2016-17 GLM course KULeuven
Exam projects
M.Sc. Statistics

Members

Names:	*Student numbers:*
Nozomi Takemura	r0649141
Bharat Ram Ammu	r0614303
Björn Rafn Gunnarsson	r0648841
Daniel Izquierdo Juncàs	r0654210
Robert Broughton	r0647509

Email addresses:
nozomi.takemura@student.kuleuven.be
bharatram.ammu@student.kuleuven.be
bjornrafn.gunnarsson@student.kuleuven.be
daniel.izquierdojuncas@student.kuleuven.be
robert.broughton@student.kuleuven.be

Supervisor
Prof. Emmanuel Lesaffre

13 JUNE 2017

1 Part 1 - Poisson regression

1.1 Introduction

In this project we study the dataset “RoadKills”, which consists of observations taken from a two-year study on amphibian road kills in a National Road of southern Portugal. The dead animals were separated by road segments (52 in total) and allocated to the coordinates of its middle point. The response variable is the total number of amphibian fatalities per segment (TOT.N). The response and covariates we consider in this analysis are presented in Table 1. We conduct first a frequentist analysis and later we fit the model obtained in a Bayesian way.

Response variable:	TOT.N:	Total number of amphibian fatalities per segment
Covariates:	OPEN.L:	Open lands (ha))
	MONT.S:	Montado with shrubs (ha)
	POLIC	Policulture (ha)
	D.PARK:	Distance to Natural Park (m)
	SHRUB:	Shrubs (ha)
	WAT.RES:	Water reservoirs (ha)
	L.WAT.C:	Length of water courses (km)
	L.P.ROAD:	Paved road length (km)
	D.WAT.COUR:	Distance to water courses

Table 1: Explanatory variables

1.2 Exploratory analysis

Table 2: Simple statistics

	TOT.N	OPEN.L	MONT.S	POLIC	D.PARK	SHRUB	WAT.RES	L.WAT.C	L.P.ROAD	D.WAT.COUR
mean	25.90	36.18	1.08	0.60	12680.89	0.23	0.32	1.56	0.96	288.68
std.dev	24.28	26.50	2.08	1.72	7327.19	0.34	0.96	1.00	0.53	281.68
var	589.30	702.30	4.34	2.95	53687724.98	0.11	0.93	1.01	0.28	79343.27
median	17.50	28.53	0.00	0.06	12719.63	0.09	0.04	1.55	0.68	196.01
min	2.00	0.74	0.00	0.00	250.21	0.00	0.00	0.00	0.57	15.18
max	104.00	97.57	9.43	11.26	24884.80	1.74	6.31	3.95	2.96	1165.00

In Table 2 the mean, standard deviation, variance, median, minimum and maximum are shown for each variable of the dataset. Most variables standard deviations are of the same order as their means, suggesting a lot of variance. In most cases medians are lower than means pointing out the existence of influential high observations and that variable distributions might be right skewed, which we visually confirm checking the boxplots (Appendix A.2). For the variables Montado with shrubs, Policulture, Shrubs and Water reservoirs the median is 0 or very close to 0, indicating that a large number of road segments have a value of 0 for these variables.

In Figure 1 we plot the histogram of the response. In agreement with the results observed in Table 2 and in the boxplots, we see that TOT.N distribution is right skewed with a large number of sectors with counts close to zero. When the response is a count of rare events we usually model the data with Poisson, negative binomial or quasi-Poisson regression.

1.3 Frequentist analysis

1.3.1 Multicollinearity

Before starting the regression, we check for multicollinearity as it can cause statistical and computational issues. Same as in linear regression, we compute the Variance Inflation Factors (VIF) of the predictors and look for large values. All VIFs are smaller than 5 so multicollinearity is not a problem (Table 3).

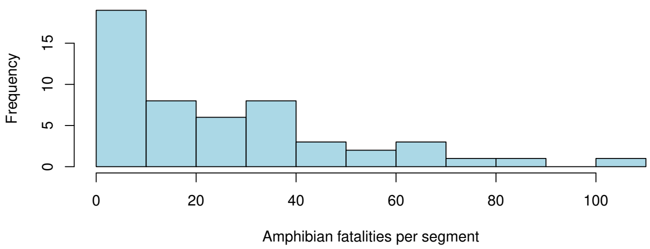

Figure 1: Histogram of response

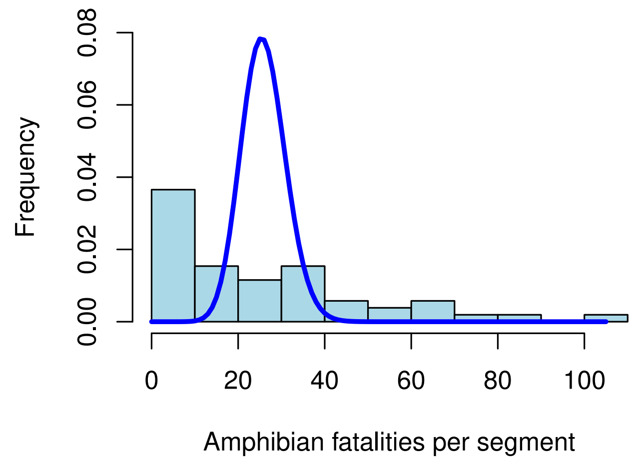

Figure 2: Histogram of frequencies and fitted Poisson distribution of the null model

Table 3: VIF's of the predictors

	OPEN.L	MONT.S	POLIC	D.PARK	SHRUB	WAT.RES	L.WAT.C	L.P.ROAD	D.WAT.COUR
1	1.24	1.15	1.31	1.80	1.55	1.30	2.03	1.12	1.72

1.3.2 Poisson regression

We start by performing a visual check comparing the histogram of frequencies of the response with a Poisson distribution with parameter λ obtained from fitting the null model (Figure 2). Even though the fitted distribution tries to accommodate to the data, the response is clearly not Poisson distributed. The real distribution has a higher frequency around zero and is overall more spread than a Poisson distribution.

Model selection

Despite the response not being Poisson distributed, we fit a generalized linear model with Poisson as the distribution part and the covariates as the systematic part. For the systematic part the canonical link (logarithm) is used. We start fitting a model whose systematic part includes all linear terms (Table 4). In this model all covariates except D.WAT.COUR are found to be significant at a 5% level.

Table 4: Coefficient estimates for the full model of Poisson regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.9950	0.1083	36.90	0.0000
OPEN.L	-0.0046	0.0015	-3.02	0.0025
MONT.S	0.0867	0.0133	6.53	0.0000
POLIC	-0.0306	0.0144	-2.12	0.0342
D.PARK	-0.0001	0.0000	-22.34	0.0000
SHRUB	-0.5758	0.1009	-5.70	0.0000
WAT.RES	0.1082	0.0291	3.72	0.0002
L.WAT.C	0.3125	0.0429	7.28	0.0000
L.P.ROAD	0.1706	0.0564	3.02	0.0025
D.WAT.COUR	0.0001	0.0001	0.66	0.5092

Table 5: Coefficient estimates for the final model of Poisson regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.0337	0.0906	44.51	0.0000
OPEN.L	-0.0048	0.0015	-3.19	0.0014
MONT.S	0.0872	0.0132	6.59	0.0000
POLIC	-0.0279	0.0138	-2.02	0.0439
D.PARK	-0.0001	0.0000	-22.97	0.0000
SHRUB	-0.5650	0.0996	-5.67	0.0000
WAT.RES	0.1013	0.0272	3.73	0.0002
L.WAT.C	0.2975	0.0364	8.18	0.0000
L.P.ROAD	0.1762	0.0557	3.16	0.0016

Next, model selection methods are applied to find the "best" model using significance tests to compare nested models. There are 3 significance tests available: the likelihood ratio test (LRT), the score test and the Wald test, which are all asymptotically equivalent. Here, we only work with the former two: LRT evaluates if the data is likely to have come from a more complex model instead of the simple one and score test evaluates if a parameter is equal to a certain value (0 in this case). Starting with the full model and using drop1 function, LRT suggests to only drop D.WAT.COUR at a 5% significance level (p-value = 0.5098). Score test gives the same results with a p-value for D.WAT.COUR of 0.5092. Using the function step in R, we perform stepwise selection based on the Akaike information criterion (AIC) and the same model is reached with AIC = 530.51.

Model adequacy

The systematic and distribution parts of the "best" model obtained with Poisson regression are

$$\begin{aligned} log(\lambda_i) &= log[E(y_i|x_i))] \\ &= \beta_0 + \beta_1 OPEN.L_i + \beta_2 MONT.S_i + \beta_3 POLIC_i + \beta_4 D.PARK_i \\ &\quad + \beta_5 SHRUB_i + \beta_6 WAT.RES_i + \beta_7 L.WAT.C_i + \beta_8 L.P.ROAD_i \end{aligned} \tag{1}$$

$$y_i \sim Poisson(\lambda_i)$$

with y_i the number of amphibian fatalities for the ith observation with regressors x_i, λ_i its expected value and β_i the coefficients shown in Table 5.

We apply the deviance goodness of fit test to compare the observed with the estimated frequencies. Under the correct model, the deviance

$$D \xrightarrow{d} \chi^2_{(n-p)}$$

and as a rule of thumb the ratio between the residual deviance and the residual degrees of freedom should be close to 1. In this case deviance is 273.12 and the residual degrees of freedom is 43 so $\frac{D}{df_{res}} = 6.3517$ with deviance χ^2-test p-value ≈ 0. The null hypothesis stating that the fitted model is correct is rejected and the ratio indicates strong overdispersion. In a Poisson distribution $E(y) = Var(y)$, which is clearly not the case as the mean response is 25.90 and its variance is 589.30 (Table 2), so overdispersion was expected as $Var(y) \gg E(y)$.

Next, we produce various plots to detect which are the most influential observations on the construction of the model (Figure 3). We see that many residuals have absolute values higher than 1.96, indicating that the model is not fitting the data well. We also notice that there are two highly influential observations: sector 8 and sector 11. Sector 8 is characterized by having the highest policulture ($11.263ha$) and sector 11 has the highest water reservoirs ($6.309ha$).

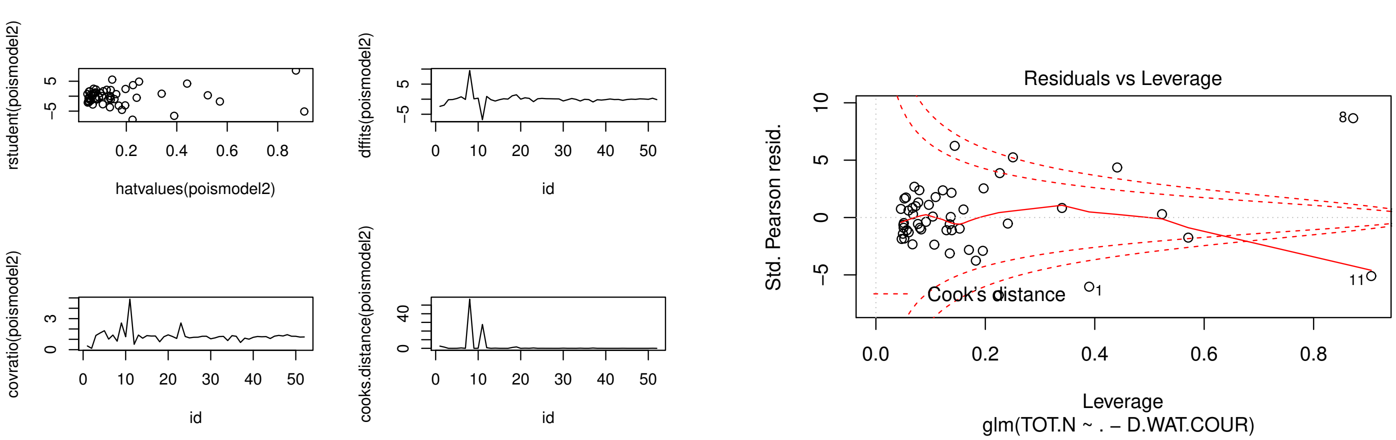

Figure 3: Influence plot for Poisson regression

For all the reasons exposed above it is concluded that Poisson regression is not a good model to fit the data so other alternatives will be explored.

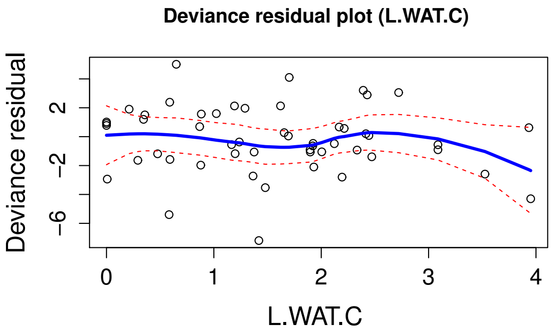

Figure 4: Deviance residual plot for Length of water courses (km)

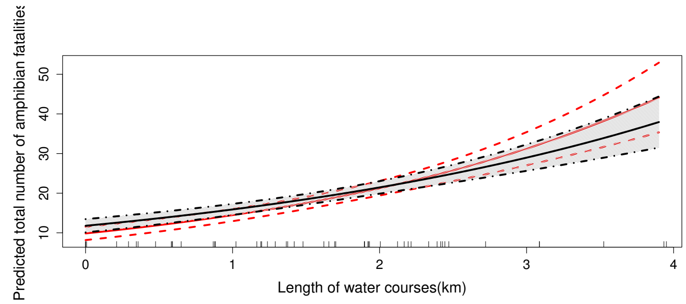

Figure 5: Predicted responses and CIs for the Poisson regression with interactions (red) vs without interactions) (black) with OPEN.L, MONT.S, POLIC, D.PARK, SHRUB, WAT.RES, and L.P.ROAD held at their means

Further comments on Poisson regression

Apart from the canonical link, that linearizes the relationship between response and covariates, Poisson distribution admits the identity and the square root as link functions. In this case, the final model obtained with the square root link $\sqrt{\lambda_i} = \boldsymbol{x}_i\boldsymbol{\beta}^T$ includes all the regressors except D.WAT.COUR and POLIC and its AIC = 516.75 indicating a better fit than with the canonical link. However, it suffers from the same overdispersion problems as its counterpart as the ratio between residual deviance and residual degrees of freedom is still larger than 5.

In the model selection interactions between regressors are considered, but model (1) is chosen as the final one. The model containing all the interactions has p $= 9 \times \frac{9-1}{2} + 9 + 1 = 46$ parameters, which even though being smaller than n = 52, it is still regarded as a small sample to work with the full interactions model. The stepwise procedure considering the interactions results in the model with 18 parameters whose AIC = 317.6, deviance residual = 42.24, and its dof = 34. However, the standard errors of estimated coefficients for main effects are larger than those for the model without interactions.(Appendix A.1). This, as can be seen in Figure 5, leads to its confidence intervals of mean responses inflating as well as, especially in the space with few observations.

It could be contemplated the use of a zero-inflated Poisson (ZIP), a model combining two processes: one governed by a Poisson distribution and the other by a binary distribution that generates zeros. Even though the histogram of the response (Figure 1) might suggest the use of a ZIP model, there are no segments with exactly zero dead animals found. Hence, this model does not make sense in our case.

1.3.3 Negative binomial regression

When overdispersion is found, the negative binomial model is a good alternative to Poisson. This distribution has larger variance than mean, which results in a longer and fatter tail. In Figure 6 we show how the negative binomial distribution fits the response compared with the Poisson one (with distribution parameters given by null models). This distribution is able to capture the observed data much better than the Poisson, following the data in the peak around zero and in the long tail. However, the predicted values for the peak around zero are still lower than the observed and it predicts higher frequencies for segments with 10 to 30 fatalities.

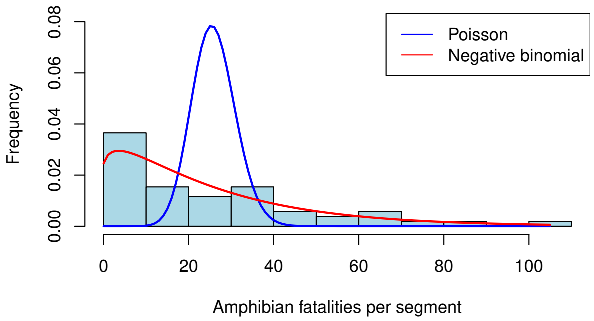

Figure 6: Histogram of frequencies, fitted Poisson distribution of the null model (blue) and fitted negative binomial distribution of the null model (red)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.2487	0.2761	15.39	0.0000
OPEN.L	-0.0099	0.0031	-3.16	0.0016
MONT.S	0.0613	0.0343	1.79	0.0737
POLIC	0.0008	0.0421	0.02	0.9851
D.PARK	-0.0001	0.0000	-9.60	0.0000
SHRUB	-0.3764	0.2451	-1.54	0.1247
WAT.RES	0.0968	0.0769	1.26	0.2078
L.WAT.C	0.1919	0.1005	1.91	0.0560
L.P.ROAD	0.2674	0.1363	1.96	0.0498
D.WAT.COUR	-0.0001	0.0003	-0.20	0.8396

Table 6: Coefficient estimates for the full model of negative binomial regression.

Model selection and interpretation

We fit a generalized linear model with the response following a negative binomial distribution and with a logarithm link function. Similar with the Poisson regression we start fitting a model whose systematic part includes all linear terms (Table 6). With this model we observe that many covariates that were significant in the Poisson model are not significant anymore. When the wrong distribution is taken the systematic part tries to accommodate for it, giving results that do not need to be true. Here, only OPEN.L, D.PARK and L.P.ROAD are significant at a 5% level (the latter barely).

Again, we use model selection methods to find the "best" model using significance tests to compare nested models. Starting with the full model and using drop1 function with the LRT as our selection criterion, we discard one variable at a time until all variables are found significant by the LRT. The final model includes the variables OPEN.L, D.PARK and L.WAT.C as regressors. We also test the final model with the score test and reach the same conclusion.

The model obtained using stepwise selection based on the AIC gives a model that apart from the 3 mentioned variables also includes L.P.ROAD with an AIC = 382.9. If we base our selection in the Bayesian information criterion (BIC), which has a stronger penalization for complicated models, the final model is also the one that only includes OPEN.L, D.PARK and L.WAT.C. Thus, we decide to choose this as our final model, with AIC = 384.25 and BIC = 390.06. Results are presented in Table 7. The shape parameter is $\theta = 4.73$ with associated standard error = 1.16; the dispersion parameter is taken to be 1 by the fit.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.4666	0.1716	26.02	0.0000
OPEN.L	-0.0107	0.0031	-3.41	0.0007
D.PARK	-0.0001	0.0000	-10.71	0.0000
L.WAT.C	0.1867	0.0791	2.36	0.0182

Table 7: Coefficient estimates for the final model of negative binomial regression.

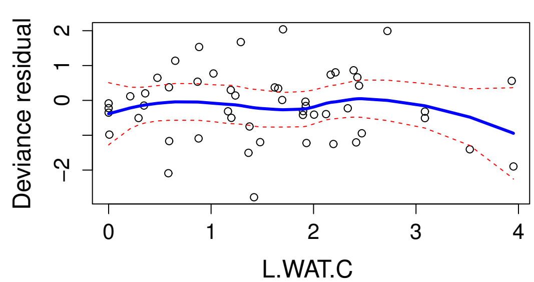

Figure 7: Deviance residual for L.WAT.C (negative binomial regression)

As shown in Table 7, the estimated coefficient of OPEN.L is -0.0107, indicating the log of an expected number of amphibian fatalities per segment decreases by 0.0107 for one unit rise in open lands(ha) when the distance to natural Park (D.PARK) and length of water courses (L.WAT.C) are kept constant. The estimate, -0.0001 for D.PARK implies $1m$ increase in the distance to Natural Park decreases the log of an expected count of fatalities by 0.0001, as long as OPEN.L and L.WAT.C are held constant. On the contrary, $1km$ rise in the L.WAT.C increases the log of an expected count of those fatalities by 0.1867. Although D.PARK seems less important than L.WAT.C due to quite a big difference in the order of magnitude of their estimates, this can be just because of the difference in their units, m and km. The intercept is the expected value of the log of counts in the situation where all covariates are 0, which in this case would imply extrapolation as the minimum for some covariates is larger than 0.

Model adequacy

The deviance goodness of fit test is used to compare the observed with the estimated frequencies. The deviance of the final model is 52.003 and the residual degrees of freedom is 48 so $\frac{D}{df_res} = 1.0834$ with deviance χ^2-test p-value ≈ 0.32. Hence, we do not reject the null hypothesis and conclude that the model fits the data well. In addition, the ratio $\frac{D}{df_res}$ is very close to the ideal value 1, so this model does not suffer from overdispersion. The magnitude of the deviance residuals for this model (Figure 7) seems to be smaller than that for the Poisson model (Figure 4). This suggests that the possible outlying observations would be far less for this model.

We produce various plots to detect which are the most influential observations on the fit (Figure 8). The most notable change compared with the Poisson counterpart is that in this model there are not too many large residuals. The Dffits plot, which shows how influential is a point in the regression, does not present extreme observations compared with the others. Similarly, in the covariance ratio and the Cook's distance plots we do not observe observations that influence too much the model (as it was observed in Figure 3).

Prediction

Figure 9 shows the mean responses and its confidence intervals of the number of fatalities over L.WAT.C when OPEN.L and D.PARK are fixed at their means. As suggested in Figure 9, the CI is much larger for the high values of L.WAT.C because there are only a few observations. Interestingly, the evolution of the mean responses over L.WAT.C appear to be roughly linear though it would be more natural to see the exponential development. This can be due to the fact the estimated coefficient for the intercept is more than 20 times larger than that for L.WAT.C.

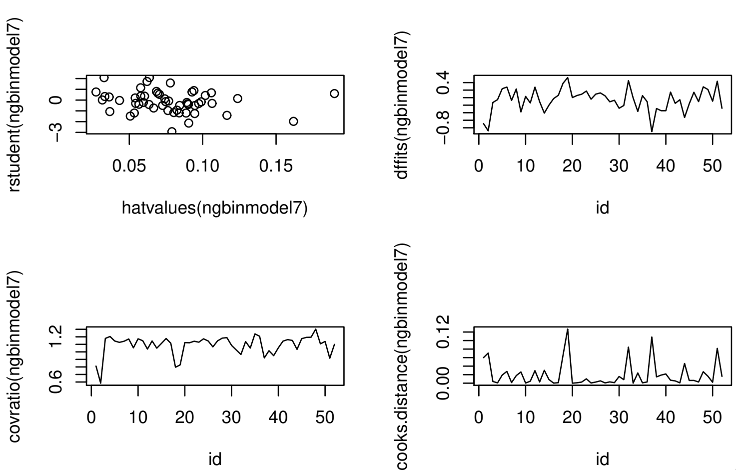

Figure 8: Influence plots for negative binomial regression

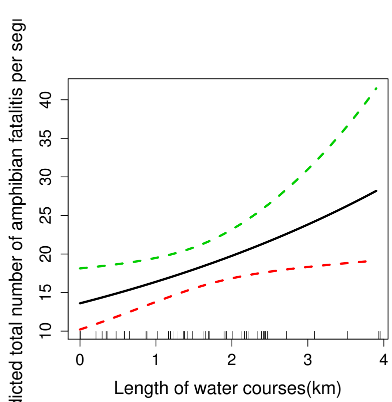

Figure 9: Prediction of TOT.N for mean OPEN.L(36.18ha) and mean D.PARK(12680.89m)

1.3.4 Quasi-Poisson regression

Model selection and interpretation

Quasi-likelihood models are characterized for specifying separately the mean and the variance function so they are used in cases with overdispersion. We fit a quasi-Poisson regression model with all covariates included. In Table 8 the coefficient estimates are presented with their robust standard errors obtained with the sandwich estimator for the covariance matrix. We see that the estimated coefficients are approximately the same as the ones obtained with Poisson regression (Table 4), but in the quasi-Poisson case the standard errors are much larger which results in non-significance of many variables that were significant for Poisson.

Table 8: Coefficient estimates for the full model of quasi-Poisson regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.9950	0.2639	15.1410	0.0000
OPEN.L	-0.0046	0.0037	-1.2383	0.2225
MONT.S	0.0867	0.0324	2.6783	0.0105
POLIC	-0.0306	0.0352	-0.8689	0.3899
D.PARK	-0.0001	0.0000	-9.1661	0.0000
SHRUB	-0.5758	0.2460	-2.3406	0.0241
WAT.RES	0.1082	0.0710	1.5244	0.1349
L.WAT.C	0.3125	0.1047	2.9857	0.0047
L.P.ROAD	0.1706	0.1376	1.2403	0.2217
D.WAT.COUR	0.0001	0.0004	0.2708	0.7879

Table 9: Coefficient estimates for the final model of quasi-Poisson regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.0253	0.1654	24.3442	0.0000
MONT.S	0.0699	0.0329	2.1228	0.0390
D.PARK	-0.0001	0.0000	-9.8649	0.0000
L.WAT.C	0.1682	0.0796	2.1133	0.0398

Since the AIC does not exist for quasi-Poisson models, the "best" model is again searched with LRT. Although SHRUB is found to be significant at the final model obtained by the LRT, it is not significant by the score test; thus, it is removed. Finally, the model consists of the following covariates: MONT.S, D.PARK, and L.WAT.C whose estimated coefficients are shown in Table 9. The estimated coefficients for D.PARK and L.WAT.C are -0.0001 and 0.1682, which are more or less same to the one described for the negative binomial final model's output as well as their standard errors. MONT.S is now identified to be important with its estimate 0.0699, meaning $1ha$ increase in the Montado with shrubs would lead to a rise in the log of an expected total number of amphibian fatalities per segment by 0.0699 when D.PARK and L.WAT.C are unchanged. The interpretation of the intercept is similar as the one carried out in negative binomial regression.

Model adequacy (Comparison with the Negative binomial regression)

To study the goodness of fit, the χ^2 test is again utilized: $\frac{D}{dof_{res}} = \frac{335.99}{48} \approx 6.9997$. The corresponding p-value ≈ 0, giving enough evidence to reject the null hypothesis that this model is correct, which was expected as the deviance test also rejected the null hypothesis in the Poisson case.

The possible influential observations are again studied from Figure 10. Similar to the negative binomial regression results, the number of observations detected to be influential is much smaller than the one obtained in the Poisson regression. There does not seem to be much difference between the quasi-Poisson and negative binomial regression outputs.

Prediction (Comparison with the Negative binomial regression)

The Figure 11 shows the confidence intervals and point estimates of expected TOT.N over the Length of water courses for quasi-poisson regression, given Montado with shrub and distance to Natural Park fixed on their means. Although such a given constrain is different from the one assumed for Figure 9, we still compare them since it would not influence the predictions to fix a regressor found not to be important. According to Figure 11, CIs and estimated mean responses seem to be roughly similar for both regression approaches, but still quasi-Poisson regression can yield slightly higher (by around 1) predicted

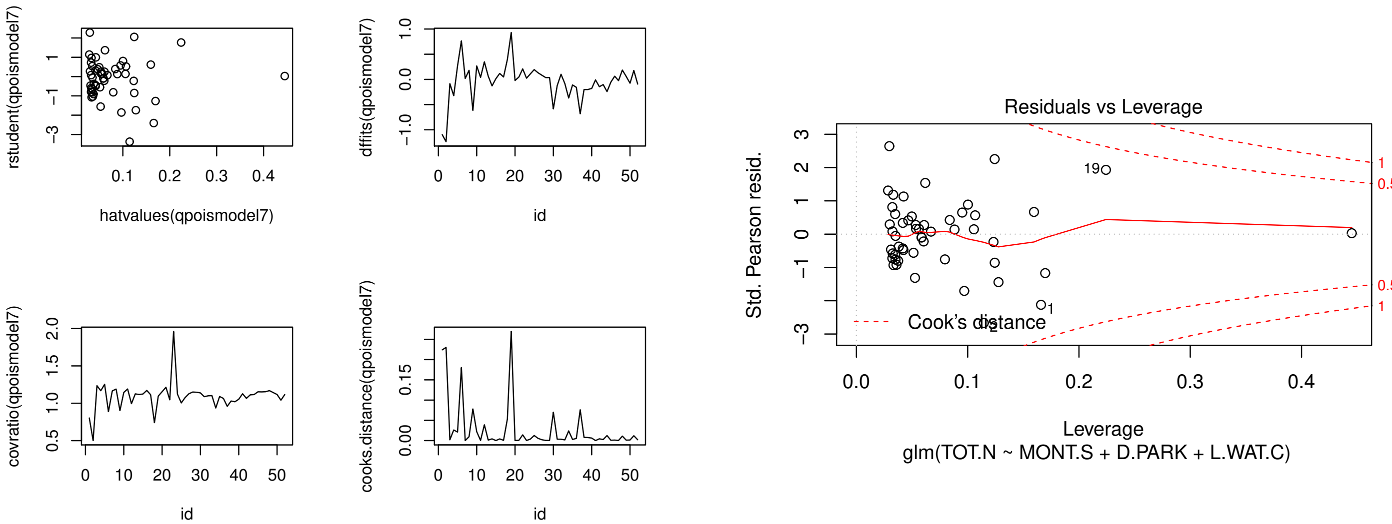

Figure 10: Influence plot for Poisson regression

responses than negative binomial regression approach on average. However, such a difference is thought to be decreased as water courses become longer.

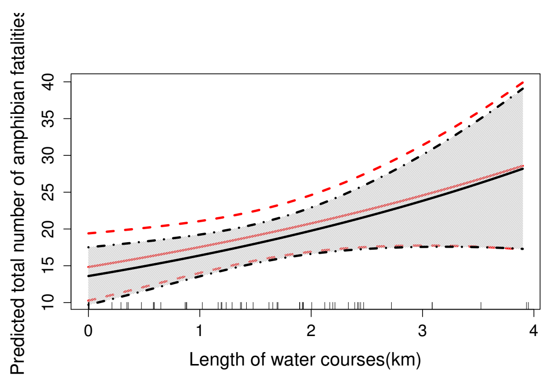

Figure 11: Predictions and CIs of means of TOT.N with MONT.S and D.PARK held at their means for Quasi-Poisson regression (red) and for negative binomial regression (black).

1.4 Bayesian analysis

We estimate the final model obtained in the Poisson regression (Model 1) in a Bayesian manner. The main reason why this model is selected for Bayesian analysis, in spite of that it would be incorrect because of the overdispersion, is the following. Firstly, there seems no available functions in *MCMCpack* which allow us to implement Bayesian negative binomial regression. Secondly, the quasi-likelihood approach does not use a likelihood, and therefore it is not capable for its model to be fitted with a standard Bayesian method which is based on the likelihood. Another possibility would be to fit a Bayesian Poisson regression for only the variables that were find significant in the quasi-Poisson approach.

As a first step, the prior distributions of the parameters (regression coefficient) need to be specified. Since there is no prior information about them, a non-informative prior would be appropriate. The model is fitted with a multivariate Normal prior on $\boldsymbol{\beta}$ with mean and variance given by the default in *MCMCpoisson*, 1000000 iterations of which first 10000 are not used (burn in).

Table 10: Output from the bayesian Poisson regression

	Mean	SD	Naive SE	Time-series SE
(Intercept)	4.0335	0.0906	0.0001	0.0005
OPEN.L	-0.0048	0.0015	0.0000	0.0000
MONT.S	0.0868	0.0132	0.0000	0.0001
POLIC	-0.0285	0.0139	0.0000	0.0001
D.PARK	-0.0001	0.0000	0.0000	0.0000
SHRUB	-0.5664	0.0999	0.0001	0.0006
WAT.RES	0.0999	0.0272	0.0000	0.0002
L.WAT.C	0.2974	0.0365	0.0000	0.0002
L.P.ROAD	0.1768	0.0559	0.0001	0.0003

Table 11: Quantiles for each variable

	2.5%	25%	50%	75%	97.5%
(Intercept)	3.8562	3.9722	4.0337	4.0948	4.2111
OPEN.L	-0.0077	-0.0058	-0.0048	-0.0038	-0.0018
MONT.S	0.0606	0.0780	0.0870	0.0958	0.1124
POLIC	-0.0562	-0.0378	-0.0283	-0.0191	-0.0018
D.PARK	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
SHRUB	-0.7661	-0.6330	-0.5653	-0.4983	-0.3735
WAT.RES	0.0457	0.0816	0.1001	0.1184	0.1526
L.WAT.C	0.2256	0.2728	0.2973	0.3218	0.3692
L.P.ROAD	0.0663	0.1393	0.1770	0.2146	0.2855

After fitting the model, to assure if the samples obtained from MCMC procedure are truly from the posterior distribution, the trace plots are studied. As seen in the Figure 12, the convergence seems to be reached (see also Appendix A.3). This is also acquired from the Table 10; more specifically, the Naive SE and Time-series SE appear to be quite close, suggesting our samples are basically independent. The Time-series SE is the estimated sampling error whereas Naive SE is the one obtained under the assumption that there is no autocorrelation among samples.

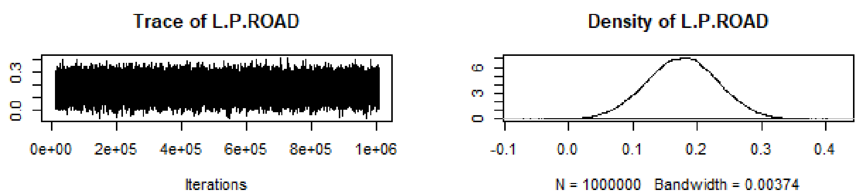

Figure 12: Trace plot (L.P.ROAD)

The estimated posterior means shown in Table 10 are roughly the same as the Poisson regression's estimated coefficients (Table 5). Although the estimated standard deviations are also quite similar to the standard errors in Table 5, their meaning is totally different; former ones describe the uncertainty we have about the estimated posterior mean (parameter) given this data, while latter ones give the variability of the estimated parameters when samples from the population are repeatedly drawn. Table 11 provides the 95 % equal tail credible interval for each parameter, all of which do not include 0, giving evidence against removing those covariates. This is the same conclusion attained from the frequentist Poisson regression.

1.5 Conclusion

In this analysis, Poisson regression model is fitted in a frequentist and in a Bayesian approach. With non-informative prior for Bayesian Poisson regression, the same conclusion is obtained for both procedures, reaching a final model that contains eight regressors. However, it is revealed that there is overdispersion, which cannot be dealt by Poisson regression, so the negative binomial regression and quasi-Poisson regression are conducted in a frequentist manner. From the χ^2 Goodness of fit test, there is not strong evidence against that the model fits the data well in negative binomial regression, supported by its p-values 0.32. The final Poisson regression model suggests keeping all covariates except D.WAT.COUR while correcting for overdispersion in the quasi-Poisson model only MONT.S, D.PARK and L.WAT.C are significant. The final model of the negative binomial regression suggests only three important covariates: it is found the open lands and distance to natural park have a negative effect on the expected total number of amphibian fatalities per segment, while length of water courses have a positive impact on it with all the other covariates kept constant.

Appendices

A Part1

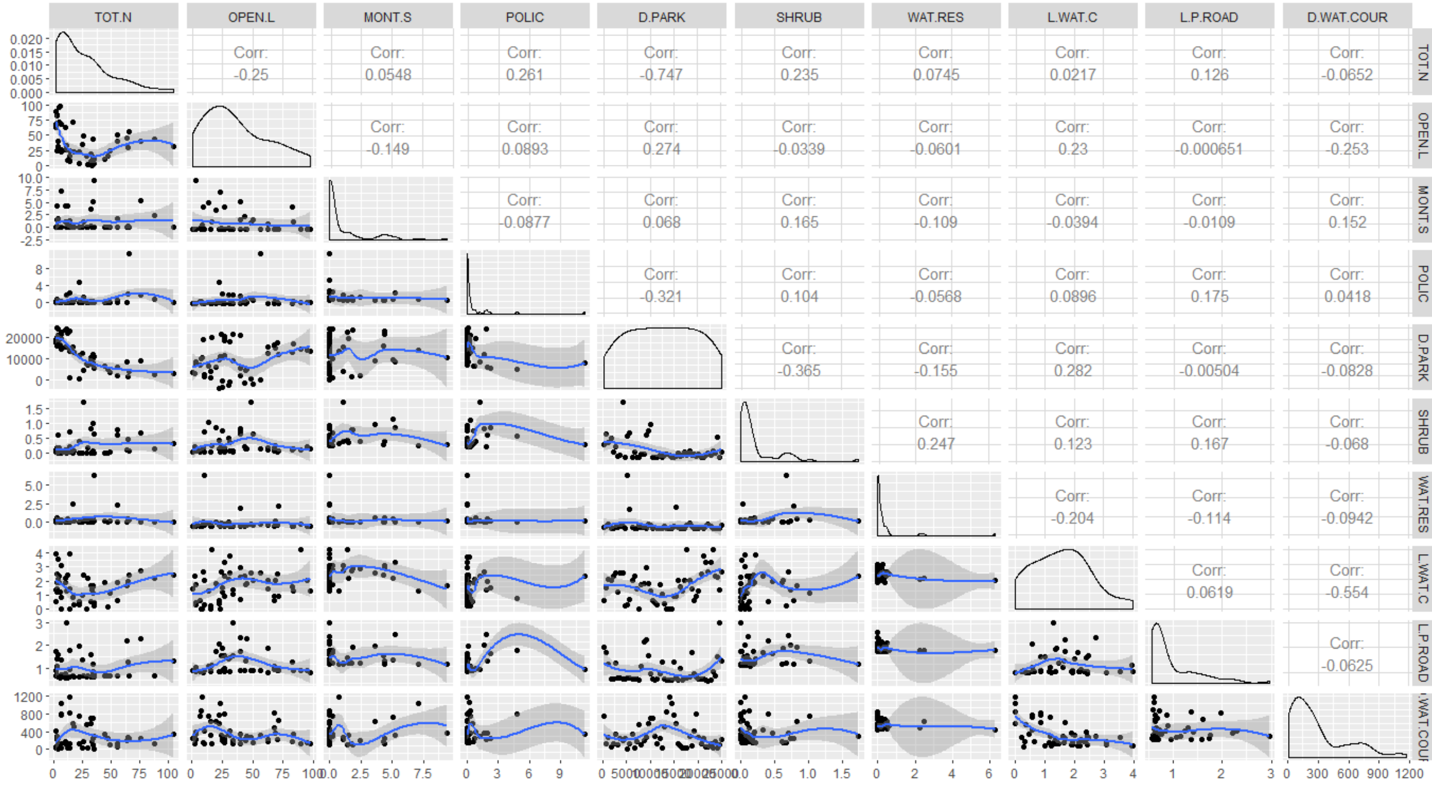

Figure A.1: Correlation charts

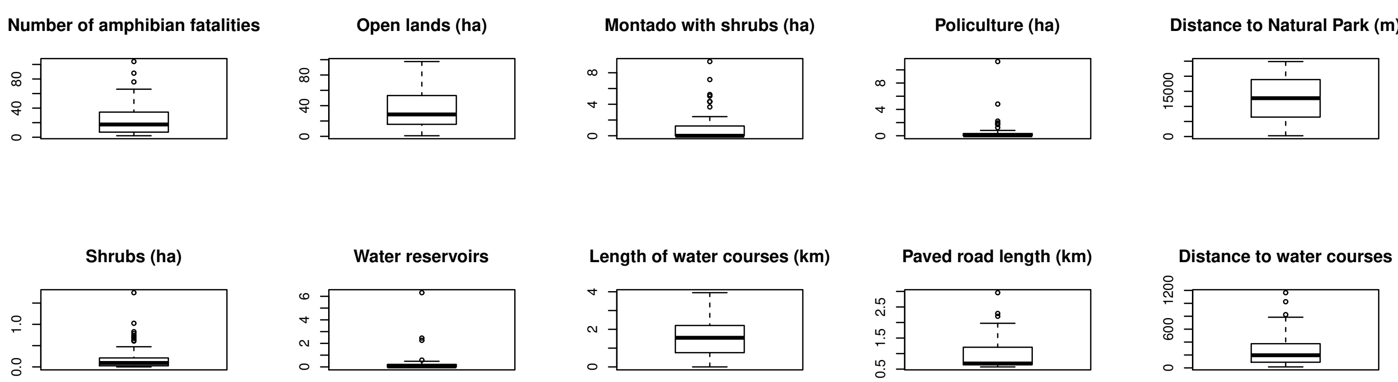

Figure A.2: Boxplots of response and regressors

Table A.1: Coefficient estimates for the final model of Poisson regression considering interactions

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.1355	0.2631	15.7178	0.0000
D.PARK	-0.0001	0.0000	-6.4825	0.0000
L.WAT.C	-0.6370	0.1172	-5.4361	0.0000
MONT.S	0.0616	0.0497	1.2388	0.2154
SHRUB	2.1096	0.3079	6.8519	0.0000
OPEN.L	0.0033	0.0063	0.5263	0.5987
POLIC	0.1184	0.0207	5.7298	0.0000
WAT.RES	-0.1319	0.0407	-3.2423	0.0012
L.P.ROAD	0.6069	0.3318	1.8294	0.0673
L.WAT.C:MONT.S	0.1529	0.0254	6.0207	0.0000
MONT.S:OPEN.L	-0.0026	0.0013	-2.0065	0.0448
L.WAT.C:L.P.ROAD	0.8885	0.1252	7.0996	0.0000
SHRUB:L.P.ROAD	-2.6917	0.3841	-7.0087	0.0000
OPEN.L:WAT.RES	0.0066	0.0017	3.9259	0.0001
SHRUB:POLIC	-0.6728	0.1147	-5.8635	0.0000
D.PARK:L.P.ROAD	-0.0000	0.0000	-2.6150	0.0089
OPEN.L:L.P.ROAD	-0.0192	0.0080	-2.4006	0.0164
D.PARK:MONT.S	-0.0000	0.0000	-1.8082	0.0706

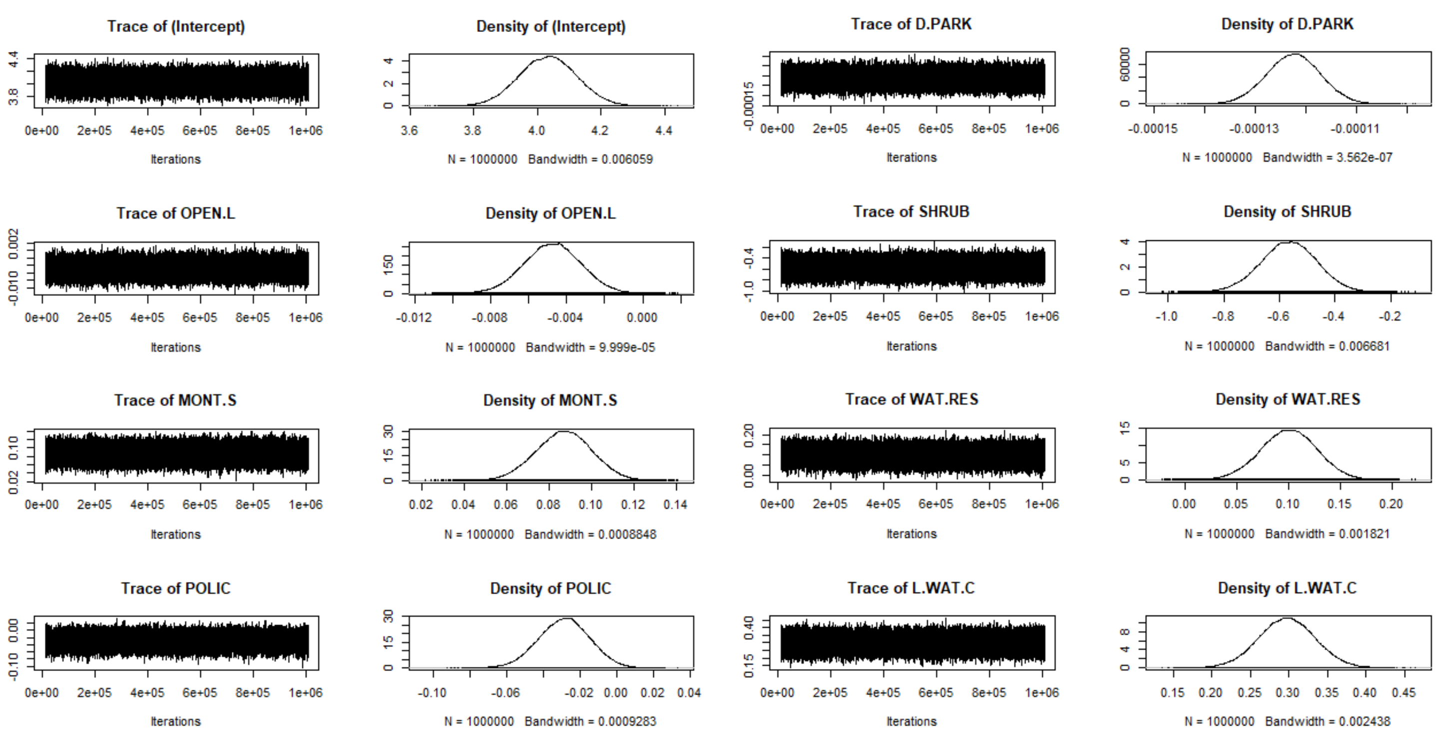

Figure A.3: Trace plots