Generalized Linear Models

Group # 01

SEMINAR 1 Djallonke sheep data

Ricardo Castañeda r0731529

Qianli Fan r0775346

Butynets Mariia r0771332

Lieven Govaerts q0152493

Meng Wang r0767603

ZHANG Yanyi r0731121

Ruiman Zhong r0767577

Kendall Brown r0773111

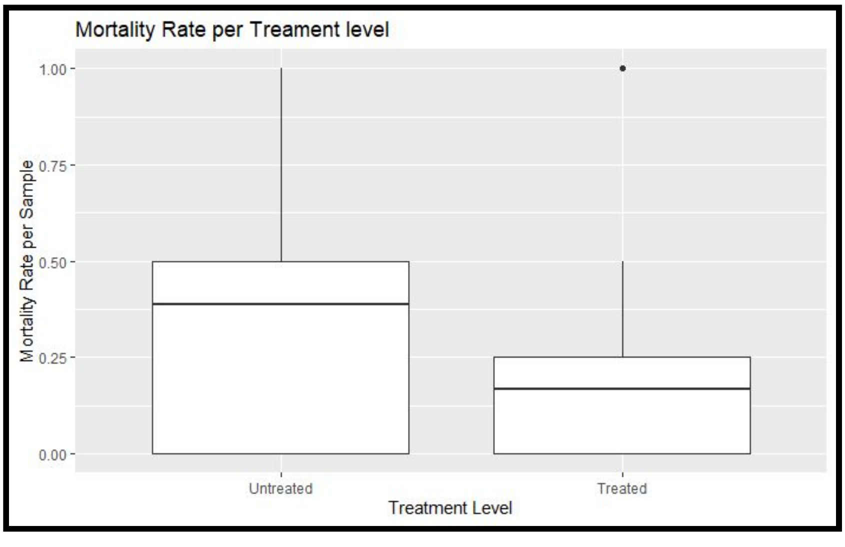
ABSTRACT

The objective of this report is to analyze the data provided for seminar 1 of KU Leuven's General Linear Models course for the second term of the academic year 2019-2020. The data set contains information regarding a set of ewes that were subject to a deworming procedure and had the mortality rates of their offspring recorded. The question of consideration is whether or not the deworming procedure significantly influences the mortality rate of the ewes' offsprings.

Three models were built and analyzed to determine the efficacy of the Descriptive Statistics deworming procedure. The logistic regression model built proved to be inadequate, failing both the Pearson chi-squared and the deviance tests, indicating that the model likely has an issue with overdispersion. Calculating the odds ratio of mortality for after assessing the performance of the logistic model, a quasi-likelihood model was built in hopes of overcoming the aforementioned issues. This model proved to be unfruitful as well, achieving a similar mortality rate. Finally, a Beta-binomial model managed to solve the overdispersion issues but in the end, all results and predictions were mostly the same.

1. DESCRIPTIVE STATISTICS

The dataset analyzed contained information regarding 75 experiments of varying sample sizes. These observations were divided into two subgroups one control(28) and one exposed to the deworming treatment(47). From the provided dataset it can be determined that the mean mortality rate of the offspring from the ewes which underwent the deworming procedure was on average 0.17 compared to 0.33 for the untreated control group. A box-plot visualization of the observed class can be seen below. The initial visual analysis shows that there is a reason to hypothesize a potential relation between receiving the deworming treatment and having a lower offspring mortality rate. This relation will be explored more formally during the model building process.



2. LOGISTIC REGRESSION

Logistic regression was built to fit the relationship between mortality rate(mr) and treatment status. The theoretical model can be expressed as:

logit[P(mr | X1)] =β0 + β1 * X1

Where X1 is a binary classifier denoting treatment status, taking a value of 1 when the group is exposed to a deworming treatment and 0 when the group is left untreated.

2.1 Fitted Model Summary

Based on the data provided, the Logistic model was fitted, and the resulting summary statistics are found below.

	Estimate	Std. Error	Z Value	P-value
Intercept	-0.5217	0.1477	-3.531	0.000414
$x1$	-0.9289	0.2028	-4.581	4.63e-06

From this table the intercept and regression coefficient of group status are obtained. Using these results the fitted model may be expressed as:

$$logit[P(mr \mid x1)] \;=\; -0.5217 \;+\; (-0.9289 * x1)\,;$$

Based on the fitted model, via a logit transformation, the probability of mortality in respect to treatment status can be obtained as follows:

P(mr)= exp(-0.5217+(-0.9289)*x1) / (1+exp(-0.5217+(-0.9289)*x1))

From the summary report the odds ratio and its 95% confidence interval, listed in the following table, are obtained.

	Estimate	2.5%	97.5%
Intercept	0.593	0.442	0.790
$x1$	0.395	0.265	0.587

2.2 Interpretation

From the table above, the odds ratio can be estimated to be 0.395. This value is smaller than 1, and thus implies a negative association between the mortality rate and the receiving a deworming treatment. To be more precise, a group that receives the deworming treatment can expect to see their offspring mortality rate drop by 26%-59%.

2.3 Logistic Regression - Prediction

Based on the fitted model, the predicted mortality can be calculated for each group, and the results are shown below. Here it can be said that when left untreated, an ewe's offspring can expect a mortality rate of 37.2%, but it decreases to 19% when the parent undergoes the deworming procedure.

	Untreated	Treated
Est. Mortality Rate	0.372449	0.189911

2.4 Goodness of fit tests (GOF)

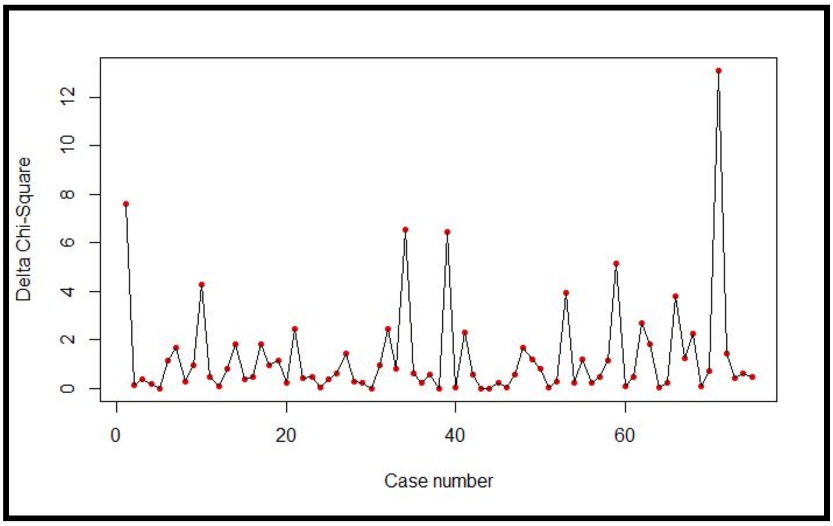
The Pearson Chi-Square test and deviance tests were performed to test the logistic model goodness-of-fit, and the resulting p-values from each test are presented in the table below.

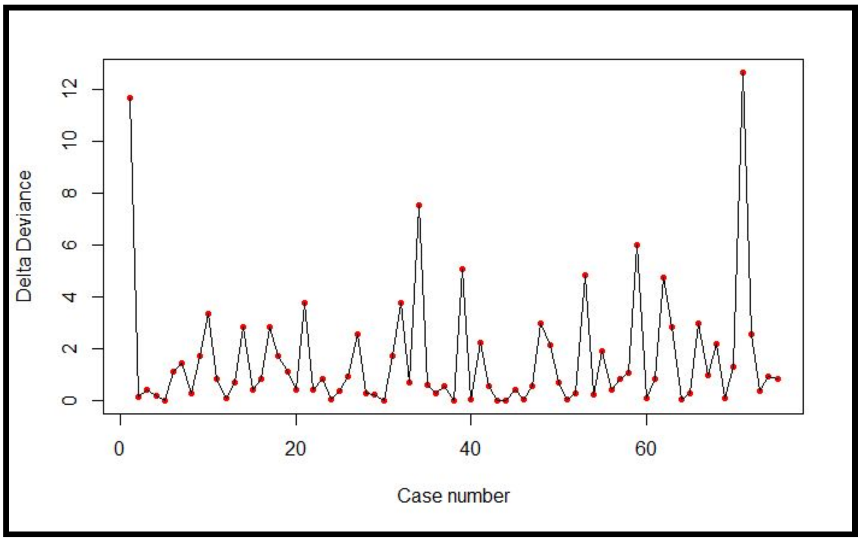
	Pearson Test	Deviance Test
p-value	0.046	0.001

In consideration of an alpha level of α=0.05, it can be seen that both tests fall below this threshold. This implies that the model fails to explain the mortality rate given the model covariates. Further analysis of the residuals will provide more details about these results.

2.5 Residual analysis

The influence of each binary case is analyzed in the Delta Chi-square and Delta Deviance residual plots as follows:





Analysis of the Delta Chi-Square plot yields that sample 71 has a high level of influence on the model. Comparing with the Delta Deviance plot indicates reaffirms this assumption in addition to signaling that observation 1 may be highly influential as well. In consideration of these findings, it may be wise to check observation 71 and find what makes this point so influential. Therefore, calculations made from the raw data verified that the mortality rate of observation 71 is 69.23%, which is much higher than the other observations. However, it is hard for an individual observation to have a great impact on the whole regression.

3. QUASI-LIKELIHOOD MODEL

As stated earlier the Pearson Chi-Squared test and Deviance test rejected the logistic regression model. We will attempt to fit the data using a quasi-logistic regression model to derive a more accurate estimation of the model parameters. Quasi-logistic regression suggests basing estimation on the penalized likelihood function. In most situations, it is used for solving the problem of complete separation. The summary statistics of this fitted model and its corresponding equation are as follows:

	estimation	std.error	t value	P-value
Intercept	-0.519	0.1681	-3.104	0.002717
$x1$	-0.9257	0.2307	-4.026	0.000137

The overdiperson parameter is 1.29.

$$Logit[P(mr|x1)] =- 0.519 - 0.9257 * x1$$

When compared to the logistic regression model, the quasi-likelihood model corresponding intercept and odds ratio show a slight increase. Additionally, the width of the confidence interval surrounding the odds ratio estimate appears to be constant between the two models.

Odds Ratio	odds ratio 2.5%	97.5%
0.396	0.265	0.588

In consideration of the calculated odds ratio, it can be claimed that the treated group will bear offspring with a mortality rate between 26%-59% that of those born to ewes without the deworming treatment.

3.1 Quasi-Likelihood model - Prediction

As it is expected from the model evaluation provided earlier, the quasi-likelihood model provides near equivalent results to the previously rejected logistic model. These results being that when left untreated, ewe's can expect their offspring to have a mortality rate of approximately 37%. If dewormed, the ewe's can expect the mortality rate to drop to 19%.

	Treat	Control
probability of death	0.1908284	0.3730964

4. BETA-BINOMIAL MODEL

According to the results obtained in the logistic regression and the quasi-likelihood models, it was observed that the predictions did not differ significantly. However, the goodness-of-fit (GOF) tests suggested not accepting the models. After analyzing the dataset, the sample size among the groups highlighted differences (TREAT = 47, CTRL =28), and this may affect the results. This inequivalence in the sample size may lead to having more variability than the assumed by the Binomial distribution, and this is known as overdispersion (when the variance is higher than the mean).

Consequently, the additional Binomial variability or overdispersion needs to be considered to improve the model's predictions and accuracy, and a Beta-Binomial regression model handles this problem. This approach is appropriate for modeling overdispersed clustered binary data such as the proportion of successes and failures as required in this scenario. This model allows the λ parameter of the Binomial distribution [B(n, λ)] to vary randomly, and in this way it considers the model's extra variability effects.

4.1 Construction of the Beta Binomial model:

Distribution part

Considering a cluster of factors (n,m), where m is the number of successful events and $m|\lambda$~Binomial (n, λ). Then, λ follows a Beta distribution Beta(a1, a2).

To calculate the marginal mean and variance, the model uses the re-parametrization $\mu=a1/(a1+a2)$, and $\phi=1/\ a1+a2+1$, where $E[m] = n*\mu$,and $Var[m] = n*\mu(1-\mu)*[1+\varphi(n-1)]$. Marginally, averaging with respect to the beta distribution for λ, Y has a beta-binomial distribution $P(m)= C(n, y)*B(a1+y, a2+n-y) / B(a1, a2)$.

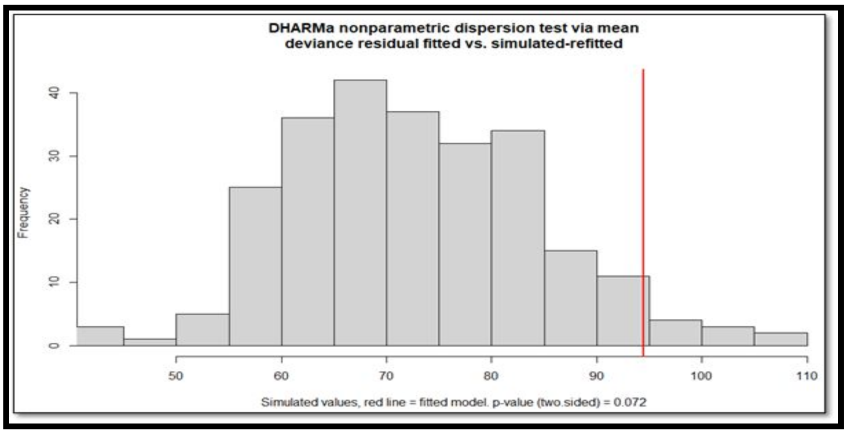
Systematic part

The response in this model is $y = m/n$. The mean is $E[y] = \mu$, defined such as $\mu = g-1\ (X*b) = g\ -1\ (v)$, indicating that g corresponds to the Logistic function, X is the Design-matrix, b stands for the Vector of fixed effects, v denotes the linear predictor (Xb), and ϕ represents the overdispersion parameter.

4.2 Overdispersion test

In this case, the package DHARMa in R tests the overdispersion in the data. It contains two tests that compare the dispersion of simulated residuals to the observed residuals. Following this, the results show insignificant p-values that suggest the data do not present overdispersion. However, the p-values are close to being significant, and this indicates that there is room for improvement. Besides, the graph illustrates the distribution does not fall outside the cut off value, but it is close to the limit, and thereby the Beta Binomial model could solve overlooked issues.

DHARMa nonparametric dispersion test via mean deviance			
Dispersion	13.011,00	P-Value	0,072
Alternative	Hypothesis:	Two,sided	



4.3 Beta-binomial Results Interpretation

The model that links the response variable with the predictors is represented just like the previous models, but in this case, to calculate the ML for the estimators and the standard errors, it is assumed that Yi follows a beta-binomial distribution.

Mu coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.5848	0.1862	-3,141e+00	1.685e-03
$x1$	-0.8520	0.2528	-3.372e+00	7,471e-04

Following this, the regression coefficients from the table above fit the model as follows:

logit[P(mr | x1)] = -0.5848 + (-0.8520*x1);
P(mr | x1)] = e (-0.5848 + (-0.8520*x1))/ [1+ e(-0.5848 + (-0.8520*x1))]

Phi coefficients: (scale = identity)	Estimate	Std. Error
Phi.(Intercept)	0.05606	0.03014

The estimator phi corresponds to the overdispersion parameter Φ= 0.056, which influences the variance. Hence, the probabilities of mortality in the ewes' offsprings according to the group have an estimated standard deviation = $\sqrt{[0.056(\mu(1-\mu))]}$. This confirms that this model considered the extra variance that the logistic and Quasi-Likelihood regression models ignored.

4.4 Beta-binomial prediction

The final prediction illustrates small differences compared to the other models. The Beta Binomial model predicts a rate of mortality that is around one percent lower (0.358) than the one obtained with the Logistic and Quasi-likelihood models (0.37). Therefore, it can be said that based on the Pearson Chi-Square test, the Beta Binomial model is more accurate, but when applied to reality the results do not show a relevant improvement.

Predicted M. Rate	Untreated	Treated
	0,3579	0,1921

4.5 Goodness of fit tests (GOF)

The Pearson Chi-squared and Deviance tests are used again to assess the model based on a confidence level of α= 0.05. Then, the Pearson test depicts a p-value considerably higher than 0.05, and this suggests that the regression coefficients explain the mortality rate accurately. This shows that the Beta Binomial model considered the extra-variance that caused overdispersion and solved the problems that made the Pearson chi-square test to reject the previous models.

However, the deviance test still indicates that the model does not explain the variance in the data accurately, as it provides a significant p-value. Therefore, before deciding on the final model, it is necessary to make some predictions and compare the results with the different models.

Person	
Xs1	p-value
68.497	0.627

Deviance		
Dev1	df1	p-value
115.57	73	0.001106

5. FINAL COMPARISONS

Predictions for Mortality Rate (MR.)				
Model	Untreated	Treated	E[m|X=0] = 28*μ	E[m|X=1] = 47*μ
1 logistic	0,3724	0,1899	10,429	8,93
2 quasi-likelihood	0,3731	0,1908	10,447	8,97
3 Beta-binomial	0,3579	0,1921	10,021	9,03
4 Unweighted M.R	0,3307	0,1727	9,260	8,12
5 Population Weighted M.R.	0,3724	0,1899	10,429	8,93

To conclude, the observed or actual mortality rates were calculated using the traditional average methods, to compare the accuracy of the predicted rates or estimates. Two types of methods were applied as follows:

Unweighted Mortality rate:

- $MR_i = y_i/n_i$
- MR Treat = mean(MR_{ix1})
- MR CTRL = mean(MR_{ix0})

Population weighted mortality rate:

- MR Treatment = sum(y_{x1})/sum(n_{x1})
- MR CTRL = sum(yx0)/sum(nx0)

On the one hand, the proportion for the unweighted mortality rate is calculated sample by sample, and then these rates are used to average the mortality rate per group. On the other hand the Weighted approach considers each observation according to its population size, and hence those samples with large n will have more significance.

Overall, the predictions are quite close to the actual mortality rates, and the results do not differ significantly among them. However, especially for the untreated group, the unweighted Mortality rate and the Beta-binomial approaches show a lower probability compared to the other methods. It may be associated with the fact that the Beta-binomial model tries to consider the extra variance caused by the differences in the sample size and the unweighted Mortality rate also reduces the sample size difference by calculating the average from all mortality rates per group. Besides, the results depicted by the weighted Mortality rate seem to match with the Logistic and Quasi-likelihood regression models that did not consider the differences in the sample size within the groups.

6. CONCLUSION

Based on the results from the model building procedure it can be concluded that of the three models built, the most adequate model proved to be the beta-binomial model. This model was able to overcome the overdispersion issue of the logistic regression model quite well and produced results which failed to reject the pearson chi-squared, unfortunately this model was still unable to pass the deviance test for fitness. Based on the results gathered from the beta-binomial model, it can be said that an untreated ewe can expect their offspring to die prematurely approximately 36% of the time. If they were to undergo the deworming procedure the mortality rate drops to slightly 19%. As this is a common trend amongst all models tested here, it can be said that there does appear to be a relation between treatment status and offspring mortality rate.

Additionally, all models appear to be over-estimated across the per group mortality rates. This is unfortunate as the untreated mortality rate is being over estimated by over 2.5% by the best model calculated. A similar overestimation of just under two percent is found in the treated group. The relative change in mortality rate does appear to be constant amongst all groups implying that the treatment is effective and should be employed in consideration for the life of ewes.