate the mean in each group for average age
new_epil = data.frame(age_c = c(0,0),
                      treatment = c("placebo","Progabide"))
exp(predict(poisson2,new_epil))
#   Estimate the mean in each group for 30 year old
new_epil = data.frame(age_c = c(30-mean(epil$age),30-mean(epil$age)),
                      treatment = c("placebo","Progabide"))
exp(predict(poisson2,new_epil))


#   Does the model fit the data well? Dealing with overdispersion
#   Check Poisson distribution
#   Overlay Poisson null model onto the distribution of the response
hist(epil$seizure.rate,breaks = 25,prob=T,xlab="",main="",cex.axis=1.2) #   Not
 ↪  zero inflated
poissonmu = exp(summary(poisson_null)$coefficients[,1])
lines(0:63,dpois(0:63,poissonmu),col="blue",lwd=3)

#   Check Negative Binomial distribution
nb_null = glm.nb(seizure.rate ~ 1,data = epil)
summary(nb_null)
nbmu = exp(summary(nb_null)$coefficients[,1])
nbsize = summary(nb_null)$theta
nbp = 1-nbmu/(nbmu+nbsize)
ngbindmft = dnbinom(0:63, prob=nbp, size=nbsize, log = FALSE)
lines(0:63,ngbindmft,col="red",lwd=3)
legend(30, 0.10, legend=c("Poisson", "Negative Binomial"),
       col=c("blue", "red"), lwd=3, cex=1)
```

```r
# Model counts with negative binomial distribution
nb1 = glm.nb(seizure.rate ~ age_c + treatment,data = epil);nb1
nb1_sum = summary(nb1);nb1_sum
#   The overdispersion seems to be better, but no variable is significant
#   Testing interaction
add1(nb1,~.+age*treatment,test='LRT')

#   Model 2: main effects and interaciont
nb2 = glm.nb(seizure.rate ~ age_c * treatment,data = epil)
nb2_sum = summary(nb2);nb2_sum

#   Testing each term
drop1(nb2,test="Chisq")
anova(nb2,test="Chisq")

#   Overall significance
anova(nb_null,nb2,test="Chisq")

##  QuasiPoisson model
qua_poisson = glm(seizure.rate~age_c*treatment,data=epil,
                  family=quasipoisson(link="log"));qua_poisson
qua_poisson_sum = summary(qua_poisson);qua_poisson_sum

#   Robust S.E.
coeftest(qua_poisson,vcov=sandwich)

#   Bayesian Poisson model
#   Quasi_poisson is not possible in Bayesian approach
poisson_bayes = MCMCpoisson(seizure.rate~age_c*treatment,data=epil)
poisson_bayes_sum = summary(poisson_bayes);poisson_bayes_sum
plot(poisson_bayes)
```