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).

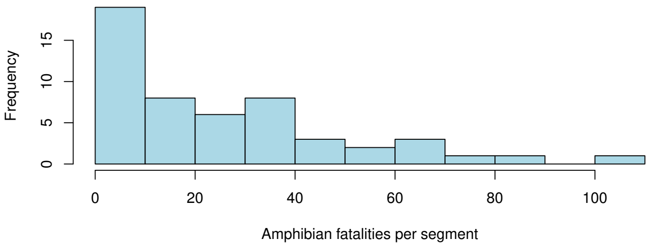


Figure 1: Histogram of response

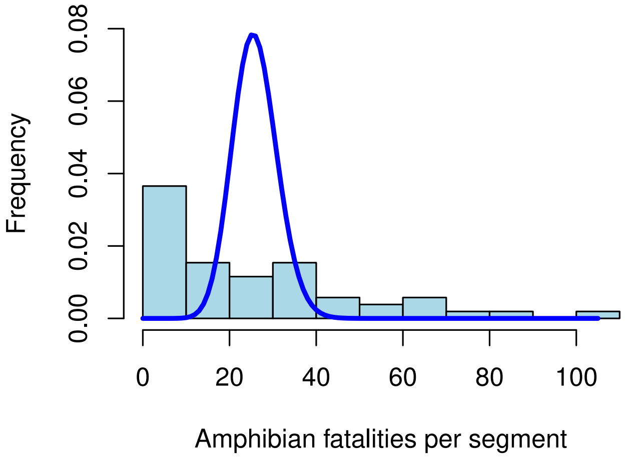


Figure 2: Histogram of frequencies and fitted Poisson distribution of the null model

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 ($6.309ha$).

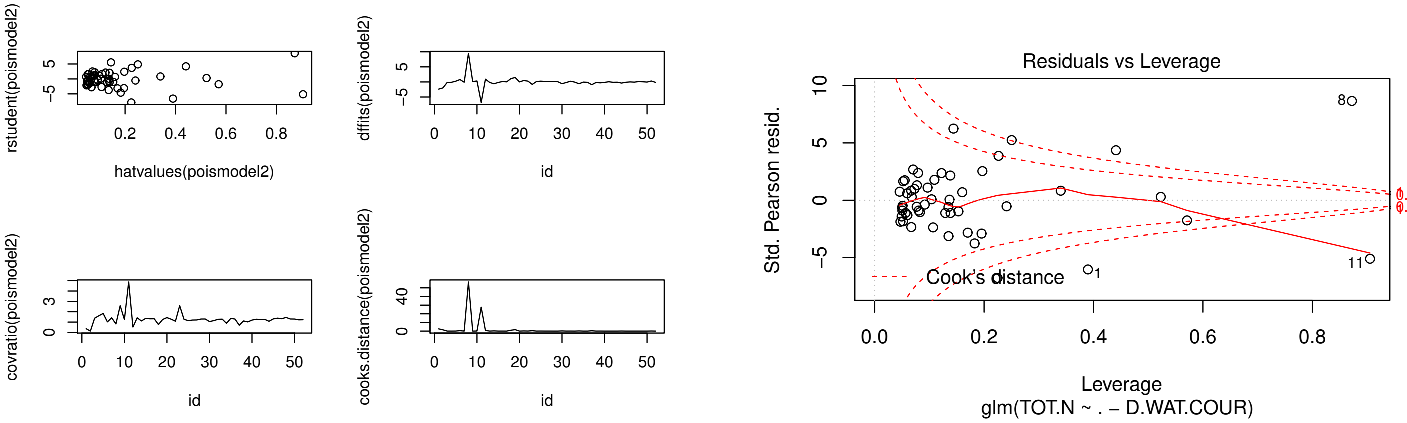


Figure 3: Influence plot for Poisson regression

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.

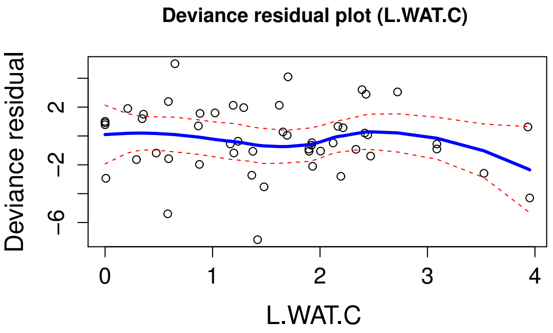


Figure 4: Deviance residual plot for Length of water courses (km)

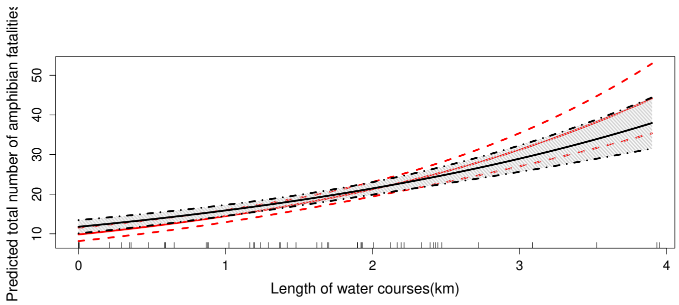


Figure 5: Predicted responses and CIs for the Poisson regression with interactions (red) vs without interactions) (black) with OPEN.L, MONT.S, POLIC, D.PARK, SHRUB, WAT.RES, and L.P.ROAD held at their means

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.

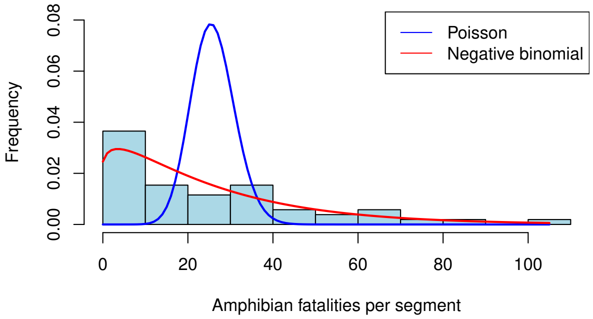


Figure 6: Histogram of frequencies, fitted Poisson distribution of the null model (blue) and fitted negative binomial distribution of the null model (red)

	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.

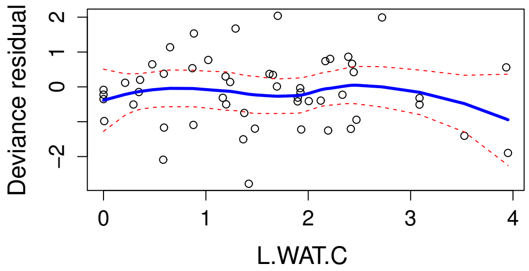


Figure 7: Deviance residual for L.WAT.C (negative binomial regression)

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.

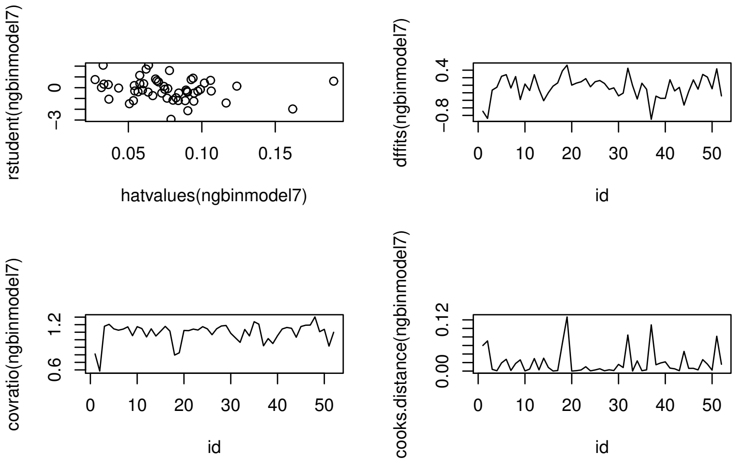


Figure 8: Influence plots for negative binomial regression

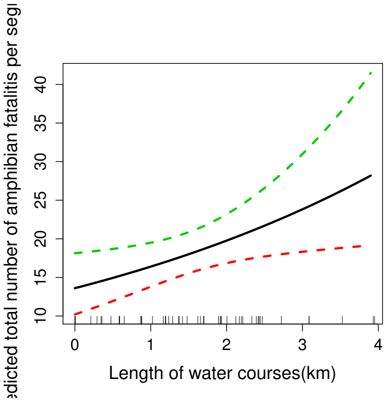


Figure 9: Prediction of TOT.N for mean OPEN.L($36.18ha$) and mean D.PARK($12680.89m$)

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

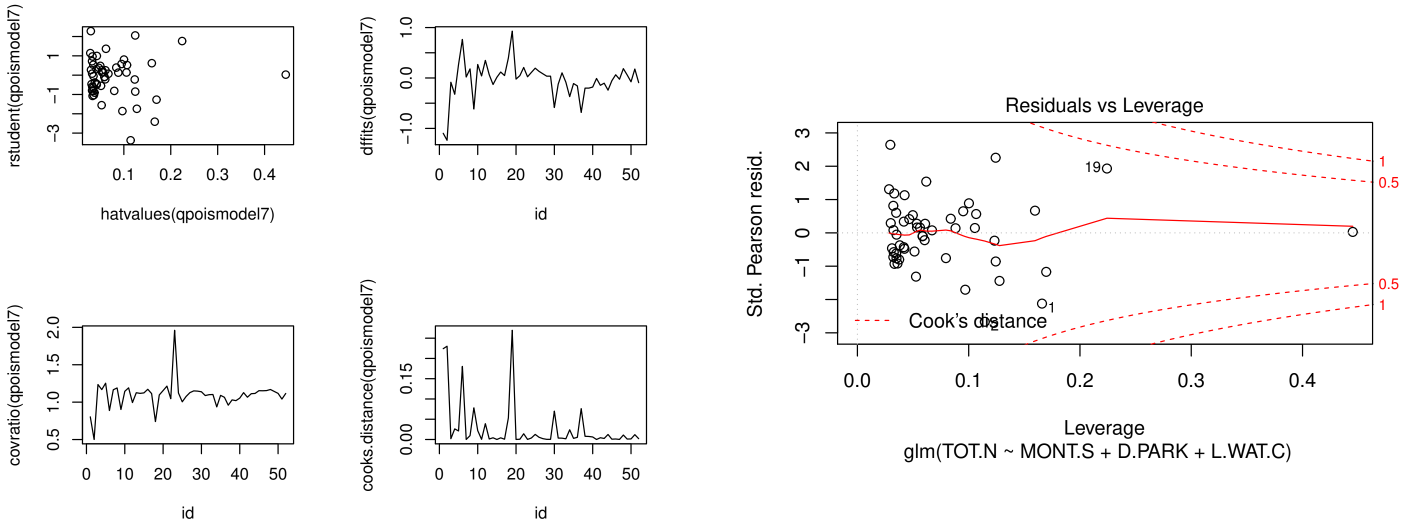


Figure 10: Influence plot for Poisson regression

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

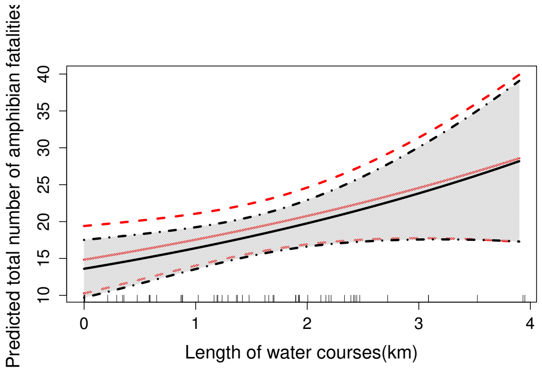


Figure 11: Predictions and CIs of means of TOT.N with MONT.S and D.PARK held at their means for Quasi-Poisson regression (red) and for negative binomial regression (black).

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.

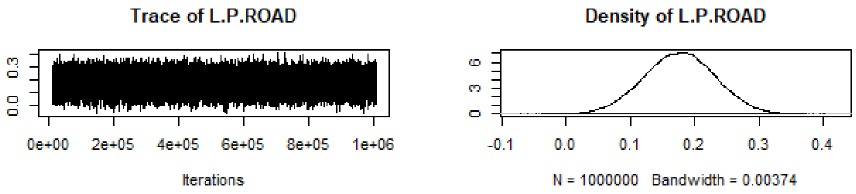


Figure 12: Trace plot (L.P.ROAD)

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.

2 Part 2 - Modeling fish growth

2.1 Introduction

In the second section of this report a data set containing information on the fish species Merluccius gayi, a species of merluccid hake, will be analyzed. The fish analyzed here were captured in Chile in the years 1984 and 2010. First, a descriptive analyses will be carried out, then different models will be constructed in an attempt to capture the relationship between the length-at-capture and age of the fish. Next, a model including all covariates will be constructed to answer two research questions: Firstly, if fish captured between July and September (end of winter in Chile) are on average smaller than the fish captured during the rest of the year. Secondly, we will try to find out if the average length of the fish captured in 2010 is different from fish captured in 1984. Lastly, fish that is of size greater or equal to 30 cm is considered of commercial interest. Therefore a model will be constructed that relates the probability of catching a fish of that size and other covariates in the data set. A description of the response variable of interest and the covariates used for this analysis is given below.

Response variable:	Length:	Length-at-capture in centimeters
Covariates:	Year:	Year of capture (1984 or 2010)
	Month:	Month of capture: 1-12
	Age	Age-at-capture in years
	Sex:	Male, female

2.2 Descriptive analysis

To get a preliminary idea of the relationship between the covariates and the response variable[1], length of capture, plots showing the effect of each covariate separately on the response variable were constructed. A indication of a nonlinear relationship between two covariates *age* and *month* and the response variable *Length* can be seen from figures 1 and 2. As can be seen the length of fish increases relatively fast at first and and then slows down. A more 'wiggly' relationship is observed between length and month.

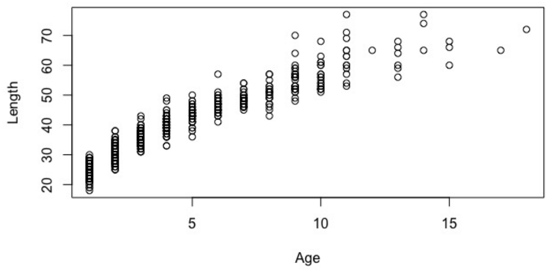


Figure 1: Length vs. Age

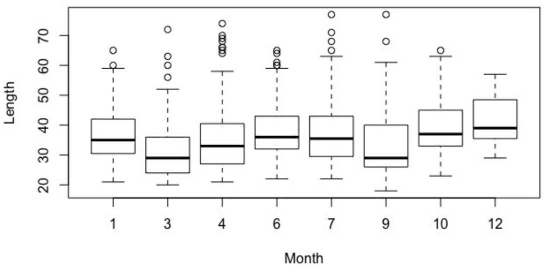


Figure 2: Length vs. Month

The relationship of *length* and the two binary covariates in the data set, *sex* and *year*, is shown for each covariate separately in figures 3 and 4. As shown, female fish seem to be larger on average then male fish and there is also an indication that fish captured in 1984 were on average larger then the fish captured in 2010.

[1] 1. Make a descriptive analysis to look at the relationship between the covariates and the length-at-capture.

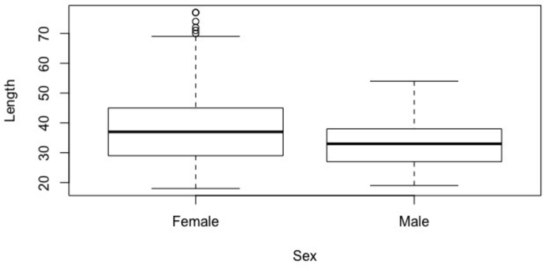


Figure 3: Length vs. Sex

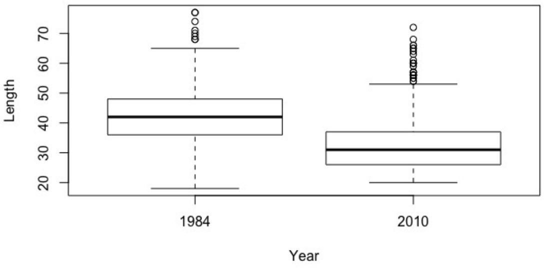


Figure 4: Length vs. Year

To further investigate if the relationship between the two non-binary covariates, i.e. age and month, and the response variable is linear or not an ordinal logistic regression model was constructed using the *polr* function in R. To construct the model the continuous response variable was recoded as a categorical variable and the covariates were split up into several dummy variables. Again nonlinearity is observed in the relationship between length and age of fish as the 'steps' are unequal. The growth rate of fish is high for the first period of it's life span, as can be seen from the first two 'steps', and then slows down, indicated by the relatively smaller final 'step'. A more complicated relationship is observed between length of fish and the month of capture, the steps are again unequal and go up and down.

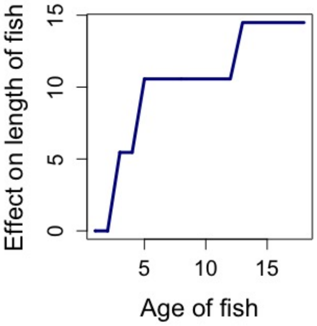


Figure 5: Length vs. age

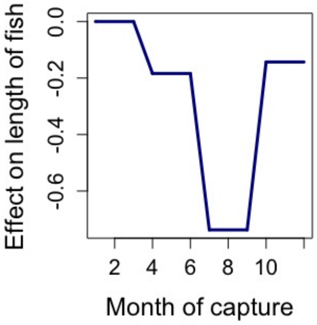


Figure 6: Length vs. month

Finally, to investigate the relation between the age and month further categorical and continuous interactions were constructed. A graph showing the continuous interaction, refereed to as statistical interaction, for the variables is given in figure 7. From the figure one can't see strong evidence for the presence of a interaction between the two variables although the difference in length between months seems to increase slightly as age increases. Continuous interaction, is linear which can be too simplistic for the data at hand. Therefore a categorical interaction of a more general type was also constructed. As can be seen from figure 8, there is some indication of a nonlinear interaction between the two variables.

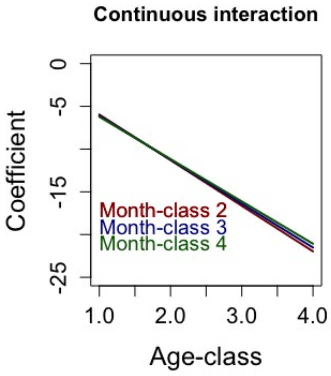


Figure 7: Length vs. age

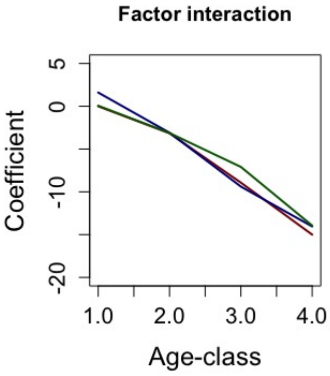


Figure 8: Length vs. month

2.3 Modeling length-at-capture as a function of age

In the previous section a nonlinear relationship between the covariate *age* and the response variable, *length*, was observed. Next, four smoothing techniques will be performed in attempt to adequately model the relationship between the variables[2]. Then, the different models will be compared and a optimal model chosen for further analysis.

2.3.1 Polynomial regression model

Polynomial regression models of various degrees were fitted, the resulting fits are shown in figure 9. As can be seen in the figure the different fits seem to follow the pattern in the majority of the data points, however the higher degree polynomials get more wiggly in interval where there are relatively few data points (after age≈12).

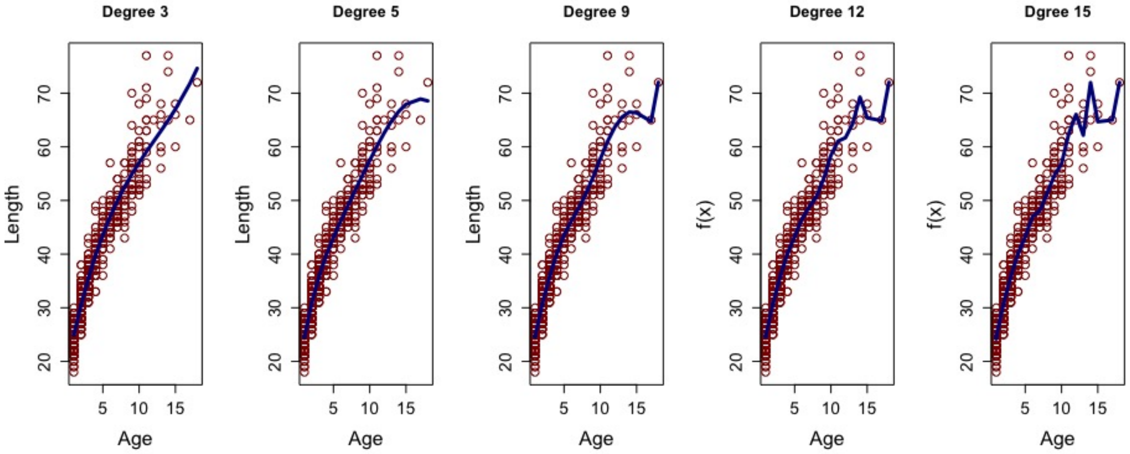


Figure 9: Polynomial of various degrees

2.3.2 Truncated polynomial splines

Next truncated polynomial splines of degree 2 were fitted. Truncated polynomial splines allows the use of a local polynomial defined by knots and also fulfill the demand for global smoothness by implementing a

[2] Model the length-at-capture as a function of age only, use the following techniques:Polynomial regression model, Truncated polynomial splines of degree 2(consider k=2, 3 and 5 knots), B-splines of degree 2 (consider m=3, 5 and 8 knots), Cubic P-splines (consider k=5, 8 and 20 knots)

global polynomial that applies for the whole interval of the data. The choice of the number of knots affects the fit. Functions with a small number of knots can lead to a very smooth fit that doesn't fit the data well and functions with large number of knots can provide a very rough estimate that are difficult to interpret. Here 2, 3 and 5 knots were fitted. Only the global polynomial affects the fit when two knots are chosen because the two knots define the interval [1,18] but no knots are positioned within that interval. When the third knot was added it was decided to position the knots with equal distance between them along the age axis, therefore the knots are positioned at the values [1, 9, 18]. The same principle was used when fitting five knots. However, there are few fish that have age larger than ≈14. Intervals with such sparse data can affect the fit of the polynomial. Therefore it was decided to alter the equidistant principle slightly and fit knots at the values [1,4,9,12,18]. The resulting fits are shown in figure 10.

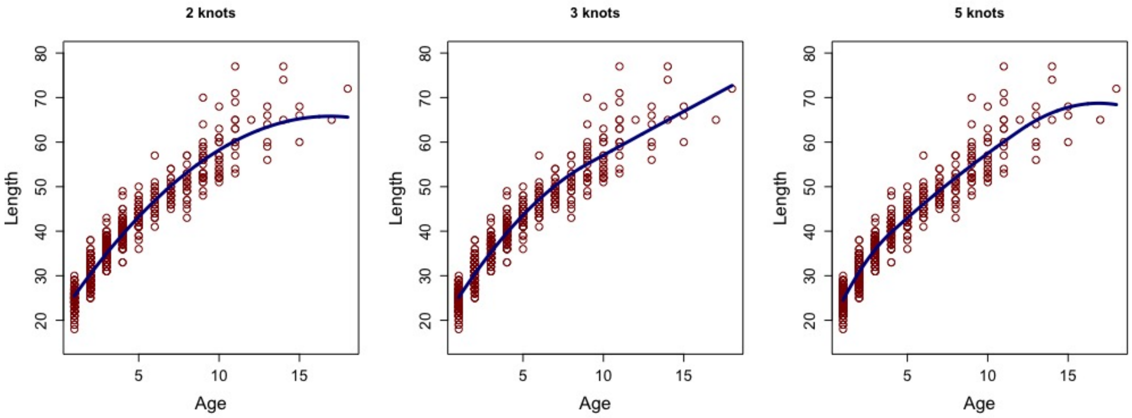


Figure 10: Truncated polynomial splines of various degrees

2.3.3 B-splines

The B-splines form a local basis where each basis function is positive in an interval defined by the l+2 knots next to it. Again equidistant distribution of knots was chosen, in that case all basis functions have the same form and are shifted along the covariate axis and then combined in an (l-1) differentiable manner to obtain the overall function. Here polynomials of degree 2 were fitted using 3, 5 and 8 knots. Plots of the resulting fits are shown in figure 11. We can see that the function with 8 knots better captures the relationship between the variables then when three and five knots are used.

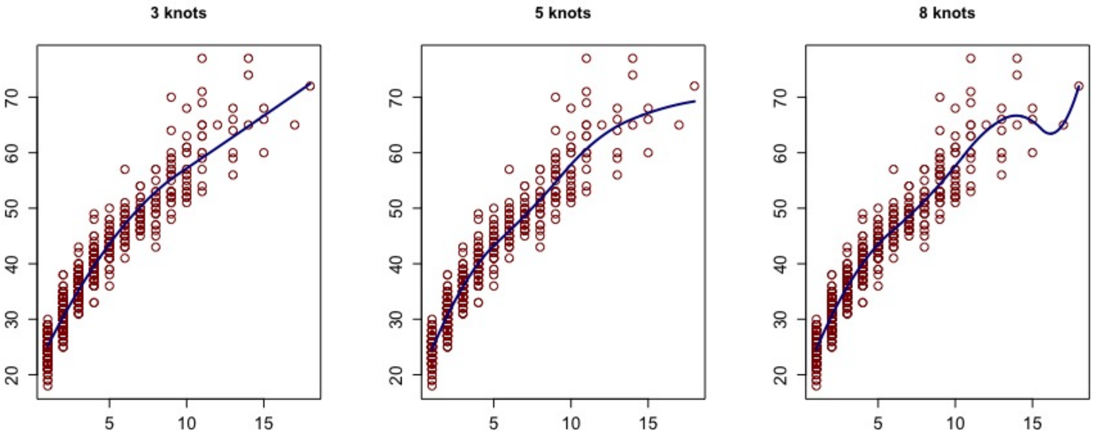


Figure 11: B-splines with 3, 5 and 8 knots

2.3.4 Cubic P-splines

Finally, cubic penalized splines (P-splines) were fitted. The advantage of the technique is that the smoothness of the estimated fit is no longer controlled by the number and the position of knots but rather by the smoothing parameter. The smoothing parameter controls the influence of the penalty and therefore an optimal value of the smoothing parameter needs to be estimated. Here, P-splines with 5, 8 and 20 knots were fitted, the resulting fit and their optimal smoothing parameter (lambda) is given in figure 12. The best fit is obtained when using 20 knots, this also seems to model the relationship between the variables better then all the previous methods.

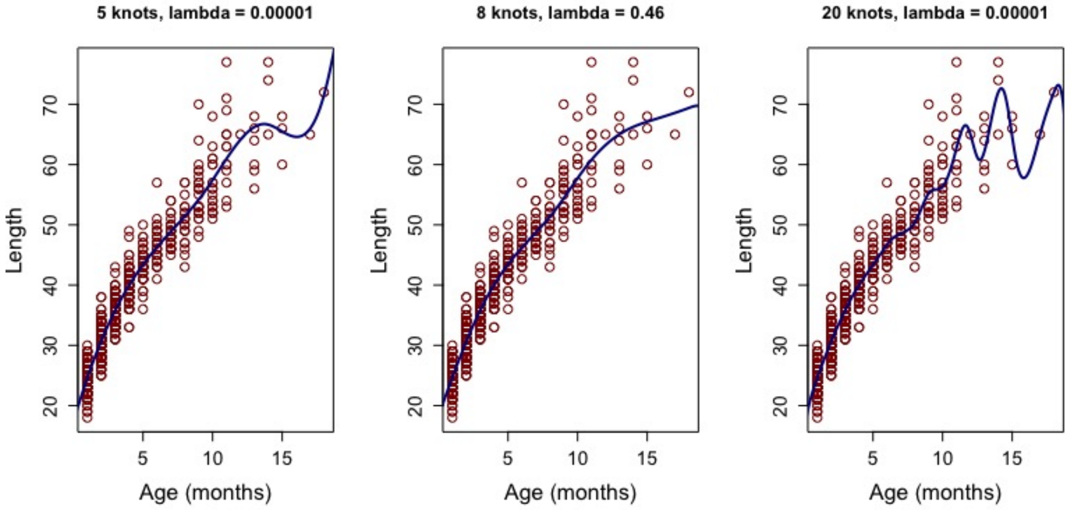


Figure 12: P-splines with 5, 8 and 20 knots

2.3.5 Selection of best model

To compare the different models the AIC of each model was computed. As sown in table 1 the optimal model based on AIC is the cubic P-splines model with 20 knots as it has the lowest value of AIC. It is also found to be the model that most effectively captures the relation between the two variable from observing the different graphs. Therefore, this model was chosen as the optimal one.

Model	AIC
Poly2	4417.878
Poly3	4377.38
Poly5	4348.58
Poly9	4342.74
Poly12	4339.59
Poly15	4319.75
Degree 2 TP2	4417.88
Degree 2 TP3	4393.88
Degree 2 TP5	4350.23
Degree 2 BS3	1864.19
Degree 2 BS5	1807.91
Degree 2 BS8	1806.08
Cubic PS5	1805.52
Cubic PS8	1805.07
Cubic PS20	1781.38

Table 1: AIC's of all models

2.4 Answering the research questions

The following two questions are of interest:

1. Is the fish captured between July and September on average smaller compared to the rest of the year?

2. Is the average length of the fish captured in 2010 different from that captured in 1984?

To answer the two research questions the three additional covariate, i.e. Year, Month and Sex, were added to the model[3]. It is important not to omit important explanatory variables since it can cause bias in estimates and hence lead to incorrect conclusions. Id is left out as it is merely a column for

[3] Question 4 - Include the other covariates in the selected model and answer the two research questions described above

identification. To answer the first research question the covariate Month was dichotomised in the following way:

$$Month_i = \begin{cases} 1, & \text{if the month in cell i is July, August or September} \\ 0, & \text{otherwise} \end{cases}$$

The variable Year was already dichotomised so no further action was needed in that regard. Next, an appropriate model was constructed to answer the two research questions.

The Model

Since a nonlinear relationship was observed between Length and Age a generalized additive model was constructed. Since the response variable, Length, is a continuous variable it is suspected that assuming a normal distribution combined with a suitable link function may give a good model. A histogram of the response variable is given in Figure 13. As shown the distribution is slightly positively skewed.

For this reason we first attempted to fit a Poisson model with the canonical log link, using smoothed Age as covariate, along with Sex, Year and Month as binary explanatory variables. However, we could not use 20 knots when smoothing Age using a P-spline as it is required that the number of knots must be lower than the number of unique levels in the Age variable. Age only has 18 levels so 17 knots were chosen. However, since we were not properly convinced from the histogram that the distribution of Length was Poisson a negative binomial model with log link and a quasi-poisson model with log link were also constructed. Furthermore, a model assuming normal response and using the identity link function was fitted. All models were compared based on their AIC values. As shown in figure 2 the lowest AIC is obtained from the linear regression model[4].

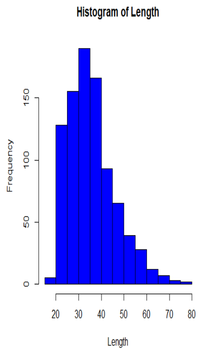


Figure 13: Histogram

Model	Poisson	Quasi-Poisson	Neg Binomial	Linear Regression
AIC	4998.18	5010.46	5002.12	4278.29

Table 2: Compare AIC's of all models

Therefore the model was chosen as the optimal model to answer our research question[5]. The model takes the form:

$$Y_i = \beta_0 + f(Age_i) + \beta_1 Year_i + \beta_2 Sex_i + \beta_3 Month_i + \epsilon_i \tag{1}$$

where month is defined as mentioned earlier. Other covariates are defined as follows:

$$Sex_i = \begin{cases} 1, & \text{if male} \\ 0, & \text{if female} \end{cases} \qquad Year_i = \begin{cases} 1, & \text{if it is 2010} \\ 0, & \text{if it is 1984} \end{cases}$$

[4] The models were also compared using plot of the response versus fitted values, all model seemed to adequately fit the data based on those plots (see apendix figure B.1)

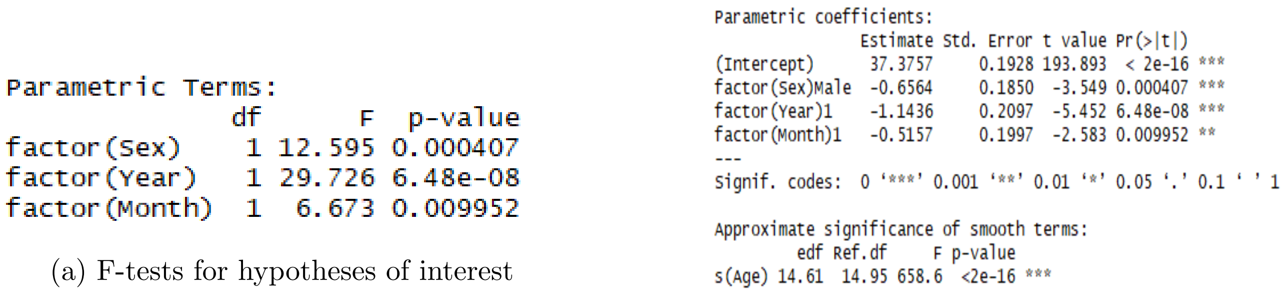
[5] **Note:** Upon analysing the fitted model it was observed that all estimated parameters were significant at the 5% level (see Figure 14b). We then performed a check that our model assumptions were satisfied before we attempted to answer the research questions and found that from checking the QQplot the assumption of normality was not satisfied. This issue will be addressed in the next question but no further action was taken to address this for the two questions at hand.

Research Question 1

Because of the parameterisation used, formally testing whether the fish captured between July and September are on average shorter than fish caught in the rest of the year equates to testing: $\beta_0 + \beta_3 < \beta_0$ or more concisely:

$$H_0 : \beta_3 < 0 \qquad (2)$$

Using the ANOVA function in R we can test the hypothesis $\beta_3 = 0$. This is comfortably rejected at the 5% level (p-value=0.00995) (see Figure 14a). Since we now know that β_3 is significantly different from 0 we now should check the sign of the estimated coefficient. In Figure 14b we clearly see that the coefficient is -0.5157 meaning that on average β_3 is significantly less than 0. Therefore we can reject the original null hypothesis and conclude that the fish caught in the months of July to September are on average shorter than those caught outside of that period. This makes sense as these are the winter months in Chile which are typically colder which may cause less availability of food for fish causing them to be shorter.



```
Parametric Terms:
                df      F  p-value
factor(Sex)      1 12.595 0.000407
factor(Year)     1 29.726 6.48e-08
factor(Month)    1  6.673 0.009952
```

(a) F-tests for hypotheses of interest

```
Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    37.3757     0.1928 193.893  < 2e-16 ***
factor(Sex)Male -0.6564    0.1850  -3.549 0.000407 ***
factor(Year)1  -1.1436     0.2097  -5.452 6.48e-08 ***
factor(Month)1 -0.5157     0.1997  -2.583 0.009952 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value
s(Age) 14.61  14.95 658.6  <2e-16 ***
```

(b) Estimated parameter coefficients

Figure 14: Answering the research questions

Research Question 2

The next research question asks whether the fish captured in 2010 have an average length different from that of the fish caught in 1984. In our model this equates to testing whether $\beta_0 + \beta_1 = \beta_0$ or otherwise written as:

$$H_0 : \beta_1 = 0$$

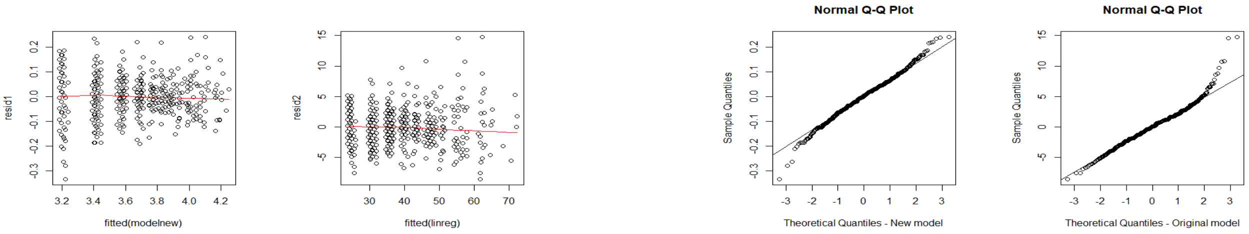
The ANOVA function was used again here as it can test this exact hypothesis. As can be seen in Figure 14a this hypothesis is rejected (p-value of $\approx$0) and we can conclude that the length of fish captured in 2010 is different from the length of those captured in 1984. This also fits well with general theory as we suspect that pollution and the change in the ocean conditions over time can impact the growth of the fish.

Taking the log of length as response

We noted that the assumption of normality was violated with our previous model. This can lead to issues when making inference based on this model. To counter this a transformed model using the logarithm of length as a response was fitted[6]. To see if this improved the models fit few things were observed. Firstly, as shown in figure 16 both models seem to adequately satisfy the assumption of homoscedasticity as seen from the straight nonparametric fit through the residuals. However, the log transformed response better satisfies the assumption of normality as shown in the QQ-plots of the two models since points better follow the line. Therefore the new transformed model is preferred.

The adjusted R^2 for the new model of 0.929 is very high. This measure gives an idea of the percentage of the variance that is explained by the covariates, penalizing for the number of covariates in the model to

[6] Question 5 - Fit the previous model but now taking as response the logarithm of length. Is this model better? You may base your decision on model diagnostics.



(a) Residuals vs fitted values (log response on left)

(b) QQ plots (log response on left)

Figure 15: Comparing the models based on how well they satisfy assumptions

avoid overfitting. This is slightly lower then the adjusted R^2 obtained from the previous model, which is 0.936. However, the difference is minimal and both models do very well in this regard. Lastly, it is useful to inspect the non-linear relationship of Age with length and log of length. This can give us an indication if our smoothing technique is working well. We can clearly see that for the new model Age and log(Length) have a much smoother relationship within the model and in fact the partial residuals seem to be closer to the blue line. This would suggest that the smoothing technique works better in the transformed response model and captures the relationship of Age and Length better.

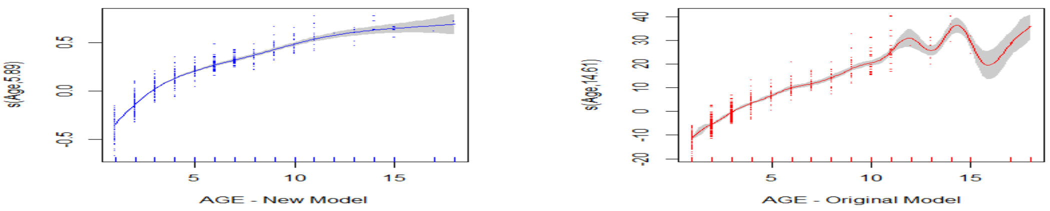


Figure 16: Relationships between Age and Length within both models - Effective degrees of freedom given in the Y label

It can be concluded that both models fit the data well suggested by their high R^2 values. However the log transformed response better satisfies the assumptions of the model and is therefore optimal. The ANOVA function was again used to check if the conclusion for the two research questions still hold, this was confirmed.

2.5 Model for plotting probability of catching a fish fit for consumption

It is also of interest to fit a model that relates the probability of catching fish fit for consumption. Fish is considered to be of commercial interest if it's length is $\geq$ 30 cm [7]. To do this a dummy variable 'Commercialfish' was created and was dichotomized in the following way:

$$Commercialfish_i = \begin{cases} 1, & \text{if the fish has length greater than or equal to 30} \\ 0, & \text{otherwise} \end{cases}$$

The goal is to model the probability of catching a fish fit for consumption. It is therefore assumed that the distributional part of our generalized linear model follows a binomial distribution as there are categorical

[7] Question 6 -Fish greater or equal than 30 cm are considered to be of commercial interest. Fit a model that relates the probability of catching a fish for consumption and the covariates considered above.

covariates in the model making $n_i > 1$. To find an appropriate link function the obtained AIC of models using different link functions were compared. The link functions compared were the 'logit', 'probit', 'Complementary log-log' and 'Cauchit'. The obtained AIC for all models were very similar, around 270. The logit function has the benefit of being more easily interpretable therefore that link function was chosen for the task at hand.

Since a non-linear relationship had been found between the covariate Age and the previous response variable Length a similar generalized additive model as before was again fitted to model the probability of catching a fish of commercial interest. Again a cubic p-spline with 17 knots was used to smoothly include the variable Age in the model. The following initial model was constructed:

$$logit(\pi_i) = \beta_0 + f(Age_i) + \beta_1 Year_i + \beta_2 Sex_i + \beta_3 Month_i \qquad (3)$$

where π_i is the probability that the fish of commercial interest.

Model building

In an attempt to simplify the model AIC of models including fewer covariates was computed. Based on the criteria it was suggested that that a model only including the covariates 'Age' and 'Month'. This model had a slightly lower value of AIC (268) then the initial model (271.3). Since the difference is only minimal the statistical significance of the two covariates, sex and year, was also observed. Both covariates were insignificant when added to the model. Therefore it was decided not to include them in the model. The possibility of an interaction term between the two remaining covariates was inspected. The interaction between the covariates was not significant and therefore not included. Hence, we selected following as our final model with January as the reference month.

$$logit(\pi_i) = \beta_0 + f(Age_i) + \beta_2 * Month2_i + \beta_3 * Month3_i + ...\beta_{12} * Month12_i \qquad (4)$$

where π_i is the probability that the fish of commercial interest.

Goodness of Fit

To check how good our model fits the data, we opted for Hosmer Lemeshow test. Hosmer Lemeshow statistic because we had the continous covariate Age in the model. This compares the observed and estimated frequencies of fish being of commercial interest for all observations in the dataset. In this test, the predicted values are arranged from lowest to highest, and then separated into 10 groups of approximately equal size. The function 'hosmlem.test' was used to perform the test in R. From that a p-value of 0.97 was obtained. Therefore we can't reject the null hypothesis that the model fits the data and can conclude that our model fits the data reasonably well.

Quality of Prediction

To measure the quality of our model in predicting the response variable, Nagelkerke's R-squared was used. To do this in R the function 'NagelkerkeR2' was used. From this we obtained a Nagelkerke R^2 of 0.86 indicating that a substantial amount (86%) of the variance in the data has been explained by our model. The concordance measure was also obtained to check the quality of our prediction. This procedure measures the number of pairs of predicted responses concord (i.e. have the same sign) with the number of pairs of observed responses. A concordance of 97.65% was obtained indicating that our model predicts the response well.

Anomalies and Influential Observations

To look for possible outliers and inspect their influence the values of Cook's distance for our observations was observed. Based on this criteria, two observations were found to be influential. These observations are number 594 (ID:79380) and 884 (ID:81513) and can be seen from from figure 17.
To see if these were in fact outliers we tried to inspect the nature of these observations. It was found that observation 594 was 1 year old at capture and had length of 30cm which was observed to be exceptionally high for fish born in April. Similarly, we found that observation 884 was 2 years old at capture and had length of 29 cm which was observed to be exceptionally low for fish born in December. Therefore, although the observations are influential there is nothing indicating that they are mismeasurements and hence were not deleted from the model.

Estimated Final Model

We can view the estimated model in Figure 18. Clearly we see that the model has significant values for only two of its levels namely '7' (July) and '9' (September). We can see that the odds of a fish being of commercial interest reduces by 0.32 and 0.12 times compared to the month of January for the fish caught during the months 'July' and 'September' respectively. This is true when fish are of the same age. Thereby, this result hints at the same result as we got for our first research questions answered previously.

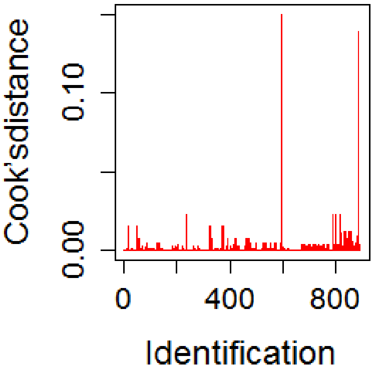
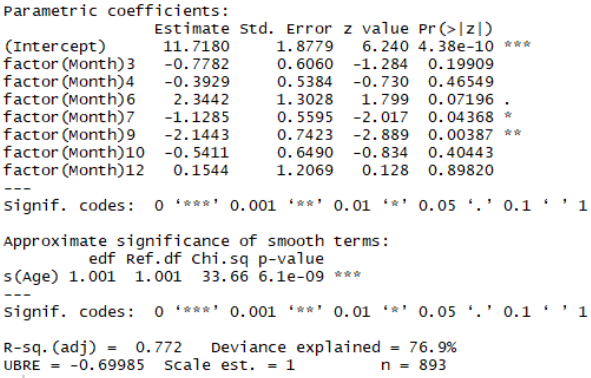


Figure 17: Cook's distance



```
Parametric coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)       11.7180     1.8779   6.240 4.38e-10 ***
factor(Month)3    -0.7782     0.6060  -1.284  0.19909    
factor(Month)4    -0.3929     0.5384  -0.730  0.46549    
factor(Month)6     2.3442     1.3028   1.799  0.07196 .  
factor(Month)7    -1.1285     0.5595  -2.017  0.04368 *  
factor(Month)9    -2.1443     0.7423  -2.889  0.00387 ** 
factor(Month)10   -0.5411     0.6490  -0.834  0.40443    
factor(Month)12    0.1544     1.2069   0.128  0.89820    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value    
s(Age) 1.001  1.001  33.66 6.1e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.772   Deviance explained = 76.9%
UBRE = -0.69985  Scale est. = 1         n = 893
```

Figure 18: Estimated logistic regression model

Response Prediction

To demonstrate the response prediction in our model, we tried to predict the probability of finding a fish of commercial interest w.r.t 'age' among 'male' fishes captured in 'April' '2010'. The resulting plot is shown in figure 19.

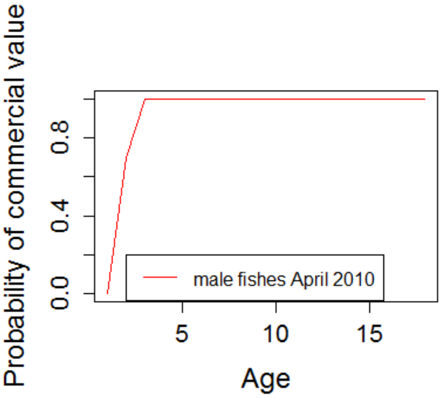


Figure 19: Predicted probability of fish

2.6 Conclusion

A general additive model was constructed to predict the probability of catching fish of commercial interest. The model constructed included two covariates, that is a non-linear function of age and the categorical covariate month. The model does a good job of predicting the probability of catching fish of commercial interest which was both indicated by a high Negelkerke's R^2 and by the

proportion of concordance. Furthermore the model fits the data
well as confirmed by the Hosmer-Lemeshow test. A more complex model was also fitted. This model included the covariates Sex and Year in addition to the covariates previously mentioned. Using this model it was concluded that fish captured in 2010 was on average not equal to those captured in 1984. Also it was found that fish captured between July and September were on average shorter then fish caught during the rest of the year. A normal distribution was assumed when constructing this model. Deviation from this assumption was observed when looking at a QQ-plot of the model, therefore a logarithmic transformation was carried out to counter this. The resulting model had a high value of adjusted R^2 indicating that it does a good job of explaining the variability in the data. As mentioned above the model includes a non-linear function of age. The non-linear function was included since the relationship between age and length was found to be nonlinear. To obtain a suitable smoothing technique to model the relationship four different techniques were performed and compared. From this comparison it was concluded that using cubic P-splines with 20 knots was optimal in capturing the relationship between the two variablesf.

Appendices

A Part1

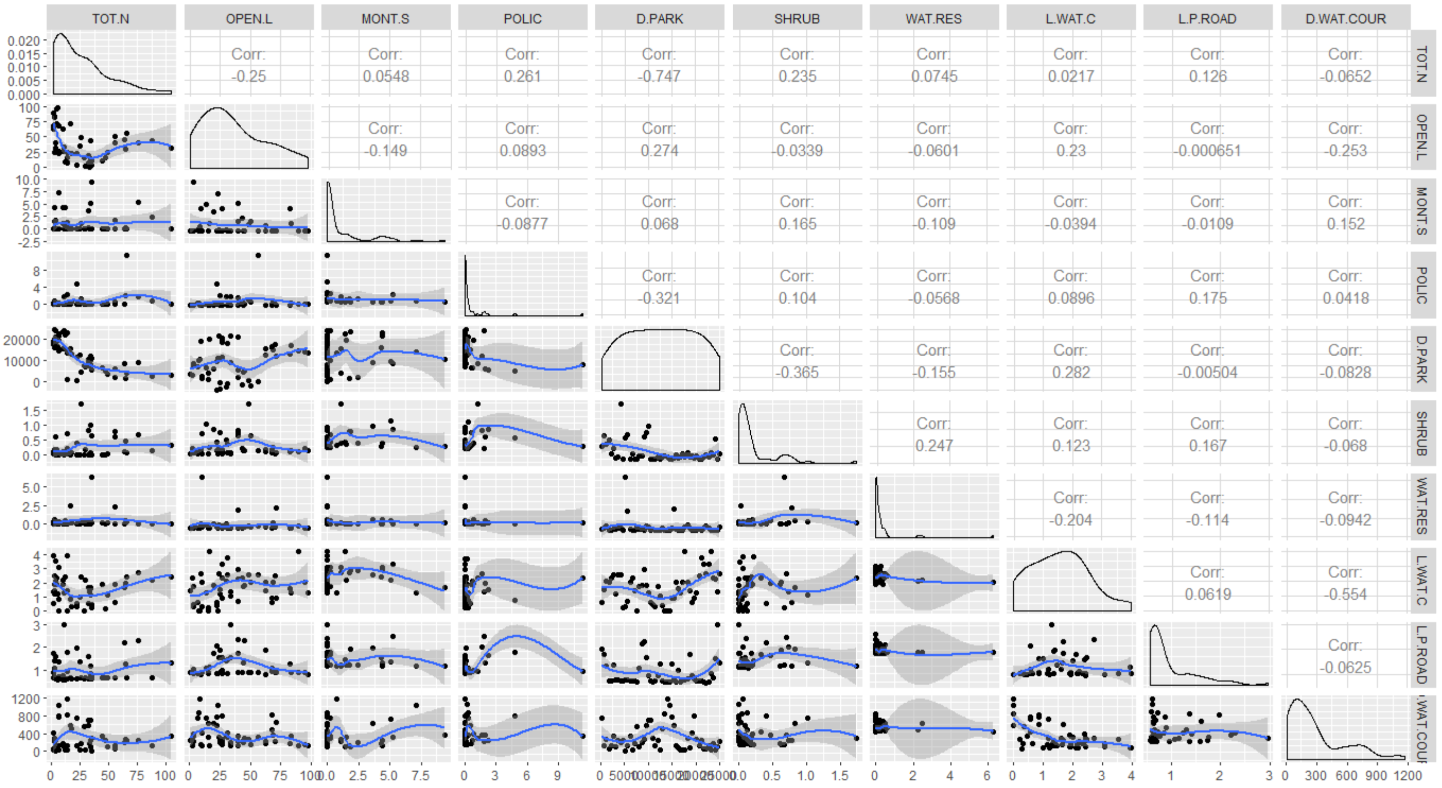


Figure A.1: Correlation charts

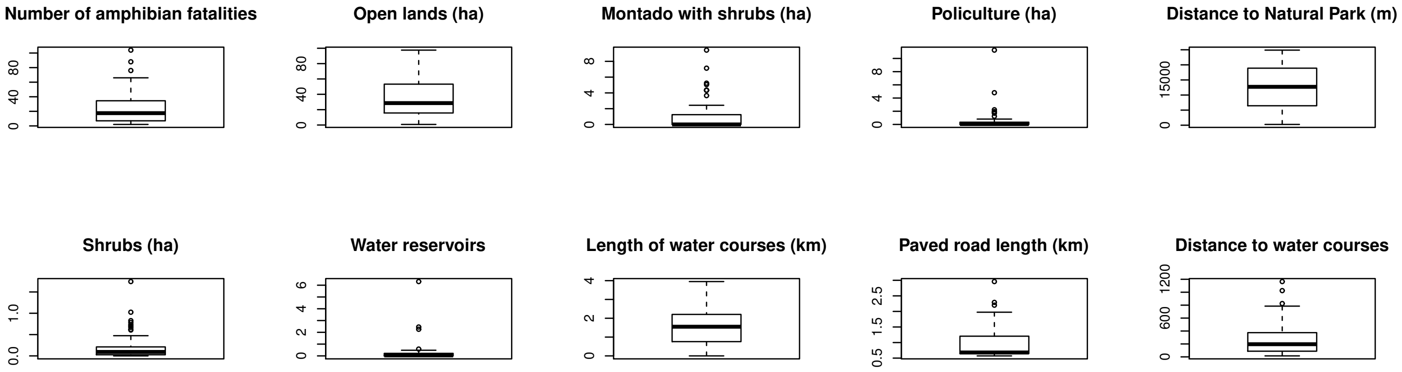


Figure A.2: Boxplots of response and regressors

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

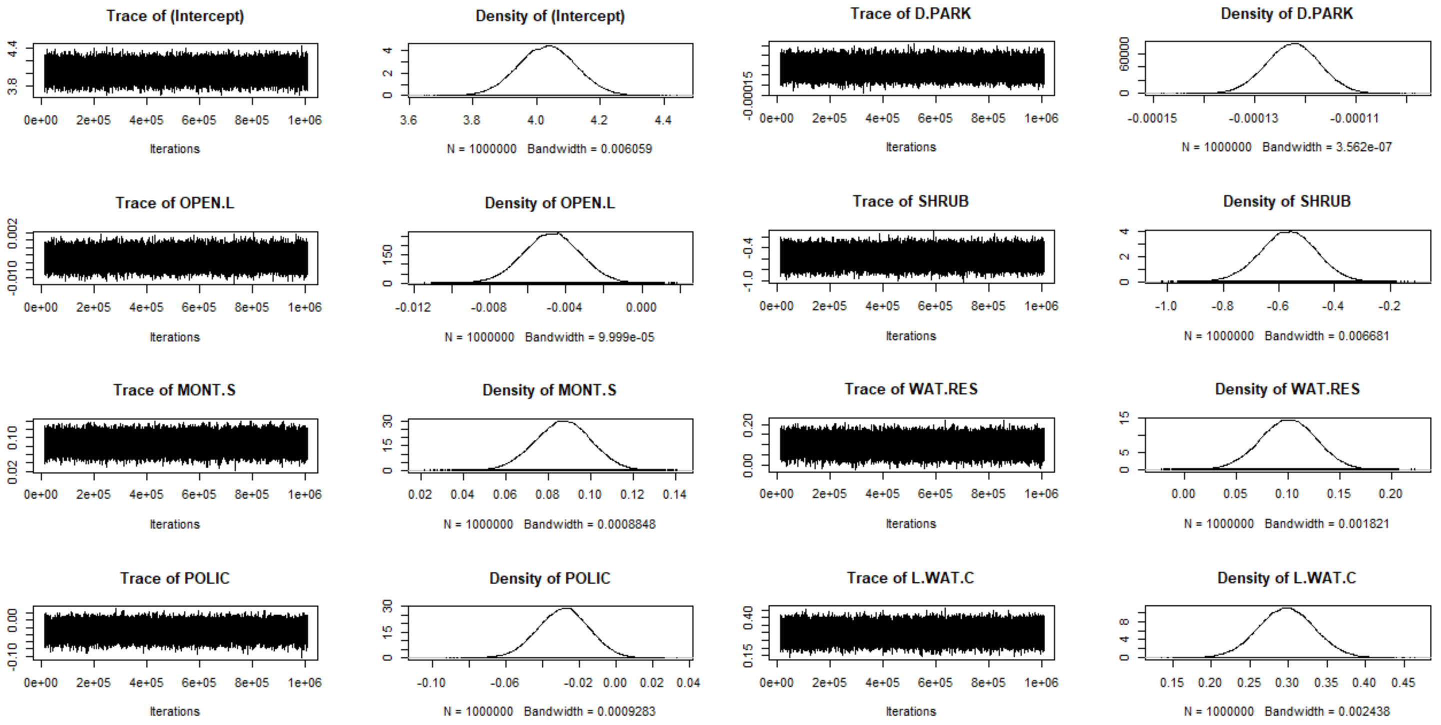


Figure A.3: Trace plots

B Part2

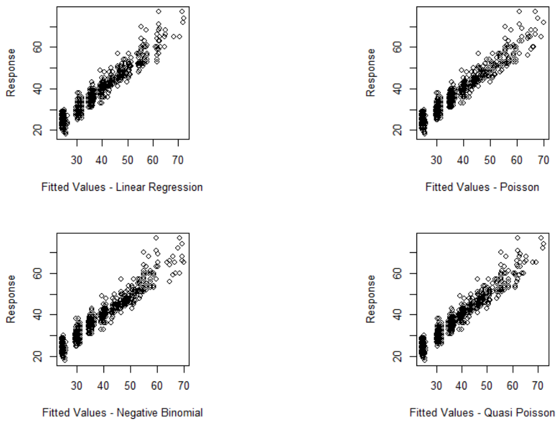
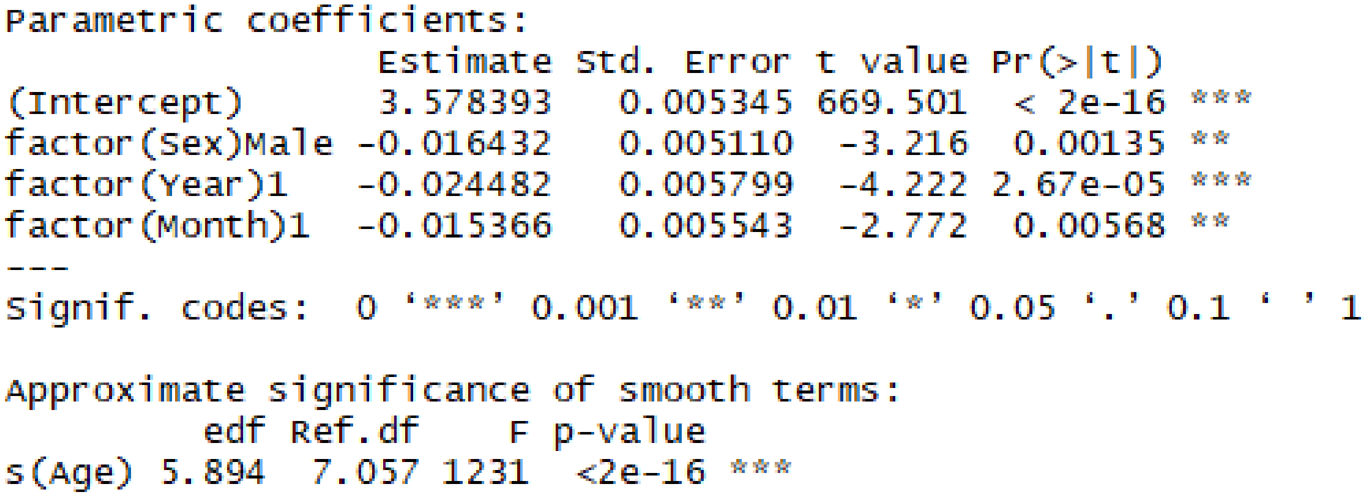


Figure B.1: Response vs Fitted values of all models



```
Parametric coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.578393   0.005345 669.501  < 2e-16 ***
factor(Sex)Male -0.016432   0.005110  -3.216  0.00135 **
factor(Year)1   -0.024482   0.005799  -4.222 2.67e-05 ***
factor(Month)1  -0.015366   0.005543  -2.772  0.00568 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df    F p-value
s(Age) 5.894  7.057 1231  <2e-16 ***
```

Figure B.2: Parameter estimates for log length as response