Assignment 1 - Generalized Linear Models (GLM)

Martin Ottosson

2024-02-18

Introduction

In this exercise the aim in part 1 is to discover whether forest amount and forest aggregation are related to species richness of forest plants and if present, what those effects are. In part 2 the effects of the same explanatory variables are analyzed on the occurence of the species Empetrum nigrum.

The data comes from a flora inventory of the historical providence of Södermanland in Sweden and it was sampled from 200 randomly selected 2.5 x 2.5 km squares. The variables are described as:

- forest - proportion of forest in a 2.5 x 2.5 km square
- forest.clumpy - aggregation index of the forest structure
- sp.richness - richness of forest species (number of species)
- Empenigr - presence/absence (1/0) of Empetrum nigrum

Method

Generalized Linear Models (GLM) can be used for statistical modeling when the response variable does not have a normal error distribution (Thulin, 2021), or in other words, a generalization of the linear model (GLM) is appropriate when the gaussian distribution cannot be assumed. In this analysis in part 1 we are going to analyze the effect of forest and forest.clumpy on sp.richness by using GLM.

Since sp.richness takes the form of count data, with non-negative integers as observed values, it is possible to use a Poisson regression model (Thulin, 2021). With this model the expected value of sp.richness, which is equal to its mean, given forest and forest.clumpy is estimated with maximum likelihood estimation.

The chosen Poisson regression model is specified as follows:

$$\log(\mathrm{E}(\text{sp.richness}_i)) = \alpha + \beta_1 \text{forest}_i + \beta_2 \text{forest.clumpy}_i$$

where

- α is the intercept term
- β_1 and β_2 are the regression coefficients for the independent variables forest_i and forest.clumpy_i respectively

The subscript i goes from 1 to n, with n being the number of observations (which is 200 in the dataset being used).

With this model the probability distribution function (PDF) is Poisson, the link function is *log* and as mentioned the goodness of fit is measured by maximum likelihood (ML), (Studiewebben, 2024).

In part 2 logistic regression will be used since the response variable is binary categorical with values 1 for presence of Empetrum nigrum and 0 for absence.

The preferred Logistic regression model is specified as:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 \text{forest}_i + \beta_2 \text{forest.clumpy}_i + \beta_3 \text{forest}_i {*} \text{forest.clumpy}_i$$

where

- α is the intercept term
- $\pi_i = 1$ when Empetrum nigrum is present and 0 when absent
- β_1 and β_2 are the regression coefficients for the independent variables forest_i and forest.clumpy_i respectively
- β_3 is the regression coefficient for the interaction term between forest_i and forest.clumpy_i

As before the subscript i goes from 1 to n, with n being the number of observations (which is 200 in the dataset being used).

The PDF for this model is binomial with a logit link function and logistic regression, being part of the GLM family, uses the ML estimator as well.

Result

Part 1

Visualizations of the relationship between each of the explanatory variables and the response variable give indications of the effects which are going to be estimated with the regression models. Therefore the analysis begins with some scatterplots.

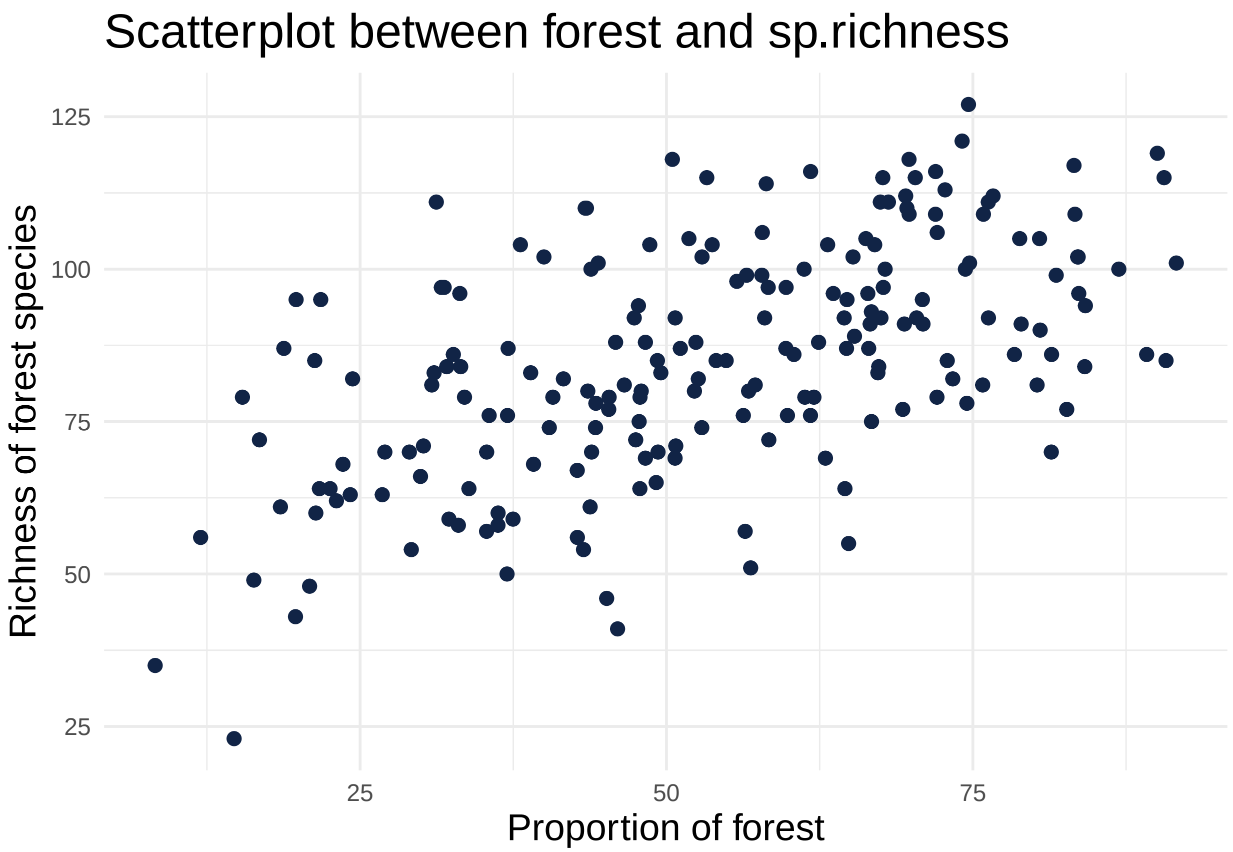

Figure 1

There seem to be a clear positive relationship between forest and sp.richness according to figure 1. The proportion of forest per 2.5 x 2.5 square km might indeed be important for explaining the number of species.

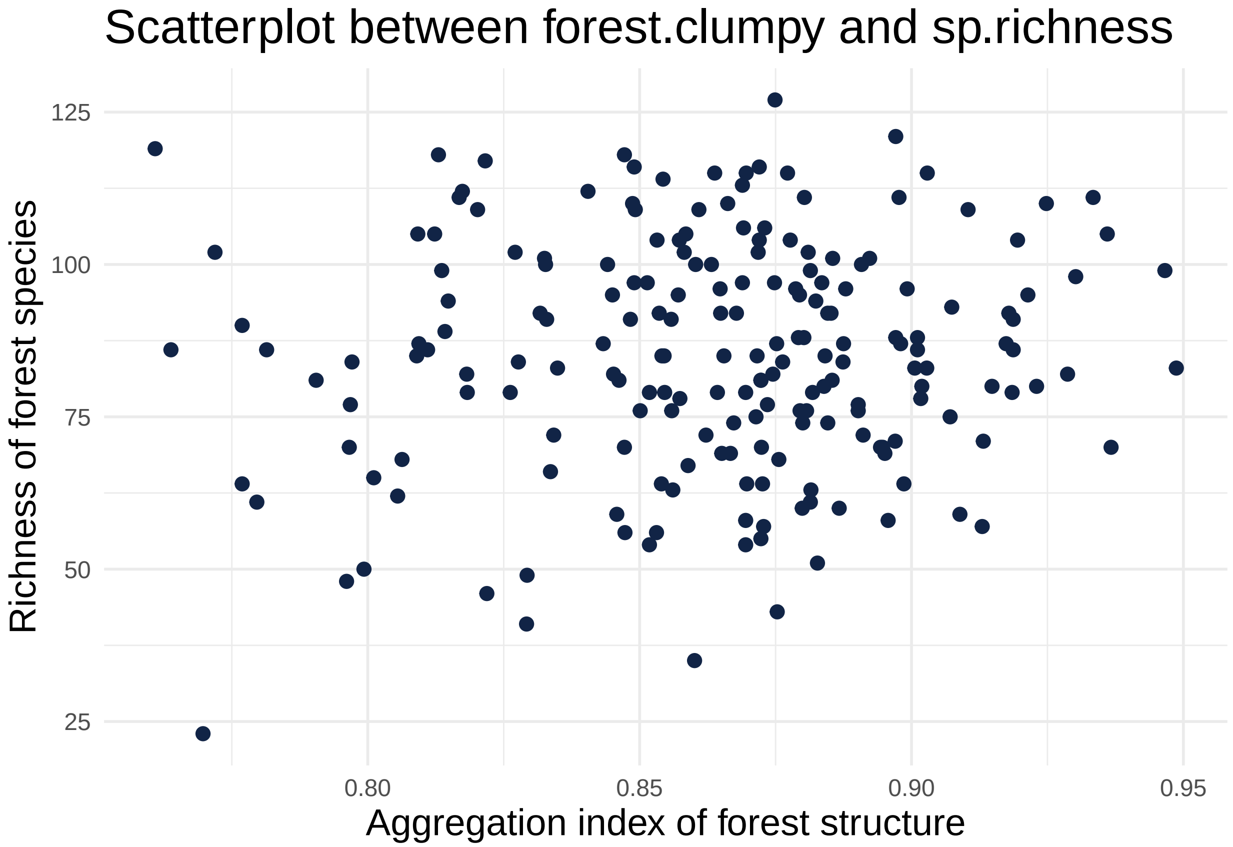

Figure 2

The relationship between forest.clumpy and sp.richness in figure 2 is not as clear, but we could suspect a positive relationship between the aggregation index of the forest structure and the amount of forest species.

We can first try an ordinary linear model estimated with ordinary least square (OLS) where the residuals are assumed to be normally distributed, thus OLS will be the maximum likelihood estimator.

Table 1 - Regression output for OLS-model

term	estimate	std.error	statistic	p.value
(Intercept)	-29.3162942	26.1141450	-1.122621	0.2629647
forest	0.5856456	0.0567975	10.311110	0.0000000
forest.clumpy	96.1547823	29.3421977	3.277014	0.0012399

The model has significant coefficient estimates for the explanatory variables at the 5%-level and R-Squared is 0.3544, but how about the assumptions underlying OLS? Diagnostic plots can reveal whether the assumptions hold.

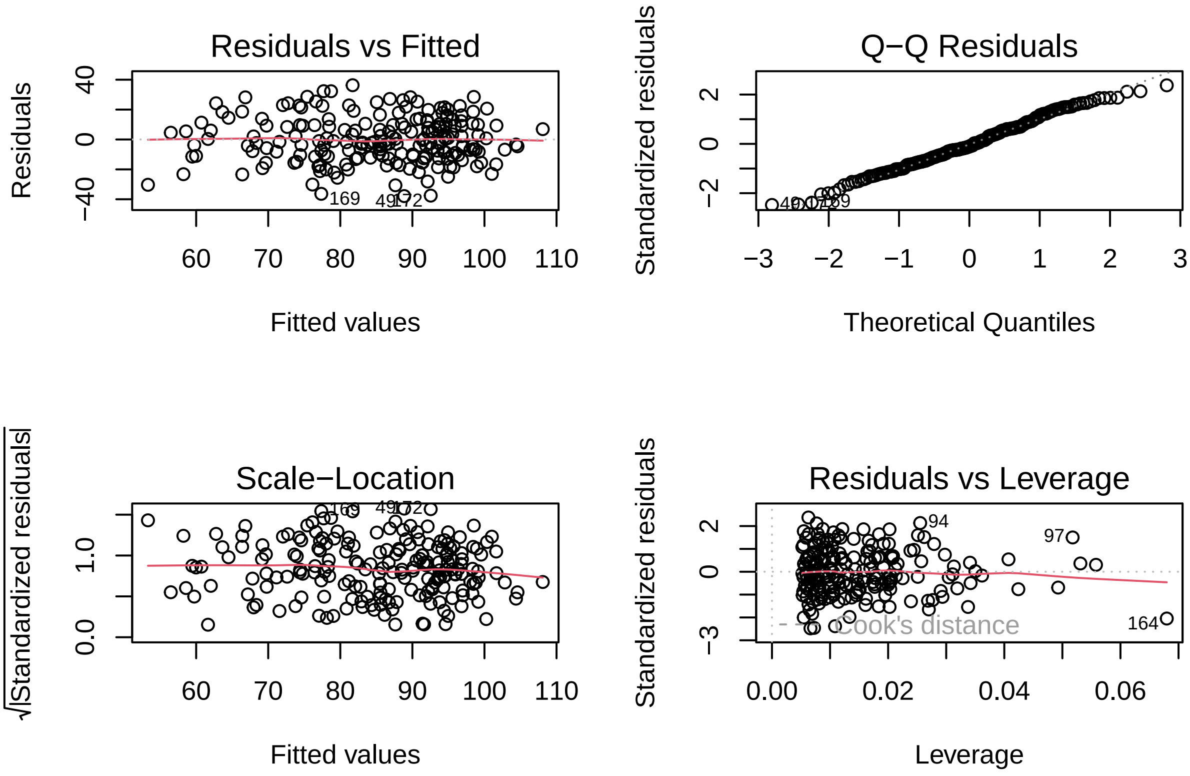

Figure 3 - Diagnostic plots for OLS-model

The Q-Q plot in figure 3 does not seem to follow the straight line very well. Instead we will run a Poisson regression with log as the link function. With this kind of model another error structure than the normal distribution is allowed.

Table 2 - Regression output for Poisson model

term	estimate	std.error	statistic	p.value
(Intercept)	2.9556051	0.1892674	15.616025	0
forest	0.0071777	0.0004181	17.165503	0
forest.clumpy	1.2671322	0.2103598	6.023642	0

In table 2 we can see that all variables have significant coefficient estimates and they are both positive (just as expected from figure 1 and 2).

The residual variance is 579.88 and there are 197 degrees of freedom. The ratio of those two should be approximately equal to 1 and not be larger than 1.2. If the ratio is larger, then the variance is increasing with the mean and the model has problems with overdispersion, which should not be the case with a poisson distribution. In this case we have:

$$\frac{\text{Residual variance}}{df} = \frac{579.88}{197} \approx 2.9$$

The model is overdispersed and we can adjust for it by changing the method from Poisson to Quasipoisson.

Table 3 - Regression output for Quasipoisson model

term	estimate	std.error	statistic	p.value
(Intercept)	2.9556051	0.3219800	9.179468	0.0000000
forest	0.0071777	0.0007114	10.090288	0.0000000
forest.clumpy	1.2671322	0.3578621	3.540839	0.0004977

The estimates are as expected still the same with the Quasipoisson method (table 3) as they were with the Poisson method (table 2), but with the added dispersion parameter there are changes in the standard errors and the p-values. All p-values are still below 0.05 so this looks as a good model.

Finally, we also want to look at the effect plots.

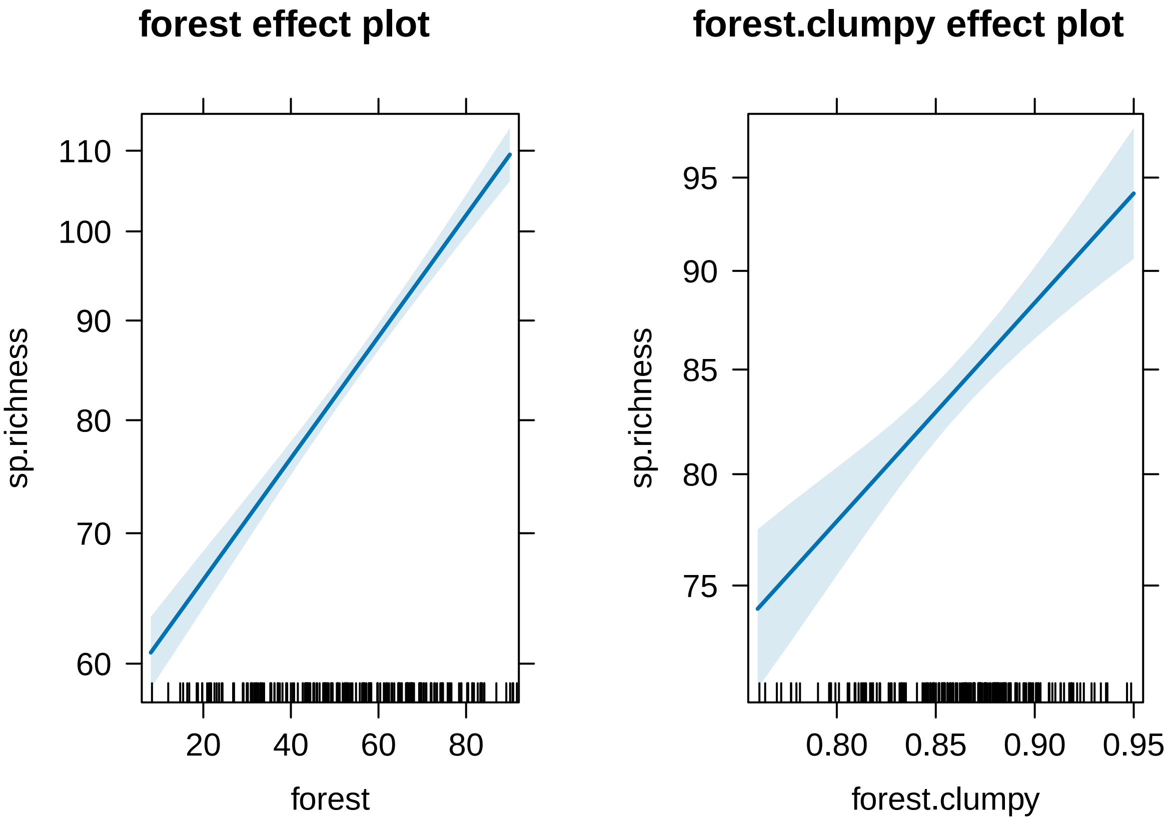

Figure 4 - Effect plots for Poisson model (y-axes at log scale)

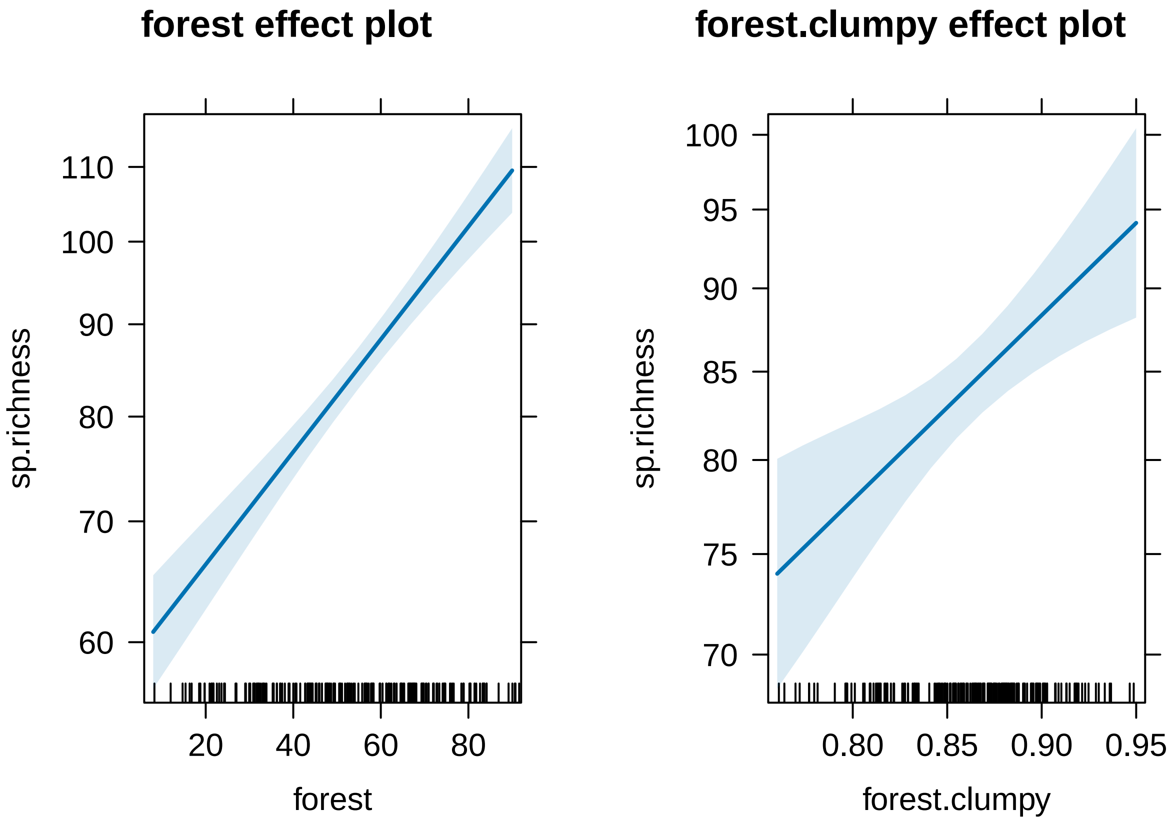

Figure 5 - Effect plots for Quasipoisson model (y-axes at log scale)

Both variables show a positive trend as seen in figure 4 and 5. The variable forest demonstrates narrower 95% confidence bands compared to forest.clumpy. When comparing the two models, the preferred model from table 3 has broader confidence bands compared to the model from table 2, meaning a somewhat higher degree of uncertainty. However, the quaispoisson model is a better choice since it does not suffer from overdispersion.

The effect plots in figure 4 and 5 have their y-axes at log scale because of the link function. Figure 6 and 7 show the plots at the original arithmetic scale, which might give a more approachable interpretation.

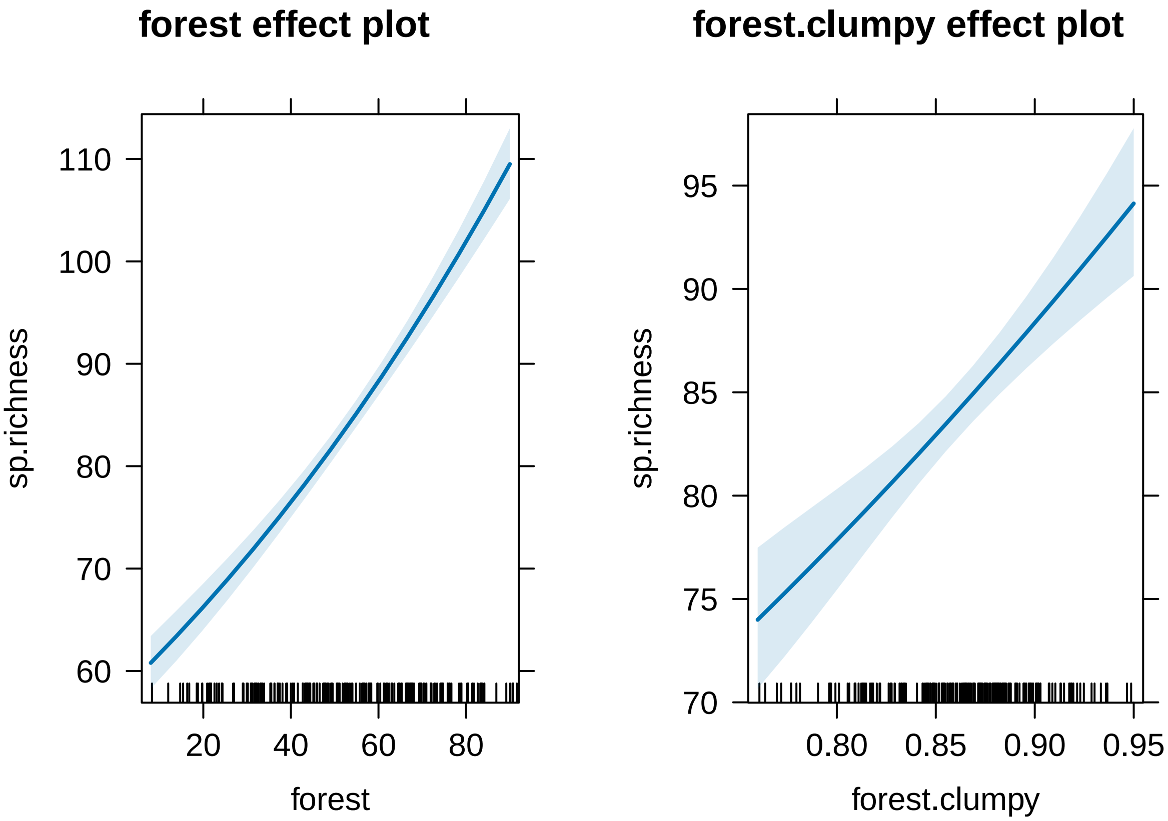

Figure 6 - Effect plots for Poisson model (y-axes at original arithmetic scale)

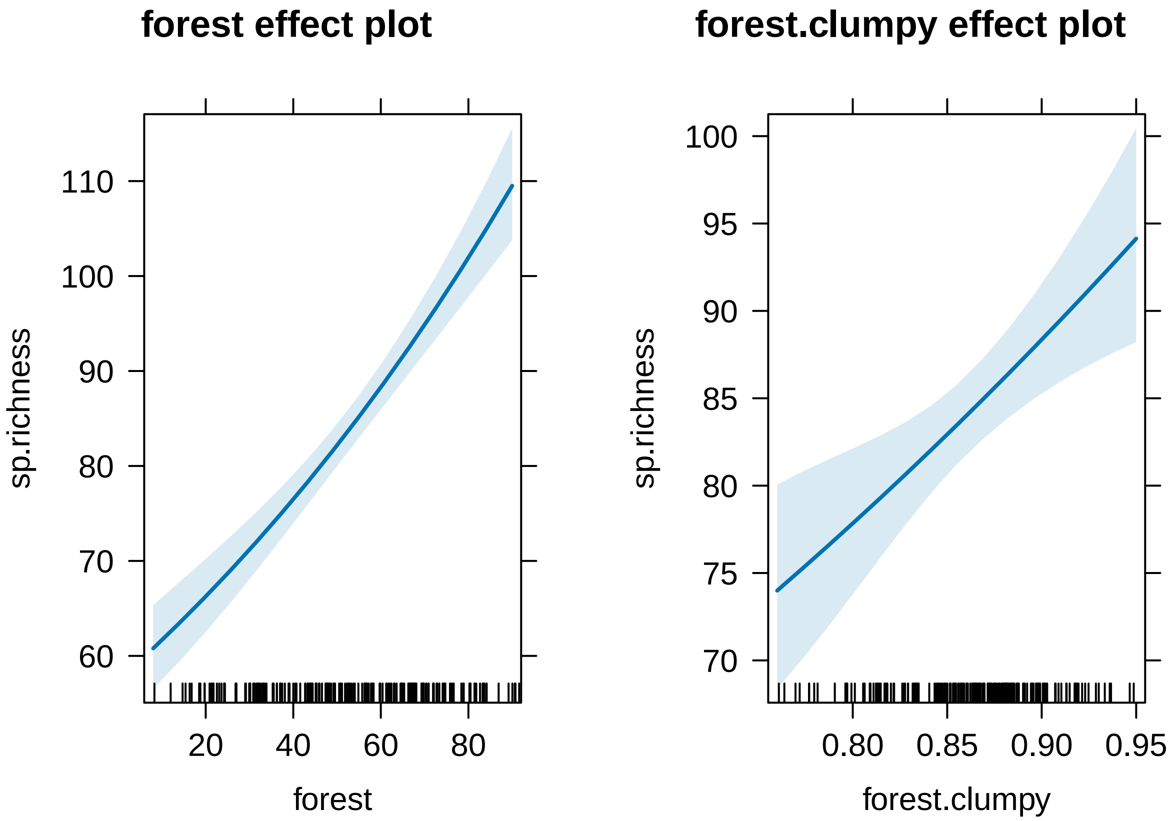

Figure 7 - Effect plots for Quasioisson model (y-axes at original arithmetic scale)

Another question is if there is an interactive effect between forest and forest.clumpy?

Table 4 - Regression output for Poisson model with an interaction term

term	estimate	std.error	statistic	p.value
(Intercept)	1.9031426	0.5269951	3.611310	0.0003047
forest	0.0253773	0.0084898	2.989164	0.0027974
forest.clumpy	2.5026639	0.6141487	4.075013	0.0000460
forest:forest.clumpy	-0.0214570	0.0099951	-2.146753	0.0318130

Table 5 - Regression output for Quasipoisson model with an interaction term

term	estimate	std.error	statistic	p.value
(Intercept)	1.9031426	0.8955945	2.125005	0.0348388
forest	0.0253773	0.0144278	1.758915	0.0801524
forest.clumpy	2.5026639	1.0437065	2.397862	0.0174294
forest:forest.clumpy	-0.0214570	0.0169860	-1.263215	0.2080126

As presented in table 4 there is a negative significant interaction effect between forest and forest.clumpy at the 5%-level in the Poisson model, but not in the Quasipoisson model in table 5. After including the interaction effect in this model it becomes worse as we also loose significance for the main effect of forest. The Quasipoisson model in table 3 without the interaction term is determined to be the best model of the ones being analyzed. A reason for the Poisson model with an interaction term not being chosen is because it still gives a residual deviance/df ratio of around 2.9.

Part 2

In the second part of this assignment we are going to study the effects of the variables forest and forest.clumpy on Empenigr using GLM. The variable we want to explain is now the presence or absence of Empetrum nigrum, which is a binary categorical variable. In this case, where the probability distribution function is binomial, the model choice falls on logistic regression with a logit link.

Table 6 - Regression model for multiple logistic model

term	estimate	std.error	statistic	p.value
(Intercept)	-2.3065734	3.8954959	-0.5921129	0.5537750
forest	0.0336824	0.0089513	3.7628513	0.0001680
forest.clumpy	-0.2140348	4.3065046	-0.0497003	0.9603612

The output from the logistic regression is presented in table 6. The variable forest is highly significant, but forest.clumpy is not significant at all with a high p-value of 0.96.

Just as for the poisson regression we should look at the ratio of residual variance over degrees of freedom.

$$\frac{\text{Residual variance}}{df} = \frac{241.67}{197} \approx 1.2$$

The ratio is not larger than 1.2 after rounding to the first decimal and there is no need to use the quasibinomial family. But since forest.clumpy was not significant another simpler model is also performed.

Table 7 - Regression output for simple logistic model

term	estimate	std.error	statistic	p.value
(Intercept)	-2.4985120	0.5179198	-4.824129	1.40e-06
forest	0.0338061	0.0086081	3.927268	8.59e-05

The simple logistic regression model in table 7 with forest as the only explanatory variable looks promising. The coefficient estimates are significant, and the ratio of residual deviance over degrees of freedom is not above 1.2. The Akaike Information Criterion (AIC) of 245.68 is also lower than the model in table 6 (247.67) which is a sign of a better goodness of fit. It is useful for comparing models estimated on the same dataset and with the same response variable.

The coefficient estimate for forest is 0.033806, meaning that the log-odds of the presence of Empetrum nigrum increases by 0.033806 when the proportion of forest in a 2.5 x 2.5 square km increases by 1 unit.

Table 8 - Regression output for multiple logistic model with an interaction term

term	estimate	std.error	statistic	p.value
(Intercept)	-30.7164745	13.4064748	-2.291167	0.0219537
forest	0.5083564	0.2100929	2.419674	0.0155344
forest.clumpy	32.8914027	15.4237860	2.132512	0.0329648
forest:forest.clumpy	-0.5557821	0.2441474	-2.276420	0.0228209

As it turns out there seem to be an interaction effect between the explanatory variables forest and forest.clumpy (table 8). The AIC goes from 245.68 in the model with forest as the only explanatory variable to 243.9 for the model with the interaction term which indicates a better goodness of fit. All estimates are significant at the 5%-level and the ratio of residual deviance over degrees of freedom is 1.2, which is sufficient. The model has positive estimates for the main effects of the variables, but the interaction effect is negative, meaning that the effect of forest on the log-odds of Empenigr is different at different values of forest.clumpy (The analysis factor, 2024) and the positive effect of forest on the log-odds of Empenigr becomes weakened when forest.clumpy increases.

Discussion

This study has aimed at finding out the relationships between the proportion of forest and the aggregation of the forest structure on the richness of forest species in Södermanland, Sweden. Furthermore, it has investigated what those effects are if being present.

Additionally, the effects of those forest variables have been analyzed on the occurrence of the species Empetrum nigrum.

Finally, the interactive effect of the explanatory forest variables have been modelled.

Model choices landed on poisson regression for the effects on species richness and logistic regression for estimating the effects on the presence of Empetrum nigrum.

Tha analysis comes to the conclusion that both of the explanatory variables should be included in a Quasipoisson model because of the overdispersion, with the variance increasing with the mean, in the standard Poisson model. Even though the model with the interaction effect gave significant estimates; it was not chosen because of the overdispersion. As indicated by scatterplots, the model showed positive main effects on amount of species from both the forest proportion and forest structure.

Regarding the effect on the presence of Empetrum nigrum, a logistic model with the interaction of the explanatory forest variables was chosen because it had significant estimates for all coefficients at the 5%-level and it had the best goodness of fit as demonstrated by the Akaike Information Criteria (AIC). The effects on the log-odds Empetrum nigrum occurrence from the forest proportion and forest structure was also positive, but the negative interaction effect weakened the positive effect from forest with increases in the aggregation index of the forest structure.

The study could have been extended to test the accuracy of the models with train-test splitting (also called cross-validation), where a part of the data is used for fitting the model and another part used for testing it (Thulin, 2021). This would make it possible to evaluate how well the model works on new data and would return the accuracy of how well the model can predict the species richness or the occurrence of the Empetrum nigrum. However, accuracy testing was not the goal of this study.

References

Thulin, M. (2021). Modern Statistics with R. Eos Chasma Press. ISBN 9789152701515.

Studiewebben (2024). Part 2. Generalized linear models (GLM):
Exercise 1. Statistical Data Analysis and Visualisation in R:II, GLM.
Södertörn University.

The analysis factor (2024). Interpreting interactions in regression.
https://www.theanalysisfactor.com/interpreting-interactions-in-regression/
#:~:text=The%20presence%20of%20an%20interaction%20indicates%20that%20the,
the%20two%20predictor%20variables%20are%20multiplied%20tests%20this.
[2024-02-17]

Appendix

In this appendix the R code from the study is presented.

```
### Libraries
knitr::opts_chunk$set(echo = TRUE)
library(faraway)
library(tidyverse)
library(esquisse)
library(ggplot2)
library(effects)
library(flextable)
library(xtable)
library(tables)
library(broom)
library(knitr)
library(kableExtra)

### Importing data
forest <- read.table("C:/Users/mart0/OneDrive/Dokument/Martin/Södertörns högskola/
Statistisk analys och visualisering i R - II/Part 2 - GLM and GAM/Assignment 1/forest.csv",
header = TRUE, sep = ",",dec=".")

### Checking data
glimpse(forest)
summary(forest)

### Figure 1
### esquisser()

ggplot(forest) +
 aes(x = forest, y = sp.richness) +
 geom_point(shape = "circle", size = 2L, colour = "#112446") +
 labs(x = "Proportion of forest", y = "Richness of forest species",
 title = "Scatterplot between forest and sp.richness") +
 theme_minimal() +
 theme(plot.title = element_text(size = 18L), axis.title.y = element_text(size = 14L),
 axis.title.x = element_text(size = 14L))

### Figure 2
### esquisser()

ggplot(forest) +
 aes(x = forest.clumpy, y = sp.richness) +
 geom_point(shape = "circle", size = 2L, colour = "#112446") +
 labs(x = "Aggregation index of forest structure", y = "Richness of forest species",
 title = "Scatterplot between forest.clumpy and sp.richness") +
 theme_minimal() +
 theme(plot.title = element_text(size = 18L), axis.title.y = element_text(size = 14L),
 axis.title.x = element_text(size = 14L))

 ### Table 1
 ols_model_1 <- lm(sp.richness ~ forest + forest.clumpy, data = forest)
summary(ols_model_1)

ols_model_1_table <- tidy(ols_model_1)
kable(ols_model_1_table)

### Figure 3
par(mfrow=c(2,2))
plot(ols_model_1)
```

```
### Table 2
glm_model_1 <- glm(sp.richness ~ forest + forest.clumpy, family=poisson(link=log), data=forest)
summary(glm_model_1)

glm_model_1_table <- tidy(glm_model_1)
kable(glm_model_1_table)

### Residual deviance / df calculation for model in table 2
579.88/197
\[
\frac{\text{Residual variance}}{df} = \frac{579.88}{197} \approx 2.9
\]

### Table 3
glm_model_2 <- glm(sp.richness ~ forest + forest.clumpy, family=quasipoisson(link=log), data=forest)
summary(glm_model_2)

glm_model_2_table <- tidy(glm_model_2)
kable(glm_model_2_table)

### Figure 4
plot(allEffects(glm_model_1))

### Figure 5
plot(allEffects(glm_model_2))

### Figure 6
plot(allEffects(glm_model_1), rescale.axis=F)

### Figure 7
plot(allEffects(glm_model_2), rescale.axis=F)

### Table 4
glm_model_3 <- glm(sp.richness ~ forest + forest.clumpy + forest:forest.clumpy,
family=poisson(link=log), data = forest)
summary(glm_model_3)

glm_model_3_table <- tidy(glm_model_3)
kable(glm_model_3_table)

### Table 5
glm_model_4 <- glm(sp.richness ~ forest + forest.clumpy + forest:forest.clumpy,
family=quasipoisson(link=log), data = forest)
summary(glm_model_4)

glm_model_4_table <- tidy(glm_model_4)
kable(glm_model_4_table)

### Table 6
logistic_model_1 <- glm(Empenigr ~ forest + forest.clumpy, family = binomial, data = forest)
summary(logistic_model_1)

logistic_model_1_table <- tidy(logistic_model_1)
kable(logistic_model_1_table)
```

```
### Residual deviance / df calculation for model in table 6
241.67/197
\[
\frac{\text{Residual variance}}{df} = \frac{241.67}{197} \approx 1.2
\]

### Table 7
logistic_model_2 <- glm(Empenigr ~ forest, family = binomial, data = forest)
summary(logistic_model_2)

logistic_model_2_table <- tidy(logistic_model_2)
kable(logistic_model_2_table)

### Table 8
logistic_model_3 <- glm(Empenigr ~ forest + forest.clumpy + forest:forest.clumpy,
family = binomial, data = forest)
summary(logistic_model_3)

logistic_model_3_table <- tidy(logistic_model_3)
kable(logistic_model_3_table)
```