Stockholms universitet

Modelling daily numbers of ringed birds with negative binomial generalized linear models

Måns Karlsson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2014:10
Matematisk statistik
September 2014

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2014:10**
http://www.math.su.se

Modelling daily numbers of ringed birds with negative binomial generalized linear models

Måns Karlsson*

September 2014

Abstract

In this thesis, we extend the mathematical framework of the generalized linear model to encompass the negative binomial distribution. Models based on the negative binomial distribution is needed when data doesn't fit a Poisson distribution due to overdispersion. We put theory into practice by analysing the inuence of weather and time on the daily number of ringed Eurasian Robins (Erithacus rubecula) in Falsterbo, Sweden. We find that increasing wind speed lowers the expected number of Robins, while drops in mean day temperature, increasing share of side wind, increasing yearly total of Robins and proximity to the median migration date all increases the expected number of Robins. These results are partly in accordance with previous studies and the results deviating from previous knowledge especially are discussed. Further improvements of the fitted model are also discussed.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail:mok.fbo@gmail.com. Supervisor: Martin Sköld.

1 Abstract

In this thesis, we extend the mathematical framework of the generalized linear model to encompass the negative binomial distribution. Models based on the negative binomial distribution is needed when data doesn't fit a Poisson distribution due to overdispersion. We put theory into practice by analysing the influence of weather and time on the daily number of ringed Eurasian Robins (*Erithacus rubecula*) in Falsterbo, Sweden. We find that increasing wind speed lowers the expected number of Robins, while drops in mean day temperature, increasing share of side wind, increasing yearly total of Robins and proximity to the median migration date all increases the expected number of Robins. These results are partly in accordance with previous studies and the results deviating from previous knowledge especially are discussed. Further improvements of the fitted model are also discussed.

Contents

2 Introduction

There is quite an abundance of unanalysed bird population monitoring data in Sweden in general and at Falsterbo Bird Observatory in particular. However, such data is often high variance count data, far from ordinary multiple regression on continuous data. Thus, the number of analyses performed is usually held back by a shortage of people with appropriate skills among researchers and amateurs collecting and owning this data.
In this thesis, we will take advantage of the surplus of data and look at the daily number of ringed Eurasian Robins in Falsterbo, trying to predict it through a function of local weather and time.
The daily number of Robins is typical bird population monitoring data in the sense that it fits an overdispersed Poisson distribution. In order to account for the higher variance, we will extend the framework of generalized linear models (GLMs) to encompass the negative binomial distribution, which is of great use when dealing with data fitting an overdispersed Poisson distribution.

The first chapter presents the data from Falsterbo Bird Observatory and the Swedish Institute for Meteorology and Hydrology. Some transformations of these data are also presented and motivated. We also discuss the collinearity issues in the weather data, including how one can deal with it.

In the second chapter of this thesis we present the theory of GLMs and expand it to encompass the exponential dispersion family, which the negative binomial distribution belongs to.

Thereafter comes, in chapter three, a presentation of the fitted models. We compare the models through measures of fit and their plausibility, with regard to previous research and possible causal relationships.

In the fourth chapter, the conclusions of the comparison of models is presented, along with a discussion of further improvements of the model(s).

Hopefully, the methods of analysis presented in this thesis can somewhat aid ecologists and ornithologists studying similar relationships in other datasets. The R-script I have developed alongside the analysis is almost entirely generalized and thus applicable on other species and weather data sets. It is available by contacting me through email.

A basic understanding of calculus, algebra, probability theory and mathematical statistics is expected from the reader.

3 Material

3.1 Background

At Falsterbo Bird Observatory, data collection for bird population monitoring has been the primary task since 1980. Species are monitored through various methods; breeding bird surveys, roosting bird surveys, counting diurnal migrating birds and ringing migrating passerines. The common goal for these time series is to detect long-term trends, primarily in population size and migration timing.

However, it can and should be questioned whether data actually is possible to use for such purposes. In order to verify the value of this data it is important to know what influences it. We will try to do this through modelling the daily number of ringed Eurasian Robins (*Erithacus rubecula*), henceforth called Robins, as a function of weather and time factors.

3.1.1 The standardized ringing

Nocturnally migrating passerines, such as the Robin, are at Falsterbo mainly monitored through standardized ringing. Birds are trapped using mistnets in fixed locations during a fixed time of the day, during a fixed interval of dates. Only weather influences the number of mistnets used each day; nets are only active when wind speed is sufficiently low and there is no heavier precipitation. This is in order to assure the safety of the birds, as they may suffer from getting caught during such circumstances. It may however also be a strong cause of variation in the collected data. The details of the standardized ringing scheme is presented in Roos and Karlsson (1981) and has, naturally, been unchanged since the first implementation.

3.1.2 The Eurasian Robin

Choosing the Eurasian Robin for this analysis has several statistical and scientific advantages. It is typical in the sense that it is the population that migrates past Falsterbo that gets monitored, which is the case for most species. There is plenty of previous research concerning the Robins migration, which could provide clues to what may influence it's presence at Falsterbo.

Also, being a nocturnal migrant, there is further research covering this migration strategy at large. The migration period during autumn is also fully covered by the ringing season for this species (Figure 3 and 4). It is a numerous species; averaging 2378 ringed birds per autumn provides a generous amount of days with ringed Robins, which hopefully mitigates some zero inflation.

The Robin is a distinct species morphologically, minimizing the risk of variation caused by species wise misidentification. It is also easy to determine the age of a Robin, separating first calendar year (1cy) birds from adult (2cy+) birds is

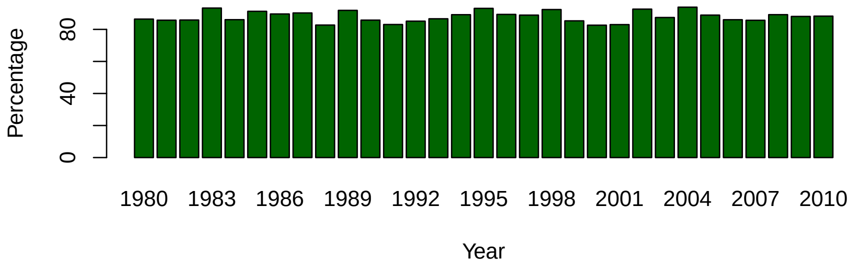


Figure 1: Percentage of first calendar year Robins

almost always possible and thus we can easily analyse age categories separately, if we have reason to believe there is a difference between them. The percentage of 1cy birds is almost constant over the years, which could indicate that there is no difference in effect of external factors between age categories.
However, it is very difficult to determine the sex of a Robin, especially during the autumn. Thus we cannot separate males and females and analyse these separately.

3.2 Response

The response data is, naturally, discrete. In it's original format it consist of the number of ringed Robins each day during the so called autumn season, i.e. the period of 21 july to the 10 november, for all years between and including 1980 and 2010. The distribution of daily totals is shown in Figure 2.

It is clear that data is heavily zero inflated and quite tail heavy to the right in it's original format. The heavy zero inflation is partly due to the distribution of Robins being heterogen over the ringing season, as is shown in Figure 3 and Figure 4.

3.2.1 Determining the migration period

If we want to model the number of birds for all dates of the autumn season, data follows the distribution shown in Figure 2. However, it would not make sense to try and find the influence of weather on the number of Robins during dates which has an expected number of Robins being close to zero.
Choosing a period of dates that can constitute the migration period of the Robin would thus be beneficial. It is also motivated by the abundance of structural

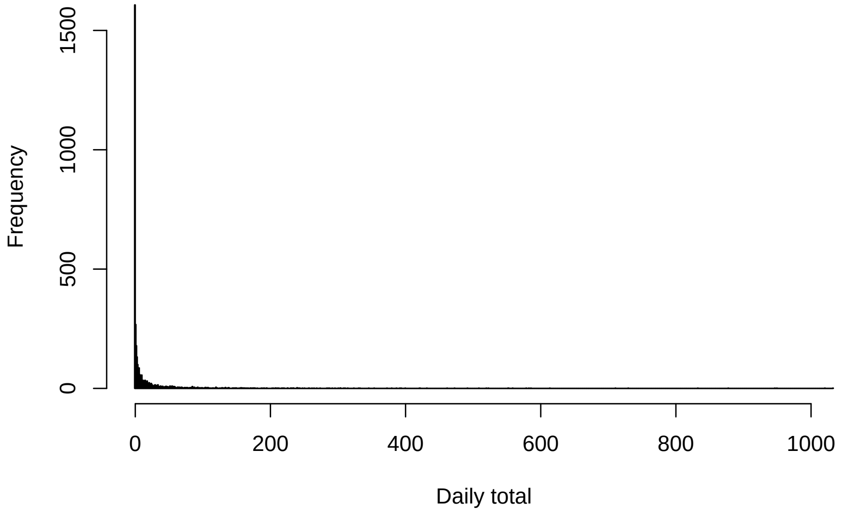


Figure 2: Distribution of daily totals of Robins

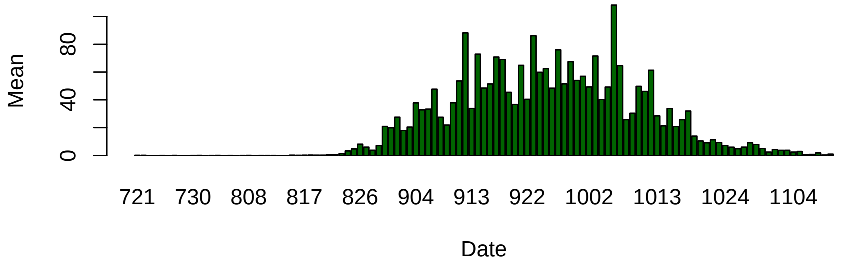


Figure 3: Mean number of Robins for each date of the autumn season, based on daily totals from the years 1980-2010.

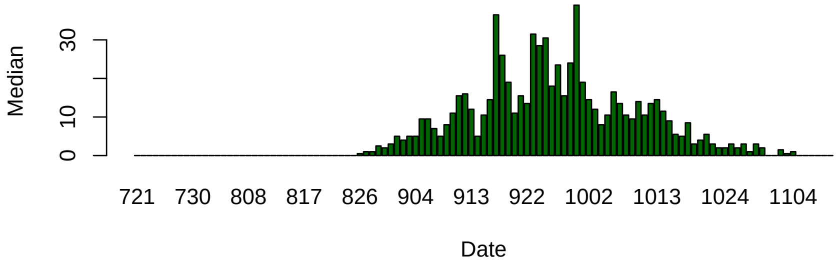


Figure 4: Median number of Robins for each date of the autumn season, based on daily totals from the years 1980-2010.

zeros in data, i.e. zeros that do not occur by chance. The days with zero Robins stemming from days outside the migration period are as good as inevitable. The zeros occurring on dates during the migration period should be considered sampling zeros, since they appear due to random effects, and they should not be excluded from our analysis.(Ridout, Demio, and Hinde, 1998)

Determining the migration period is somewhat arbitrary. One could take a period of dates with a mean number of birds exceeding c birds, but how would one determine the constant c?
A similar process could be revolving around the median number of birds. Again, choosing c is quite arbitrary.

The perhaps most rigorous way of determining the migration period would be to calculate between which dates a certain percentile of the migrating Robins are ringed. This would account for the variation in migration timing between years. However, what determines the percentile?
For each year, I calculated the dates on which 2.5 % and 97.5 % of the yearly sum of Robins had been ringed. I then chose, for each year separately, the daily totals of ringed Robins between and during these dates. The distribution of our selected daily totals is presented in Figure 5.

The number of zeros is now lower and the ones remaining should be sampling zeros.

3.2.2 Migrating and local populations

All Robins in our data has been categorized by age, they are either 1cy or 2cy+. By grading the extent of postjuvenile moult (i.e. to what extent the bird has lost it's juvenile plumage) among 1cy birds, locally bred birds may be separated

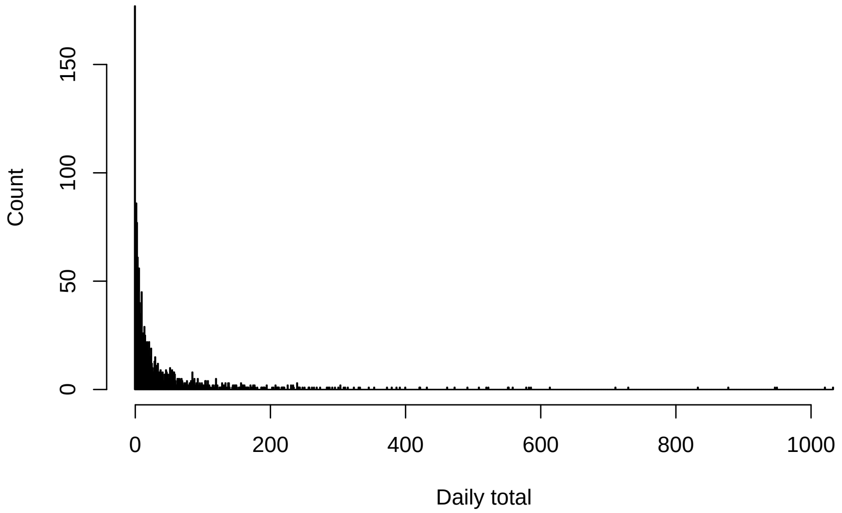


Figure 5: Distribution of selected daily totals

	Local birds				Migrating birds			
Score	0	1	2	3	4	5	6	NA
Count	9	10	3	10	459	14056	5090	48523
Percentage	0.01	0.01	<0.01	0.01	0.67	20.60	7.45	71.22

Table 1: Distribution of post juvenile moult among young Eurasian Robins ringed at Falsterbo during the autumn.

from birds having been bred elsewhere and arrived at Falsterbo during migration. Birds graded 4 or more may be considered migrating and the distribution of post juvenile moult is shown in Table 1.

An analogous criterion for separating migrating from local birds among adults is the extent of secondary (a section of the wing) moult. In Table 2 the distribution of secondary moult is shown. Birds with a score of 27 or more can definitely be regarded as migrating.

During days with more than 100 Robins, several measurements are excluded due to time shortage. Therefore, the number of NAs is high. These days only occur during the migration period. We know this since even the sum of all local birds registered in our data is too small to force measurement exclusion if all

	Local birds					Migrating birds				
Score	0	9	24	25	26	27	28	29	30	NA
Count	6	1	4	3	3	36	143	136	1979	7166
Percentage	0.06	0.01	0.04	0.03	0.03	0.38	1.51	1.44	20.88	75.61

Table 2: Distribution of secondary moult among adult Eurasian Robins ringed at Falsterbo during the fall.

were caught the same day.
The most important conclusion of Table 1.1 and Table 1.2 is that the absolute majority of Robins ringed in Falsterbo belong to the migrating population. This is the population of interest for our modelling.

3.3 Weather factors

Weather data has been downloaded from the open database at the Swedish Institution of Meteorology and Hydrology. (SMHI, 2014) During the whole period of standardised ringing, weather data has been collected at the very same location that the ringing has been conducted at.
There are several weather factors to choose from, and most of them are measured several times a day. According to Zehnder et al. (2001), wind speed, wind direction, air pressure and air pressure change accounted for 66 % of the variation in number of nocturnally migrating birds over Falsterbo. Temperature, temperature change, humidity, dew point temperature, distance to cloud base and amount of cloud cover had no significant influence on the number of birds observed. Precipitation was not included, due to it causing the equipment not to register the number of birds correctly.

Using this knowledge, I chose to use data from the measurements presented in Table 3 in some way. Air pressure data is not freely accessible, thus I could not include this variable. However, change in air pressure is very strongly reflected in wind speed; the quicker the change of air pressure the stronger the wind becomes.
These measures are taken 8 times daily, except for precipitation, which is taken once per day. I chose to include temperature, having personal observations of drops in temperature causing migratory movement among birds. Although I have not observed this among Robins I cannot rule out there being a possible correlation. Also, I find support for this in Alerstam et al. (1973), relating cold front passages to migration movement.

Constituting a very small part of data, the instances of NAs will be omitted in the analysis and thus we exclude the corresponding data for the number of ringed Robins.

Measurement	Observations	NAs	NA %
Temperature	12736	9	0.07
Precipitation	1592	0	0
Wind direction	12736	16	0.13
Wind speed	12736	16	0.13

Table 3: Missing values (NAs) in weather data.

In Appendix 1, complete correlation matrices for the daily measurements of temperature, side wind component, tail wind component and wind speed are presented, along with the eigen-values of the matrices. The smallest eigen-values are (rounded to two decimals) 0.02, 0.26, 0.08 and 0.08 respectively. Being so close to zero, they all indicate severe to moderate collinearity, except for sidewind component, which is acceptable. (Sundberg, 2002)

3.3.1 Temperature

The air temperature is measured in °C every third hour, starting at 00:00 UTC (Coordinated Universal Time). The strong indications of collinearity signals that we should be restrictive when selecting temperature variables for the initial model. I will use the temperature at each time, the mean temperature based on the 00:00 and 03:00 UTC and the change in this mean temperature between days as possible predictors.
Temperature variables will be denoted "TEXX", where "XX" is replaced by the UTC time of measurement, and the mean temperature will be denoted by "mTE" or "mTEDIFF" if the change in temperature is used.

3.3.2 Precipitation

The precipitation is measured in kg/m^2 during the 24h period between 06:00 UTC the day before the registered measurement and 06:00 UTC the day of measure. I do not have any data on duration of precipitation nor when during the day precipitation was received, making the measurement quite blunt.
The precipitation variable will be denoted "P".

3.3.3 Wind direction

The wind direction is defined as the direction from which the wind blows. Originally in data, wind direction was given in degrees, which is problematic for our models. Wind from 10° is very similar to wind from 360°, however our regression would interpret it as vastly separated. Also, no wind is noted with 0°.
Considering the mean track direction of migrating birds (225°) and that this

was generally independent of wind direction (Zehnder et al., 2001) I chose to split the wind direction in two components, side- and tailwind is denoted s_w and t_w respectively. They were calculated according to

$$s_w(d) = \left|\sin\left(\frac{(d-45)\pi}{180}\right)\right| \qquad t_w(d) = \cos\left(\frac{(d-45)\pi}{180}\right)$$

where d is the wind direction in degrees. Nothing indicates that sidewind from either right or left should be advantageous, thus the absolute value of this component should be used. With the rotation I use, perfect tailwind yields $t_w = 1$ and consequently, perfect headwind yields $t_w = -1$. This split also helps mitigate collinearity. Interactions with wind speed will be investigated.

It should be noted that Sandberg, Pettersson, and Alerstam (1988) experimentally found that Robins in Falsterbo during sunset and an hour thereafter oriented with a mean heading of 273° and 332° with clear skies and simulated covered skies respectively. If this is the actual heading of Robins during migration over Falsterbo, we should consider ignoring the absolute value of the side wind component, as it might actually be the true tail wind component. Also, we should take the absolute value of the current tail wind component in such an alternative analysis.

$$s_{alt}(d) = |t_w(d)| \qquad t_{alt}(d) = \sin\left(\frac{(d-45)\pi}{180}\right)$$

The correlation between the number of birds aloft close to dawn and the number of ringed birds the following morning (Zehnder and Karlsson, 2001) is probably not an argument strong enough to exclude all observations of wind direction besides 03:00 UTC and 06:00 UTC. However, when also considering the collinearity, the need for excluding some measurements is prominent. We will do this in various ways, such as taking means of several measurements.

Side wind component variables will be denoted "SWCXX", where "XX" is replaced by the UTC time of observation, and the mean side wind component will be denoted "mSWC". Tail wind component will be denoted analogously with "TWC".

3.3.4 Wind speed

The wind speed is measured in m/s. The magnitude of the collinearity problem calls for variable selection and/or combining. Several measurements concern wind speeds during a part of the day when neither migration nor ringing of Robins occur, such as the wind speeds in the afternoon. These should be possible to exclude, while the remaining measurements can be combined into one variable. All in all, the handling of these measurements will to great extent be similar to the handling of the wind direction variables, with which interactions should be investigated.

Wind speed variables will be denoted "WSXX", where "XX" is replaced by the UTC time of observation, and mean wind speed will be denoted "mWS".

3.4 Time factors

Clearly, the number of Robins varies over time in several ways. There is variation between years and over dates within years, which should reflect fluctuations in migration population and migration timing respectively. Ideally this variation is measured perfectly, since it is strongly associated with the purpose of the measurements; to detect long term trends in population size and migration timing.

3.4.1 Year

One of the purposes of the ringing in Falsterbo is to monitor long term population trends. For this it is assumed that the yearly totals is a sample of the migrating population of Robins, albeit somewhat distorted. By using the yearly total as an explanatory variable, we can investigate the relation between daily totals and yearly totals. If the yearly total mostly is an effect of (or lack of) singular days with extreme amounts of Robins, it should not explain the daily totals of Robins in an effective way. Although, if the seasonal total is an effect of the general amount of Robins, it should better explain the daily totals.
This variable will be denoted "Year".

3.5 Combined factors

In order to handle the collinearity problem for some data, I combined predictors as is shown in Table 4. I also calculated the difference in mean temperature between the day i and $i-1$, in order to be able to investigate the effect of temperature change.

3.5.1 Date

In order to investigate the influence of point in time during the annual migration cycle, I created an explanatory variable d_{ij} based on dates. I calculated the mean date $\bar{d_i}\cdot$ for a ringed Robin each year i and took the absolute value of the number of days of difference between $\bar{d_i}\cdot$ and all other dates d_{ij} for all i .

$$d_{ij} = |\bar{d_i}\cdot - d_j|$$

If the migration is uniformly distributed over my chosen date intervals, this factor should be insignificant. We will denote it "Date" in our models.

New predictor	Combined predictors
Mean wind speed	Wind speed 00:00 UTC Wind speed 03:00 UTC Wind speed 06:00 UTC
Mean side wind component	Side wind component 00:00 UTC Side wind component 03:00 UTC
Mean head wind component	Head wind component 00:00 UTC Head wind component 03:00 UTC
Mean temperature	Temperature 00:00 UTC Temperature 03:00 UTC

Table 4: Combination of predictor variables.

One could use date as an explanatory variable. I deem that suboptimal due to the variation in migration timing between years. We want to investigate the importance of where we are temporally in the migration cycle, not temporally on the year. Albeit there is correlation between the two, the latter is an attempt at constructing the former.

4 Methods

Unless stated otherwise, the information in this chapter is gathered from Agresti (2013), mainly chapters 4 and 14, with some slight own additions to clarify results and relations.

4.1 Regression for count data

When performing statistical inference on categorical data, logistic regression and loglinear models are common choices of method. These models are both generalized linear models and in the special case of one integer valued response variable, the models are even equivalent.
Count data, such as the number of birds present at a certain location, fits the special situation. A loglinear model with explanatory variables $\mathbf{x}$ is thus

$$\log(\mu(\mathbf{x})) = \alpha + \beta_1 x_1 + ... + \beta_p x_p$$

with α as the intercept and β_i as the p model parameters. The intercept describes the logarithm of the expected number of birds being counted when all explanatory variables are at their zero-level. The remaining model parameters describe the influence of an explanatory variable on the outcome, through the change in log odds ratio for a successful outcome induced by the level of the

explanatory variable. This is indeed the model structure we will use, i.e. we will model the expected number of birds μ given $\mathbf{x}$.
Commonly, a Possion distribution of the response is assumed for these situations. However, when assuming Poisson distribution, due to the defintion of the Poisson distribution, one also assumes $\mu = E(Y) = \text{var}(Y)$. In our case, the number of ringed Robins during one day is the response. Estimation yields $\hat{\mu} = 46$ and $\hat{\text{var}}(\mathbf{Y}) = 9549$. Clearly, assuming data is Poisson distributed is not preferred.
When count data has the characteristic of $\mu < \text{var}(\mathbf{Y})$, it is said to be overdispersed.
Overdispersion is rather common, especially for data in the field of ecology, and there are ways of compensating for the higher variance. Our choice will be to instead assume a negative binomial distribution for our data and fit the model accordingly. The negative binomial distribution allows for unequal mean and variance, but it's parametrization is not entirely standardized. We let the probability mass function for the negative binomial distribution be defined as

$$f(y|\mu, \gamma) = \frac{\Gamma(y + 1/\gamma)}{\Gamma(1/\gamma)\Gamma(y+1)} \left(\frac{1/\gamma}{\mu + 1/\gamma}\right)^{1/\gamma} \left(1 - \frac{1/\gamma}{\mu + 1/\gamma}\right)^{y}$$

where Γ is the gamma function. Although it may seem like a unnecessarily complex definition, we get convenient expressions for the expected value and the variance:

$$E(Y) = \mu, \qquad \text{var}(Y) = \mu + \gamma\mu.$$

However, the switch from a distribution (such as the Poisson) in the natural exponential family to a distribution in the exponential dispersion family has some consequences for the generalized linear model framework. The dispersion parameter γ has to be assumed fixed for a negative binomial regression model to be a GLM. Also, the likelihood equations for maximum likelihood estimation of the model parameters is a special case of those for an ordinary GLM. The remainder of this chapter is mainly concerned with deriving the general likelihood equations for a GLM for which the data distribution is assumed to be in the exponential dispersion family. Once the likelihood equations are determined, we can calculate maximum likelihood estimates of the model parameters.

4.2 Generalized linear models

The ordinary regression models can be extended to also allow for data with certain other distributions within the exponential family than the normal distribution. Such models are classified as generalized linear models, abbreviated GLMs, and consists of three components:

- The random component states the probability distribution of the response variable Y.
- The systematic component specifies explanatory variables $\boldsymbol{\beta}$ used in a linear predictor function.
- The link function connects the random and systematic components by specifying the function of $E(Y)$ that the model equates to the linear predictor.

4.3 The random component

The random component is the observed, independent values of the response variable Y, i.e. the response vector $\mathbf{y} = (y_1, ..., y_N)^T$, from a distribution in the natural exponential family. The probability mass function for a categorical data point i takes the general form of

$$f(y_i|\theta_i) = a(\theta_i)b(y_i)\text{exp}\left(T(y_i)Q(\theta_i)\right). \tag{1}$$

In (1), $a(\theta_i)$ is the normalizing factor, $b(y_i)$ is a function of the observation y_i, $T(y_i)$ is a sufficient statistic of y_i and $Q(\theta_i)$ is the natural parameter. (Liero and Zwanzig, 2012)
When dealing with distributions in the exponential dispersion family, such as the negative binomial, we have to extend the random component to

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right). \tag{2}$$

Here, θ_i denotes the natural parameter. One can identify the following parts of (2), as elements of the probability mass function for a natural exponential family:

- $\theta_i/a(\phi)$ corresponds to $Q(\theta_i)$.
- $\exp\left(-b(\theta_i)/a(\phi)\right)$ corresponds to $a(\theta_i)$.
- $\exp\left(c(y_i, \phi)\right)$ corresponds to $b(y_i)$

Also, we can let $T(y_i)$ be the observation y_i itself, since it is a sufficient statistic. When the dispersion parameter ϕ is known, (2) simplifies to (1).

4.4 The systematic component

The systematic component relates the parameters $\boldsymbol{\eta} = g(\boldsymbol{\mu})$ to the explanatory variables in the model matrix $\mathbf{X}$ and the model parameter vector $\boldsymbol{\beta}$ through the linear relation

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{3}$$

Each row i in the model matrix $\mathbf{X}$ contains the value of the explanatory variables x for observation i and each column contains the value of predictor j for every observation. The values of the regression parameters are contained in $\boldsymbol{\beta}$. Thus, for a model with N observations and k predictor variables, $\mathbf{X}$ is a $N \times k$ matrix. Having k predictor variables also renders $\boldsymbol{\beta}$ a k-dimensional column vector and N observations renders $\boldsymbol{\eta}$ an N-dimensional column vector.

4.5 The link function

The function g, introduced above, is the link function. It relates the parameters $\boldsymbol{\eta}$ to $\boldsymbol{\mu} = E(\mathbf{Y})$ through

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} \tag{4}$$

Since g is invertible, we can express

$$\boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta}). \tag{5}$$

In our data, $E(\hat{\mathbf{Y}})] = \mathbf{y}$ since we only have one measurement per day and thus $\hat{\boldsymbol{\eta}} = g(\mathbf{y})$.

4.6 Likelihood equations

The likelihood function L for the model parameter vector $\boldsymbol{\beta}$ is defined as

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_i P(Y_i = y_i). \tag{6}$$

Let $l(\boldsymbol{\beta}|\mathbf{y}) = \log(L(\boldsymbol{\beta}|\mathbf{y}))$ be the log likelihood, where log is the natural logarithm. Remebering (2), the form of the log likelihood for an exponential dispersion family is

$$l(\boldsymbol{\beta}|\mathbf{y}) = \sum_i \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_i c(y_i, \phi). \tag{7}$$

As per usual, the maximum likelihood estimates are the solutions to the likelihood equations, which are

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_i \frac{\partial l_i}{\partial \beta_j} = 0 \tag{8}$$

for all j. Differentiating (8) with the chain rule gives us

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{9}$$

If we can find expressions for each fraction in (9), we have explicit likelihood equations for the exponential dispersion family. Starting from the right, we find that $\partial\eta_i/\partial\beta_j$ is the derivative of (3) with respect to β_j and thus $\partial\eta_i/\partial\beta_j = x_{ij}$. Next is $\partial\mu_i/\partial\eta_i$, i.e. inversion of the derivative of the link function g. We leave this factor as it is since we are treating the general case and the link function is dependant on distribution and choice.
In order to find expressions for the remaining factors, we have to differentiate l.

Let

$$l_i = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \tag{10}$$

be the contribution of observation y_i to the log likelihood. Differentiation yields

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} \tag{11}$$

$$\frac{\partial^2 l_i}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{a(\phi)} \tag{12}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ is the first and second derivative of b evaluated at θ_i. We note that (11) replaces the first factor in (9). Since the regularity conditions hold in the exponential dispersion family we have that

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \implies \mu_i = E(Y_i) = b'(\theta_i) \tag{13}$$

$$-E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = E\left(\frac{\partial l}{\partial \theta}\right)^2 \implies \text{var}(Y_i) = b''(\theta_i)a(\phi) \tag{14}$$

Through (13) we find that $\partial\mu_i/\partial\theta_i = b''(\theta_i)$ and from (14) that $b''(\theta_i) = \text{var}(Y_i)/a(\phi)$. Since this is the inverse of the only factor we have not determined, we conclude that $\partial\theta_i/\partial\mu_i = \text{var}(Y_i)/a(\phi)$. We get the likelihood equations

$$0 = \sum_{i=1}^{N} \frac{y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{\mathrm{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{\mathrm{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \tag{15}$$

for $j = 0, 1, 2, ...$

4.7 The negative binomial GLM

The likelihood equations in (16) depend on Y_i only through the variance and the mean, which are connected through

$$\mathrm{var}(Y_i) = v(\mu_i). \tag{16}$$

The function v, called the variance function, characterizes the distribution of data. In our case we assume negative binomial distribution and due to our parametrization (as is stated in XX)

$$v(\mu) = \mu + \gamma \mu^2 \tag{17}$$

where γ is assumed to be constant.
For a negative binomial GLM with multiple predictor variables, the most common link function is the natural logarithm and that is what we will use in the next chapter. Fixing the link function to the natural logarithm, we can express the likelihood equations for a negative binomial GLM as

$$\frac{\partial^2 l}{\partial \beta_j \partial \gamma} = -\sum_i \frac{(y_i - \mu_i) x_{ij}}{(1 + \gamma \mu_i)^2 g'(\mu_i)} = 0 \tag{18}$$

for each j. The solutions to these equations are the maximum likelihood estimates of $\boldsymbol{\beta}$.

4.8 Tools for model and variable selection

We will use R to find, compare and analyse models. The R-package MASS (Venables and Ripley, 2002) has several useful tools, among which we will use glm.nb for our negative binomial regression and stepAIC for AIC-based variable elimination in fitted models.

4.8.1 The Akaike information criterion

Let k denote the number of parameters in a model and let L denote the value of the maximized likelihood function for the model. The Akaike information criterion is then defined as

$$AIC = -2\left(\log(L) - k\right)$$

One use of AIC is to compare the loss of information between models. Suppose we have a set of models $M_1, ..., M_N$ among which M_k is the model with the lowest AIC value AIC_{min}. The most probable model to lower the estimated information loss is then M_k. We can calculate how much less probable another model M_i is to lower the estimated information loss. Denote the probability that M_i lowers the estimated loss of information with $P(M_i \implies \text{Min. info. loss})$ and let AIC_i be the AIC value of M_i. Then

$$P(M_i \implies \text{Min. info. loss}) = exp\left(\frac{\mathrm{AIC}_{min} - \mathrm{AIC}_i}{2}\right). \tag{19}$$

Since we will be dealing with a surplus of possible predictor variables, the AIC is a suitable measure of fit for our models. The AIC penalizes a model with many predictor variables and thus we can single out more parsimonious model with it, giving us a general understanding. Using the procedure stepAIC we may perform variable elimination based on the AIC of the model, rather than the degree of explanation R^2, since it allows for comparison of nested models. The stepAIC-procedure will be based on backwards elimination of variables in our case, i.e. the algorithm tries to lower the AIC by removing insignificant explanatory variables.

4.8.2 Variation inflation factor

Let $\mathbf{C}$ denote the sample correlation matrix for the explanatory variables. The variation inflation factor (VIF) for a variable x_i is defined as $\mathbf{C}^{-1}_{ii}$. It indicates how much the variance of the corresponding regression coefficient is inflated by the presence of other, correlated variables. (Sundberg, 2002) It will be useful for finding collinearity problems in data.

As will the eigen-values e_i of C. Let $e_{\min}$ denote the smallest eigen-value of C. Strong collinearity is indicated by $e_{\min} < 0.05$ and moderate by $e_{\min} < 0.10$. (Sundberg, 2002)

5 Results

For our selected data, the mean number of birds $\hat{\mu} = 46.1$ and the variance $\hat{\text{Var}}(Y_i) = 9548.97$. Using (16) and (17) we can estimate $\hat{\gamma} = 4.47$. With such an overdispersion, the negative binomial GLM is preferred instead of the Poisson GLM.
Also, $\gamma = \exp(1/\theta)$, where θ is the parameter of the negative binomial distribution possible to estimate (without numerical methods). In the regression performed in R, θ is estimated according to above.

5.1 All predictors without interactions

In Appendix 2, the negative binomial GLM with all available predictor variables, without interactions, is presented. Due to the heavy collinearity in data, p-values are not really to be trusted. We should fit a model with less collinearity in data before investigating how reasonable it is that certain variables have a significant effect or not.
Looking at some measures of fit, we find that AIC = 13403. This can be compared to fitting a GLM with normal or Poisson distributions, which have AIC = 18721 and AIC= 95468 respectively. The lower value for the negative binomial model is an indication of our choice of distribution being appropriate.

Using stepAIC to eliminate variables lowers the AIC to 13380. The resulting model, denoted Model 1, is presented in Appendix 3. The correlation matrix for Model 1 has the smallest eigen-value (rounded to two decimals) 0.03, which indicates strong collinearity. Alas, we did not eliminate collinearity by ordinary variable exclusion and thus we still can not trust the parameter estimates as much as we would like. At least, using (19) we find that it is <0.001 times as probable for the model with all predictors to minimize the loss of information compared to Model 1.

Some diagnostics plots are also presented in Appendix 3. The residuals are loosely following a normal distribution and there is no observation with considerable leverage.

Looking at the fitted values in Figure 6 we find that Model 1 has the maximum fitted value of 390, compared to there being 25 observations exceeding 400 in data (Figure 5). It has, also compared to data in Figure 5, too low frequency of zeros, ones and twos. Otherwise, the distribution of the fitted values resembles the distribution of data rather much.

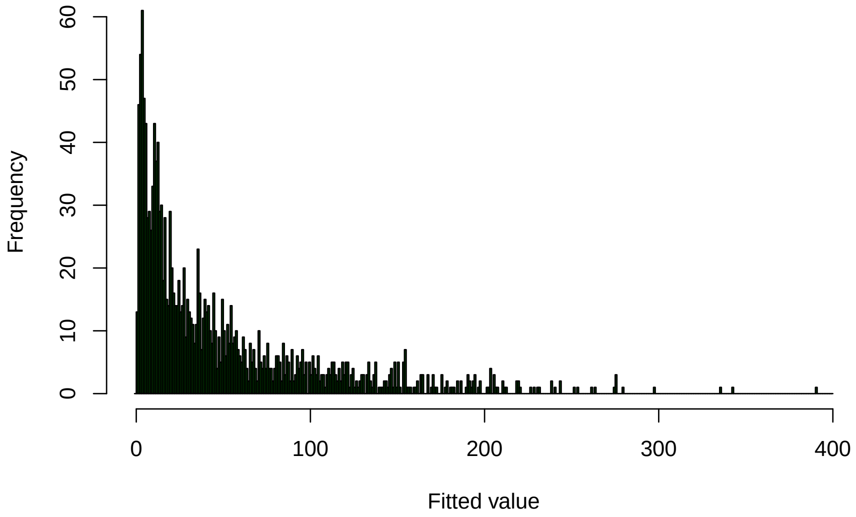


Figure 6: The fitted values of model 1

5.2 Combined predictors with interactions

5.2.1 Model specific results

In Appendix 4, the model with combined predictors is presented. The smallest eigen-value of the correlation matrix for this data is (rounded to two decimals) 0.46, indicating that we have eliminated the collinearity problem. Also, the AIC = 13377.49 is a bit lower than for Model 1.

Eliminating variables with stepAIC gives us the model in Appendix 5, denoted Model 2. The smallest eigen-value of the correlation matrix for the remaining data is (rounded to two decimals) 0.53, indicating no problematic collinearity. The only explanatory variable remaining with a p-value > 0.03 is Tail Wind component. Actually, we find that for Model 2 AIC = 13375.51 so there has been a further decrease in AIC. Comparing the AIC values with (19) gives us that it is 0.3717 times as probable for the model forming the basis for Model 2 to minimize the loss of information.

Some diagnostics plots are also presented in Appendix 4. Most important to note is that the residuals are somewhat following a normal distribution and that there is one observation (index 1477) that has quite some leverage. Looking into the details of the observation, it shows a relatively high number of Robins (380)

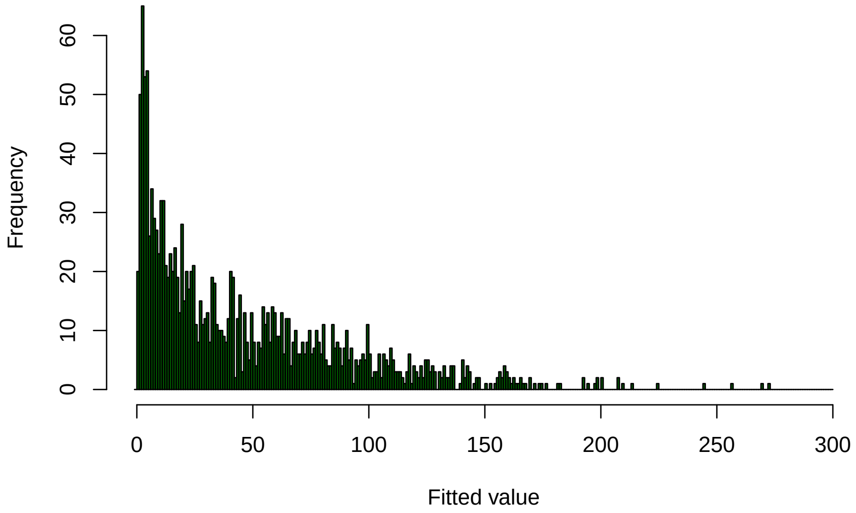


Figure 7: The fitted values of model 2

when only 20 previous had been ringed, thus making this the first selected date of the year 2009. Usually some earlier dates with fewer Robins are selected, that is the normal migration process, but this was an exception and presumably therefore this observation provides leverage.

Looking at the fitted values of Model 2 in Figure 7 we find that the maximum value is lower than for Model 1, at 273. The frequencies of zeros, ones and twos are still a bit low compared to data (Figure 5), but at least two is the most common value.

5.2.2 Variable specific results

Mean wind speed Estimated at -0.226 with a p-value close to zero, we find that the mean wind speed has a highly significant negative effect on the number of Robins. This is in accordance with Zehnder et al. (2001) where the wind speed has a significant negative effect on the number of birds aloft over Falsterbo in the autumn nights. The Pearson correlation between mean wind speed over the selected dates for each year and the yearly total of Robins is -0.262, although with a p-value of 0.153.

Mean side wind component Estimated at 0.918 with a p-value close to zero we find that the mean side wind component has a significant positive effect on the number of Robins. There were no particular expectations regarding this parameter. Possible reasons for the high significance is presented in discussion. The Pearson correlation between mean side wind component over the selected dates for each year and the yearly total of Robins is -0.146, with a p-value of 0.434.

Mean tail wind component Estimated at 0.142 with a p-value of 0.273, the mean tail wind component has an insignificant positive effect on the number of Robins. The tail wind component has a highly significantly positive effect on the number of birds in the airspace above Falsterbo. (Zehnder et al., 2001) This does not seem to apply for Robins. Although the effect is positive, it is rather insignificant. The Pearson correlation between mean tail wind component over the selected dates for each year and the yearly total of Robins is 0.472, with a p-value of 0.007.

Difference in mean temperature Estimated at -0.036 with a p-value of 0.026, the difference in mean temperature has a significant negative effect on the number of Robins. However, the variable is constructed with a drop in temperature represented by negative values. Thus, a drop in temperature increases the number of Robins. The result is in accordance with Alerstam et al. (1973); drops in temperature spur migratory movement.

Date Estimated at -0.026 with a p-value close to zero, the temporal distance from median migratory date has a significant negative effect on the number of Robins. In other words, the number of Robins is dependent on how temporally close we are to peak migration. Figure 4 indicates that the migration of Robins is not uniformly distributed over time, which is in accordance with this result.

Year Estimated at < 0.001 with a p-value close to zero, the yearly total of Robins has a significant positive effect on the daily number of Robins.

Side wind - wind speed interaction The interaction between side wind component and wind speed has, with a p-value of 0.026, a significantly negative effect on the number of Robins and is estimated to -0.083. Although the side wind component has a significant positive effect, the wind speed is presumably superior in the sense that even if side wind component is optimal, too high wind speeds won't allow for migration and/or ringing. (Zehnder et al., 2001) (Roos and Karlsson, 1981)

Tail wind - wind speed interaction The interaction between side wind component and wind speed has, with a p-value close to zero, a significantly positive effect on the number of Robins and is estimated to 0.111. It seems tail wind is even more favorable if the wind speed is high.

Side wind - tail wind interaction The interaction between side wind component and head wind component has, with a p-value of 0.030, a significantly negative effect on the number of Robins and is estimated to 0.111. Considering that the tail wind component is negative during headwind, this indicates a positive effect of head wind. However, it is hard to interpret this variable.

6 Discussion

6.1 Zero inflation

The number of predicted zeroes and ones should ideally be higher for Model 2, in order to fit data even better. Achieving a higher frequency of zeros can be done by fitting a zero inflated negative binomial generalized linear model (ZINB). Zero inflated models are of the mixture model class and combines a count mass and a point mass at zero. (Phang and Loh, 2013) Such model states that $P(Y_i = 0) = q_i$ and $P(Y_i \text{ has some distribution without zero inflation}) = 1 - q_i$, i.e. a binomial trial is conducted and either we get the response zero or a response from a probability distribution. (Lord, Washington, and Ivan, 2004)
A natural way to implement zero inflation would be to relate it to the number of days that no attempt at ringing Robins was performed. The assumption behind cancelling the ringing is that the weather is so extreme that very few or no birds are possible to catch (Roos and Karlsson, 1981). Viewing these days as results of binomial trials with a year specific probability of success would seem appropriate.
It is important to note that we wish to fit a model representing the real chain of events, not only find the best possible fit according to AIC (or something equivalent). Since attempts at ringing Robins and other birds sometimes are made in the same weather as ringing is cancelled in, we would represent the randomness in the conducted ringing fairly well with a ZINB.

6.2 Wind direction

From the results in Zehnder et al. (2001) I expected the tail wind component to be significantly positive for the number of Robins, especially when interacting with wind speed. The results in Sandberg, Pettersson, and Alerstam (1988) concerned Robins taking off in Falsterbo, their mean heading could be influenced by local factors such as city lights. The mean direction of movement for Robins ringed in Falsterbo during autumn and recovered elsewhere later in the autumn is southwestern (Karlsson, 2014). The direction coincides with the mean migratory direction in Zehnder et al. (2001). However, the Robins need not have a southwestern heading regionally over Falsterbo. In conclusion, there are uncertainties as to how the (presumed) tail wind component affects the number of Robins.

The significant positive effect of the side wind component could be due to Robins normally not migrating over Falsterbo drifting with the wind and ending up in Falsterbo. The scenario would be similar to the results in Gezelius and Hedenstrom (1988) and could be further investigated similarly. It could also form basis for an analysis of whether direction of the side wind component matters.

6.3 The Robin population

Investigating the relation between the number of Robins in Falsterbo and breeding bird surveys such as the Swedish Bird Survey could provide insight into how the number of Robins in Falsterbo is influenced by the population at large. However, this would demand that we knew the general geographical origin of the Robins ringed in Falsterbo, in order to know what population is supposedly measured. This would then form the basis for filtering data from the Swedish Bird Survey.

Currently, adequate knowledge of where the ringed Robins originate geographically is unavailable, although it is most likely derivable from data of recovered ringed Robins at Falsterbo Bird Observatory. Bird surveys is considered the most efficient way of measuring populations (Svensson, 1978), thus it would be interesting to replace the "Year"-factor in our model with estimated population sizes from these surveys in order to see if Falsterbo data reflect population changes at large.

7 Acknowledgements

My sincerest gratitude is dedicated to Martin Sköld, my supervisor during this thesis. After hinting at what methods I should use, has he given me almost complete freedom and trust in figuring out what theory to study as well as in handling the modelling. I believe this is the best way for me to learn mathematics and I could not have wished for a better scenario.

Also, I am very thankful that Falsterbo Bird Observatory let me use their data for free.

References

Agresti, Alan (2013). *Categorical Data Analysis*. Wiley.

Alerstam, Thomas et al. (1973). "Nocturnal passerine migration and cold front passages in autumn - a combined rada and field study". In: *Ornis Scandinavia* 4.2, pp. 103–111.

Gezelius, Lars and Anders Hedenstrom (1988). "Vindens inverkan pa fangsten av rodhake Erithacus rabecula och kungsfagel Regulus regulus vid Ottenby". In: *Var Fagelvarld* 47, pp. 9–14.

Karlsson, Lennart (2014). Falsterbo Bird Observatory. URL: http://www.falsterbofagelstation.se/arkiv/aterfynd/fynduttag2.php.

Liero, Hannelore and Silvelyn Zwanzig (2012). *Introduction to the theory of statistical inference*. CRC Press.

Lord, Dominique, Simon P. Washington, and John N. Ivan (2004). "Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory". In: *Accident Analysis & Prevention*.

Phang, Y. N. and E. F. Loh (2013). "Zero Inflated Models for Overdispersed Count Data". In: *International Journal of Mathematical, Computational, Physical and Quantum Engineering* 7.8, pp. 829–831.

Ridout, Martin, Clarice G.B. Demio, and John Hinde (1998). "Models for count data with many zeros". In: URL: http://sada2013.sciencesconf.org/conference/sada2013/pages/ridout98ZIPexcesszerosmodelsreview.pdf.

Roos, Gunnar and Lennart Karlsson (1981). "Ringmarkningsverksamheten vid Falsterbo Fagelstation 1980". In: *Anser* 20, pp. 99–108.

Sandberg, Roland, Jan Pettersson, and Thomas Alerstam (1988). "Why do migrating robins, Erithacus rubecula, captured at two nearby stop-over sites orient differently?" In: *Animal Behaviour*.

SMHI (2014). URL: http://opendata-download-metobs.smhi.se/explore/.

Sundberg, Rolf (2002). "Collinearity". In: *Encyclopedia of Environmetrics* 1, pp. 365–366.

Svensson, Soren (1978). "Efficiency of Two Methods for Monitoring Bird Population Levels: Breeding Bird Censuses Contra Counts of Migrating Birds". In: *Oikos* 30.2, pp. 373–386.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: http://www.stats.ox.ac.uk/pub/MASS4.

Zehnder, Susanna and Lennart Karlsson (2001). "Do ringing numbers reflect true migratory activity of nocturnal migrants?" In: *Journal fr Ornithologie* 142, pp. 173–183.

Zehnder, Susanna et al. (2001). "Nocturnal autumn bird migration at Falsterbo, South Sweden". In: *Journal of Avian Biology* 32.3, pp. 239–248.

8 Appendices

8.1 Appendix 1 - correlation matrices

Correlation matrices of the daily wind speed measurements.

	WS00	WS03	WS06	WS09	WS12	WS15	WS18	WS21
WS00	1	0.86	0.78	0.69	0.61	0.55	0.53	0.5
WS03	0.86	1	0.86	0.77	0.67	0.6	0.58	0.55
WS06	0.78	0.86	1	0.85	0.74	0.67	0.66	0.62
WS09	0.69	0.77	0.85	1	0.83	0.74	0.72	0.67
WS12	0.61	0.67	0.74	0.83	1	0.81	0.76	0.69
WS15	0.55	0.6	0.67	0.74	0.81	1	0.84	0.73
WS18	0.53	0.58	0.66	0.72	0.76	0.84	1	0.84
WS21	0.5	0.55	0.62	0.67	0.69	0.73	0.84	1

Eigen-values: $(5.70, 1.18, 0.44, 0.22, 0.15, 0.12, 0.10, 0.08)$

Correlation matrices of the daily side wind component measurements.

	SWC00	SWC03	SWC06	SWC09	SWC12	SWC15	SWC18	SWC21
SWC00	1	0.68	0.55	0.39	0.26	0.22	0.21	0.21
SWC03	0.68	1	0.67	0.47	0.31	0.24	0.26	0.23
SWC06	0.55	0.67	1	0.61	0.42	0.33	0.29	0.29
SWC09	0.39	0.47	0.61	1	0.61	0.47	0.37	0.34
SWC12	0.26	0.31	0.42	0.61	1	0.65	0.48	0.38
SWC15	0.22	0.24	0.33	0.47	0.65	1	0.59	0.47
SWC18	0.21	0.26	0.29	0.37	0.48	0.59	1	0.67
SWC21	0.21	0.23	0.29	0.34	0.38	0.47	0.67	1

Eigen-values: $(3.93, 1.52, 0.81, 0.50, 0.36, 0.31, 0.30, 0.26)$

Correlation matrices of the daily head wind component measurements.

	HWC00	HWC03	HWC06	HWC09	HWC12	HWC15	HWC18	HWC21
HWC00	1	0.89	0.81	0.73	0.63	0.56	0.56	0.52
HWC03	0.89	1	0.88	0.8	0.69	0.63	0.61	0.57
HWC06	0.81	0.88	1	0.88	0.77	0.69	0.68	0.64
HWC09	0.73	0.8	0.88	1	0.86	0.77	0.75	0.7
HWC12	0.63	0.69	0.77	0.86	1	0.85	0.8	0.73
HWC15	0.56	0.63	0.69	0.77	0.85	1	0.87	0.77
HWC18	0.56	0.61	0.68	0.75	0.8	0.87	1	0.88
HWC21	0.52	0.57	0.64	0.7	0.73	0.77	0.88	1

Eigen-values: $(6.15, 0.93, 0.32, 0.21, 0.14, 0.09, 0.09, 0.08)$

Correlation matrices of the daily temperature measurements.

	TE00	TE03	TE06	TE09	TE12	TE15	TE18	TE21
TE00	1.00	0.97	0.94	0.86	0.80	0.81	0.85	0.82
TE03	0.97	1.00	0.96	0.85	0.79	0.80	0.85	0.83
TE06	0.94	0.96	1.00	0.89	0.82	0.84	0.88	0.86
TE09	0.86	0.85	0.89	1.00	0.96	0.94	0.92	0.89
TE12	0.80	0.79	0.82	0.96	1.00	0.97	0.90	0.86
TE15	0.81	0.80	0.84	0.94	0.97	1.00	0.94	0.89
TE18	0.85	0.85	0.88	0.92	0.90	0.94	1.00	0.96
TE21	0.82	0.83	0.86	0.89	0.86	0.89	0.96	1.00

Eigen-values: $(7.16, 0.46, 0.19, 0.07, 0.05, 0.03, 0.02, 0.02)$

8.2 Appendix 2 - basis for Model 1

```
## 
## Call:
## glm.nb(formula = Robins ~ P + WS00 + WS03 + WS06 + WS09 + WS12 + 
##     WS15 + WS18 + WS21 + SWC00 + SWC03 + SWC06 + SWC09 + SWC12 + 
##     SWC15 + SWC18 + SWC21 + TWC00 + TWC03 + TWC06 + TWC09 + TWC12 + 
##     TWC15 + TWC18 + TWC21 + TE00 + TE03 + TE06 + TE09 + TE12 + 
##     TE15 + TE18 + TE21 + Year + Date, init.theta = 0.7233815393, 
##     link = log)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -2.760  -1.129  -0.532   0.109   4.950  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  4.132207   0.215791   19.15  < 2e-16 ***
## P            0.012416   0.008117    1.53   0.1261    
## WS00        -0.027000   0.019851   -1.36   0.1738    
## WS03        -0.175425   0.025039   -7.01  2.4e-12 ***
## WS06        -0.068691   0.024680   -2.78   0.0054 ** 
## WS09        -0.020742   0.023706   -0.87   0.3816    
## WS12        -0.014338   0.022165   -0.65   0.5177    
## WS15        -0.034752   0.020352   -1.71   0.0877 .  
## WS18         0.001036   0.021257    0.05   0.9611    
## WS21         0.052075   0.017517    2.97   0.0030 ** 
## SWC00        0.230313   0.140265    1.64   0.1006    
## SWC03        0.501288   0.159961    3.13   0.0017 ** 
## SWC06       -0.194433   0.159141   -1.22   0.2218    
## SWC09        0.180341   0.152073    1.19   0.2357    
## SWC12        0.051928   0.153529    0.34   0.7352    
## SWC15        0.076159   0.150906    0.50   0.6138    
## SWC18        0.166010   0.153700    1.08   0.2801    
## SWC21       -0.299266   0.142886   -2.09   0.0362 *  
## TWC00        0.238254   0.101413    2.35   0.0188 *  
## TWC03        0.319291   0.126484    2.52   0.0116 *  
## TWC06       -0.074888   0.122735   -0.61   0.5418    
## TWC09       -0.062725   0.127025   -0.49   0.6214    
## TWC12        0.161849   0.119456    1.35   0.1755    
## TWC15        0.032288   0.116931    0.28   0.7825    
## TWC18       -0.164637   0.126038   -1.31   0.1915    
## TWC21        0.127608   0.099586    1.28   0.2001    
## TE00         0.027739   0.044596    0.62   0.5339    
## TE03         0.031141   0.058974    0.53   0.5975    
## TE06        -0.095319   0.051134   -1.86   0.0623 .
## TE09        -0.028405   0.047837   -0.59   0.5527
## TE12         0.054493   0.045544    1.20   0.2315
## TE15         0.017923   0.049136    0.36   0.7153
## TE18         0.107624   0.056137    1.92   0.0552 .
## TE21        -0.115653   0.042857   -2.70   0.0070 **
## Year         0.000322   0.000037    8.68  < 2e-16 ***
## Date        -0.029517   0.004003   -7.37  1.7e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.7234) family taken to be 1)
##
##     Null deviance: 3232.0  on 1581  degrees of freedom
## Residual deviance: 1834.3  on 1546  degrees of freedom
## AIC: 13403
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.7234
##           Std. Err.:  0.0255
##
##  2 x log-likelihood:  -13328.9800
```

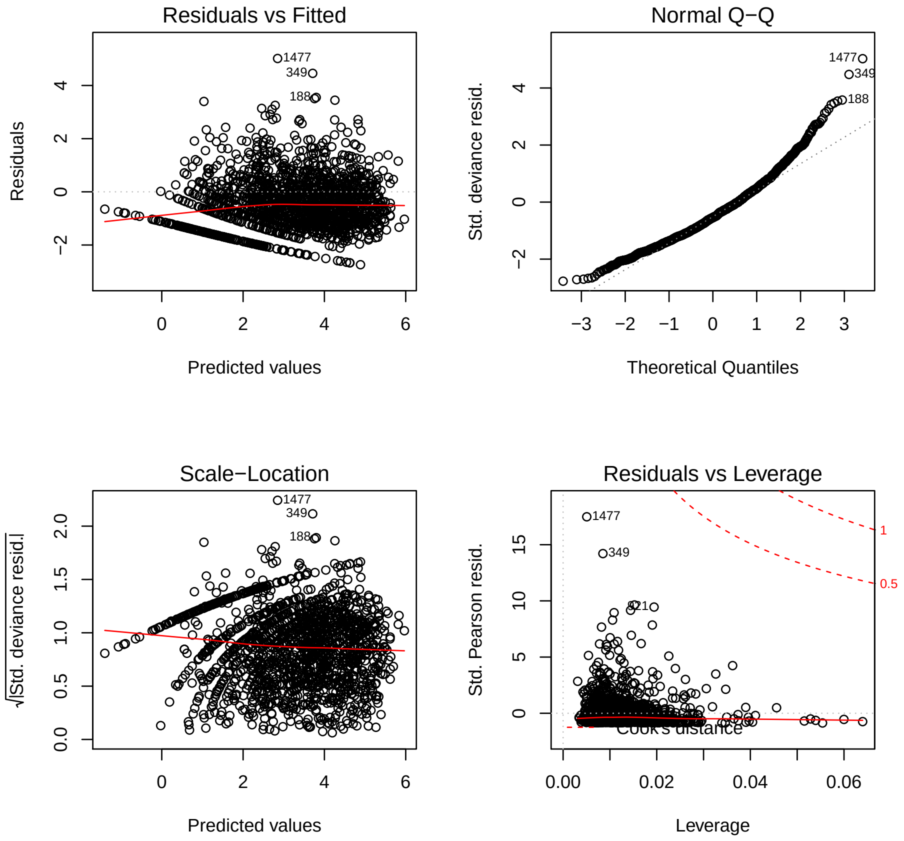
8.3 Appendix 3

8.3.1 Model 1

```
##
## Call:
## glm.nb(formula = Robins ~ P + WS03 + WS06 + WS15 + WS21 + SWC00 +
##     SWC03 + SWC18 + SWC21 + TWC00 + TWC03 + TWC12 + TE06 + TE12 +
##     TE18 + TE21 + Year + Date, init.theta = 0.7186111949, link = log)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.740  -1.131  -0.539   0.115   5.015
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.21e+00   2.09e-01   20.12  < 2e-16 ***
## P            1.29e-02   8.06e-03    1.60  0.10985
## WS03        -1.88e-01   1.85e-02  -10.20  < 2e-16 ***
## WS06        -9.99e-02   1.93e-02   -5.18  2.2e-07 ***
## WS15        -4.62e-02   1.50e-02   -3.09  0.00202 **
## WS21         4.79e-02   1.40e-02    3.41  0.00065 ***
## SWC00        2.53e-01   1.34e-01    1.89  0.05923 .
## SWC03        4.84e-01   1.38e-01    3.52  0.00043 ***
## SWC18        2.01e-01   1.37e-01    1.48  0.13998
## SWC21       -2.52e-01   1.38e-01   -1.82  0.06810 .
## TWC00        2.23e-01   9.61e-02    2.32  0.02009 *
## TWC03        2.46e-01   1.05e-01    2.34  0.01922 *
## TWC12        1.11e-01   6.95e-02    1.60  0.11017
## TE06        -5.07e-02   2.56e-02   -1.98  0.04739 *
## TE12         5.71e-02   2.49e-02    2.30  0.02160 *
## TE18         1.24e-01   4.71e-02    2.63  0.00849 **
## TE21        -1.34e-01   3.89e-02   -3.44  0.00059 ***
## Year         3.11e-04   3.69e-05    8.42  < 2e-16 ***
## Date        -2.97e-02   3.98e-03   -7.47  8.3e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.7186) family taken to be 1)
##
##     Null deviance: 3212.9  on 1581  degrees of freedom
## Residual deviance: 1834.8  on 1563  degrees of freedom
## AIC: 13380
##
## Number of Fisher Scoring iterations: 1
```

```
##
##
##               Theta:  0.7186
##           Std. Err.:  0.0253
##
##  2 x log-likelihood:  -13339.8790
```

8.3.2 Diagnostics plots for Model 1



8.4 Appendix 4 - Basis for Model 2

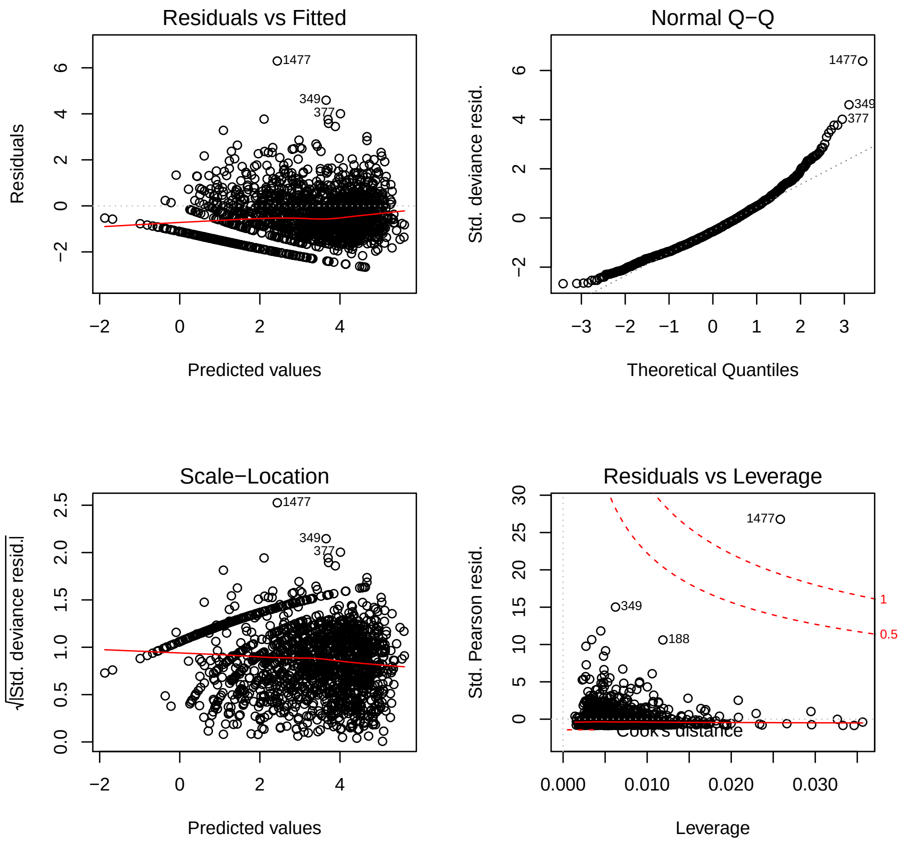
```
##
## Call:
## glm.nb(formula = Robins ~ P + mWS * mSWC * mTWC + mTEDIFF + Date +
##     Year, init.theta = 0.7127791292, link = log)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.698  -1.119  -0.538   0.132   6.293
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.31e+00   1.82e-01   23.68  < 2e-16 ***
## P              9.17e-03   8.04e-03    1.14    0.254
## mWS           -2.22e-01   2.84e-02   -7.80  6.2e-15 ***
## mSWC           9.49e-01   2.41e-01    3.93  8.5e-05 ***
## mTWC          -4.68e-02   2.16e-01   -0.22    0.828
## mTEDIFF       -3.50e-02   1.60e-02   -2.18    0.029 *
## Date          -2.60e-02   3.97e-03   -6.55  5.6e-11 ***
## Year           2.76e-04   3.67e-05    7.53  5.1e-14 ***
## mWS:mSWC      -9.46e-02   4.01e-02   -2.36    0.018 *
## mWS:mTWC       1.44e-01   3.68e-02    3.91  9.3e-05 ***
## mSWC:mTWC     -3.97e-02   4.43e-01   -0.09    0.929
## mWS:mSWC:mTWC -7.02e-02   6.74e-02   -1.04    0.298
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.7128) family taken to be 1)
##
##     Null deviance: 3189.5  on 1581  degrees of freedom
## Residual deviance: 1833.7  on 1570  degrees of freedom
## AIC: 13377
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.7128
##           Std. Err.:  0.0251
##
##  2 x log-likelihood:  -13351.4880
```

8.5 Appendix 5

8.5.1 Model 2

```
##
## Call:
## glm.nb(formula = Robins ~ mWS + mSWC + mTWC + mTEDIFF + Date +
##     Year + mWS:mSWC + mWS:mTWC + mSWC:mTWC, init.theta = 0.7120380614,
##     link = log)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.663  -1.116  -0.555   0.134   6.297
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.31e+00   1.81e-01   23.82  < 2e-16 ***
## mWS         -2.26e-01   2.68e-02   -8.45  < 2e-16 ***
## mSWC         9.18e-01   2.39e-01    3.85  0.00012 ***
## mTWC         1.42e-01   1.29e-01    1.10  0.27279
## mTEDIFF     -3.56e-02   1.60e-02   -2.22  0.02614 *
## Date        -2.55e-02   3.96e-03   -6.44  1.2e-10 ***
## Year         2.78e-04   3.66e-05    7.59  3.2e-14 ***
## mWS:mSWC    -8.34e-02   3.75e-02   -2.22  0.02632 *
## mWS:mTWC     1.11e-01   1.77e-02    6.27  3.6e-10 ***
## mSWC:mTWC   -4.61e-01   2.12e-01   -2.17  0.02986 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.712) family taken to be 1)
##
##     Null deviance: 3186.5  on 1581  degrees of freedom
## Residual deviance: 1834.0  on 1572  degrees of freedom
## AIC: 13376
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.7120
##           Std. Err.:  0.0250
##
##  2 x log-likelihood:  -13353.5080
```

8.5.2 Diagnostics plots for Model 2


Residuals vs Fitted
Residuals
Predicted values
−2
0
2
4
6
1477
349
377
Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
−3
−2
−1
0
1
2
3
1477
349
377
Scale−Location
√|Std. deviance resid.|
Predicted values
0.0
0.5
1.0
1.5
2.0
2.5
1477
349
377
Residuals vs Leverage
Std. Pearson resid.
Leverage
0.000
0.010
0.020
0.030
0
5
10
15
20
25
30
1477
349
188
1
0.5
Cook's distance