

Generalised Linear Models 1:

Intro to GLMs and Modelling Count Data

Michael Noonan

March 14, 2021

Biol 520C: Statistical modelling for biological data

1. Housekeeping
2. The Gaussian Assumption
3. Generalised Linear Models (GLMs)
4. Gaussian Linear Regression as a GLM
5. GLMs for count data
6. GLMs for count data in R

Housekeeping



- Papers have been returned.

- Papers have been returned. Quality was very high.

- Papers have been returned. Quality was very high. A few notes:

- Papers have been returned. Quality was very high. A few notes:

Always cite packages *and* package versions.

- Papers have been returned. Quality was very high. A few notes:

Always cite packages *and* package versions.

Units.

- Papers have been returned. Quality was very high. A few notes:

Always cite packages *and* package versions.

Units.

Try not to confuse function names with statistical tests (`anova()` on a pair of models is a likelihood ratio test).

- Papers have been returned. Quality was very high. A few notes:

Always cite packages *and* package versions.

Units.

Try not to confuse function names with statistical tests (`anova()` on a pair of models is a likelihood ratio test).

Try to find a balance between keeping the key results in the main text and leaning on the appendix for elements related to proof of validity of the statistical approach.

The Gaussian Assumption





We started this course with simple linear regression and we saw how:

We started this course with simple linear regression and we saw how:

- increasing the number of parameters can soak up variance and improve a model's explanatory power,

We started this course with simple linear regression and we saw how:

- increasing the number of parameters can soak up variance and improve a model's explanatory power,
- we can use mixed effects models to account for hierarchical data structures,

We started this course with simple linear regression and we saw how:

- increasing the number of parameters can soak up variance and improve a model's explanatory power,
- we can use mixed effects models to account for hierarchical data structures,
- we can modify the variance structure to account for heteroskedasticity,

We started this course with simple linear regression and we saw how:

- increasing the number of parameters can soak up variance and improve a model's explanatory power,
- we can use mixed effects models to account for hierarchical data structures,
- we can modify the variance structure to account for heteroskedasticity,
- we can modify the correlation matrix to correct for autocorrelation.

We started this course with simple linear regression and we saw how:

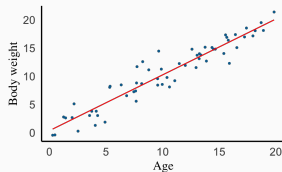
- increasing the number of parameters can soak up variance and improve a model's explanatory power,
- we can use mixed effects models to account for hierarchical data structures,
- we can modify the variance structure to account for heteroskedasticity,
- we can modify the correlation matrix to correct for autocorrelation.
- ... but there's an elephant in the room we've been ignoring...



Let's say we're interested in the relationship between age and weight in a species.

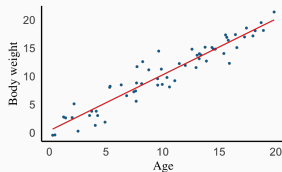
Let's say we're interested in the relationship between age and weight in a species. A simple linear relationship for this would contain an intercept (β_0) and a parameter linking weight and age (β_1):

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$



Let's say we're interested in the relationship between age and weight in a species. A simple linear relationship for this would contain an intercept (β_0) and a parameter linking weight and age (β_1):

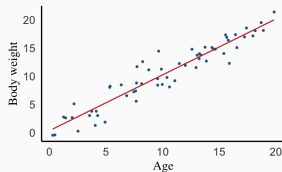
$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$



What this model is saying is that for any age_i, weight_i will be normally distributed, with $\mu_i = \beta_0 + \beta_1 \text{age}_i$

Let's say we're interested in the relationship between age and weight in a species. A simple linear relationship for this would contain an intercept (β_0) and a parameter linking weight and age (β_1):

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

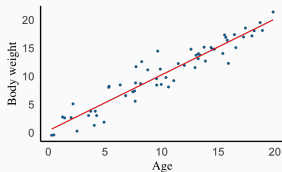


What this model is saying is that for any age_i, weight_i will be normally distributed, with $\mu_i = \beta_0 + \beta_1 \text{age}_i$

The Gaussian distribution is defined as: $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Let's say we're interested in the relationship between age and weight in a species. A simple linear relationship for this would contain an intercept (β_0) and a parameter linking weight and age (β_1):

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$



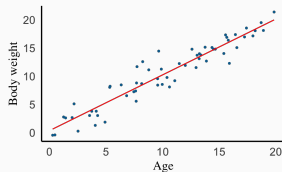
What this model is saying is that for any age_i, weight_i will be normally distributed, with $\mu_i = \beta_0 + \beta_1 \text{age}_i$

The Gaussian distribution is defined as: $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

so rearranging, we get: $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma}\right)^2}$

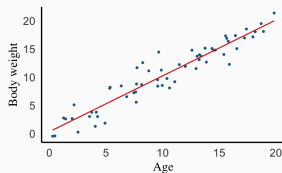
$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma}\right)^2}$$



$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

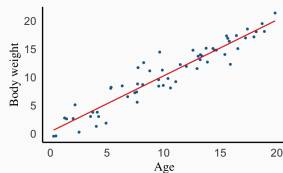
$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma} \right)^2}$$



Setting the problem up this way allows us to calculate the probability of obtaining any specific weight

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

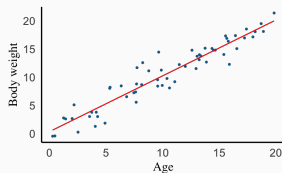
$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma} \right)^2}$$



Setting the problem up this way allows us to calculate the probability of obtaining any specific weight, and what range of weights are possible.

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma} \right)^2}$$

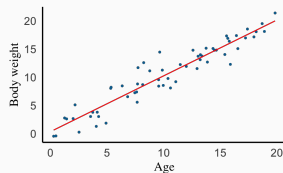


Setting the problem up this way allows us to calculate the probability of obtaining any specific weight, and what range of weights are possible.

So let's say $\sigma = 2$, $\beta_0 = 0$, $\beta_1 = 1$, $\text{age} = 1$

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma}\right)^2}$$

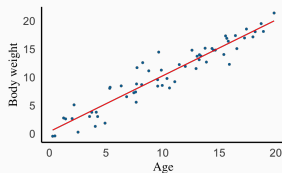


Setting the problem up this way allows us to calculate the probability of obtaining any specific weight, and what range of weights are possible.

So let's say $\sigma = 2$, $\beta_0 = 0$, $\beta_1 = 1$, age = 1, here the probability of a weight of 1g ~ 0.20

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma}\right)^2}$$

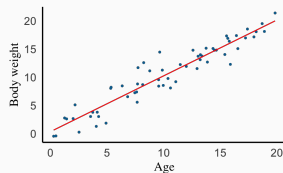


Setting the problem up this way allows us to calculate the probability of obtaining any specific weight, and what range of weights are possible.

So let's say $\sigma = 2$, $\beta_0 = 0$, $\beta_1 = 1$, age = 1, here the probability of a weight of 1g ~ 0.20 , of 2g ~ 0.18

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma}\right)^2}$$

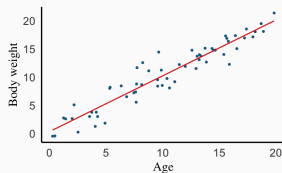


Setting the problem up this way allows us to calculate the probability of obtaining any specific weight, and what range of weights are possible.

So let's say $\sigma = 2$, $\beta_0 = 0$, $\beta_1 = 1$, age = 1, here the probability of a weight of 1g ~ 0.20 , of 2g ~ 0.18 , of 3g ~ 0.12

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma}\right)^2}$$

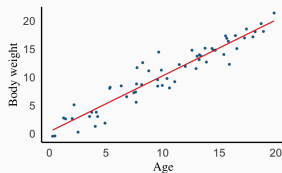


Setting the problem up this way allows us to calculate the probability of obtaining any specific weight, and what range of weights are possible.

So let's say $\sigma = 2$, $\beta_0 = 0$, $\beta_1 = 1$, age = 1, here the probability of a weight of 1g ~ 0.20 , of 2g ~ 0.18 , of 3g ~ 0.12 , and so on...

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma}\right)^2}$$



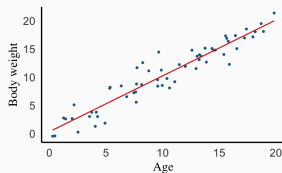
Setting the problem up this way allows us to calculate the probability of obtaining any specific weight, and what range of weights are possible.

So let's say $\sigma = 2$, $\beta_0 = 0$, $\beta_1 = 1$, age = 1, here the probability of a weight of 1g ~ 0.20 , of 2g ~ 0.18 , of 3g ~ 0.12 , and so on...

What's the range of the Gaussian distribution?

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma}\right)^2}$$



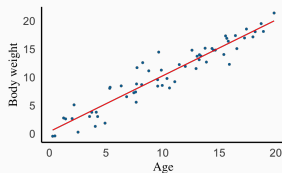
Setting the problem up this way allows us to calculate the probability of obtaining any specific weight, and what range of weights are possible.

So let's say $\sigma = 2$, $\beta_0 = 0$, $\beta_1 = 1$, age = 1, here the probability of a weight of 1g ~ 0.20 , of 2g ~ 0.18 , of 3g ~ 0.12 , and so on...

What's the range of the Gaussian distribution? $-\infty, \infty$

$$\text{weight}_i = \beta_0 + \beta_1 \text{age}_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\text{weight}_i - (\beta_0 + \beta_1 \text{age}_i)}{\sigma}\right)^2}$$



Setting the problem up this way allows us to calculate the probability of obtaining any specific weight, and what range of weights are possible.

So let's say $\sigma = 2$, $\beta_0 = 0$, $\beta_1 = 1$, age = 1, here the probability of a weight of 1g ~ 0.20 , of 2g ~ 0.18 , of 3g ~ 0.12 , and so on...

What's the range of the Gaussian distribution? $-\infty, \infty$

This means that if we set up our problem this way our model is telling us that there's some chance of getting a weight of -1g (~ 0.12)



When we fit any of the models we've been working with so far, we are assuming our residuals should be normally distributed and that our response can take any value between $-\infty, \infty$

When we fit any of the models we've been working with so far, we are assuming our residuals should be normally distributed and that our response can take any value between $-\infty, \infty$

What can you do if this is not a reasonable assumption for your data?

When we fit any of the models we've been working with so far, we are assuming our residuals should be normally distributed and that our response can take any value between $-\infty, \infty$

What can you do if this is not a reasonable assumption for your data?

- **Nothing.** If the residuals are normally distributed and the spread isn't bad, this isn't a terrible assumption (remember, no model is going to be correct).

When we fit any of the models we've been working with so far, we are assuming our residuals should be normally distributed and that our response can take any value between $-\infty, \infty$

What can you do if this is not a reasonable assumption for your data?

- **Nothing.** If the residuals are normally distributed and the spread isn't bad, this isn't a terrible assumption (remember, no model is going to be correct).
- **Transform your data.** Shoe-horning your data to fit the assumptions of normality can work, but it changes the relationship between your response and your predictors.

When we fit any of the models we've been working with so far, we are assuming our residuals should be normally distributed and that our response can take any value between $-\infty, \infty$

What can you do if this is not a reasonable assumption for your data?

- **Nothing.** If the residuals are normally distributed and the spread isn't bad, this isn't a terrible assumption (remember, no model is going to be correct).
- **Transform your data.** Shoe-horning your data to fit the assumptions of normality can work, but it changes the relationship between your response and your predictors.
- **Choose another distribution.**

Generalised Linear Models (GLMs)



"In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution." — Wikipedia

"In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution." — Wikipedia

So if our models are of this form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

it should simply be a matter of swapping out this bit $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ right?

"In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution." — Wikipedia

So if our models are of this form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

it should simply be a matter of swapping out this bit $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ right?

The challenge is that not all distributions have a mean and a variance so you can't simply interchange them.

In 1972, Nelder & Wedderburn (1972) worked out a generalisation of the linear regression model.

In 1972, Nelder & Wedderburn (1972) worked out a generalisation of the linear regression model. They extended the models we've been working with so far to any member of the family exponential distributions (Gaussian, Poisson, binomial, negative binomial, gamma, etc.).

In 1972, Nelder & Wedderburn (1972) worked out a generalisation of the linear regression model. They extended the models we've been working with so far to any member of the family exponential distributions (Gaussian, Poisson, binomial, negative binomial, gamma, etc.).

They showed how all of these distributions can be expressed by the general formulation:

$$f(Y; \theta, \phi) = e^{\frac{y \times \theta - b(\theta)}{a(\phi)}} + c(y, \theta)$$

In 1972, Nelder & Wedderburn (1972) worked out a generalisation of the linear regression model. They extended the models we've been working with so far to any member of the family exponential distributions (Gaussian, Poisson, binomial, negative binomial, gamma, etc.).

They showed how all of these distributions can be expressed by the general formulation:

$$f(Y; \theta, \phi) = e^{\frac{y \times \theta - b(\theta)}{a(\phi)} + c(y, \theta)}$$

I.e., if we fix certain pieces of this formulation to different values we can get back any member of the family exponential distributions.

In 1972, Nelder & Wedderburn (1972) worked out a generalisation of the linear regression model. They extended the models we've been working with so far to any member of the family exponential distributions (Gaussian, Poisson, binomial, negative binomial, gamma, etc.).

They showed how all of these distributions can be expressed by the general formulation:

$$f(Y; \theta, \phi) = e^{\frac{y \times \theta - b(\theta)}{a(\phi)} + c(y, \theta)}$$

I.e., if we fix certain pieces of this formulation to different values we can get back any member of the family exponential distributions.

This means that a single set of equations (and estimators) can be used for all of these different distributions.



Now that we have a general expression for the stochastic component of our model, we just need to find a way to 'link' the expectation value of our model with the expectation value of the distribution.

Now that we have a general expression for the stochastic component of our model, we just need to find a way to 'link' the expectation value of our model with the expectation value of the distribution.

To do this we need to carry out 3 steps when fitting GLMs:

Now that we have a general expression for the stochastic component of our model, we just need to find a way to 'link' the expectation value of our model with the expectation value of the distribution.

To do this we need to carry out 3 steps when fitting GLMs:

1. Make a distributional assumption on the response variable Y_i . This also defines the mean and variance of Y_i .

Now that we have a general expression for the stochastic component of our model, we just need to find a way to 'link' the expectation value of our model with the expectation value of the distribution.

To do this we need to carry out 3 steps when fitting GLMs:

1. Make a distributional assumption on the response variable Y_i . This also defines the mean and variance of Y_i .
2. Specify the deterministic part of the model.

Now that we have a general expression for the stochastic component of our model, we just need to find a way to 'link' the expectation value of our model with the expectation value of the distribution.

To do this we need to carry out 3 steps when fitting GLMs:

1. Make a distributional assumption on the response variable Y_i . This also defines the mean and variance of Y_i .
2. Specify the deterministic part of the model.
3. Formally specify the 'link' between the mean of Y_i and the deterministic part based on your distributional assumption.

Gaussian Linear Regression as a GLM

Step 1: Distributional Assumption



THE UNIVERSITY OF BRITISH COLUMBIA
Okanagan Campus

The first step of a GLM is to make a distributional assumption on the response variable Y_i .

The first step of a GLM is to make a distributional assumption on the response variable Y_i .

For standard, Gaussian linear regression this assumption is that

$$Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

The first step of a GLM is to make a distributional assumption on the response variable Y_i .

For standard, Gaussian linear regression this assumption is that

$$Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

This means that:

$$E(Y_i) = \mu_i \quad \text{and} \quad \text{var}(Y_i) = \sigma^2$$

Step 2: Specify deterministic part



THE UNIVERSITY OF BRITISH COLUMBIA
Okanagan Campus

The second step of a GLM is to specify deterministic part, also called the linear predictor:

$$\eta = X\beta.$$

η is expressed as linear combinations of unknown parameters β and the matrix of independent variables X

The second step of a GLM is to specify deterministic part, also called the linear predictor:

$$\eta = X\beta.$$

η is expressed as linear combinations of unknown parameters β and the matrix of independent variables X

In long form, this would look something like:

$$\eta = \beta_0 + \beta_1 X_{i1} \dots + \beta_n X_{in}$$

Step 3: Specify the link



THE UNIVERSITY OF BRITISH COLUMBIA
Okanagan Campus

Step 3: Specify the link



The third step of a GLM is to specify the 'link' between the expected value of Y_i and the deterministic part based on your distributional assumption.

Step 3: Specify the link



The third step of a GLM is to specify the 'link' between the expected value of Y_i and the deterministic part based on your distributional assumption.

The expected value that Y_i should take (without the stochastic component) is:

$$E(Y_i) = \mu_i.$$

Step 3: Specify the link



The third step of a GLM is to specify the 'link' between the expected value of Y_i and the deterministic part based on your distributional assumption.

The expected value that Y_i should take (without the stochastic component) is:

$$E(Y_i) = \mu_i.$$

So now we need to link η and $E(Y_i)$ based on our distributional assumption.

Step 3: Specify the link



The third step of a GLM is to specify the 'link' between the expected value of Y_i and the deterministic part based on your distributional assumption.

The expected value that Y_i should take (without the stochastic component) is:

$$E(Y_i) = \mu_i.$$

So now we need to link η and $E(Y_i)$ based on our distributional assumption.

What's the expected value of a Gaussian distribution?

Step 3: Specify the link



The third step of a GLM is to specify the 'link' between the expected value of Y_i and the deterministic part based on your distributional assumption.

The expected value that Y_i should take (without the stochastic component) is:

$$E(Y_i) = \mu_i.$$

So now we need to link η and $E(Y_i)$ based on our distributional assumption.

What's the expected value of a Gaussian distribution? The mean, μ .

Step 3: Specify the link



The third step of a GLM is to specify the 'link' between the expected value of Y_i and the deterministic part based on your distributional assumption.

The expected value that Y_i should take (without the stochastic component) is:

$$E(Y_i) = \mu_i.$$

So now we need to link η and $E(Y_i)$ based on our distributional assumption.

What's the expected value of a Gaussian distribution? The mean, μ .

What's the expected value of the deterministic model?

Step 3: Specify the link



The third step of a GLM is to specify the 'link' between the expected value of Y_i and the deterministic part based on your distributional assumption.

The expected value that Y_i should take (without the stochastic component) is:

$$E(Y_i) = \mu_i.$$

So now we need to link η and $E(Y_i)$ based on our distributional assumption.

What's the expected value of a Gaussian distribution? The mean, μ .

What's the expected value of the deterministic model? $\eta = X\beta$.

Step 3: Specify the link



The third step of a GLM is to specify the 'link' between the expected value of Y_i and the deterministic part based on your distributional assumption.

The expected value that Y_i should take (without the stochastic component) is:

$$E(Y_i) = \mu_i.$$

So now we need to link η and $E(Y_i)$ based on our distributional assumption.

What's the expected value of a Gaussian distribution? The mean, μ .

What's the expected value of the deterministic model? $\eta = X\beta$.

So here $\mu_i = \eta = X\beta$.

Step 3: Specify the link



The third step of a GLM is to specify the 'link' between the expected value of Y_i and the deterministic part based on your distributional assumption.

The expected value that Y_i should take (without the stochastic component) is:

$$E(Y_i) = \mu_i.$$

So now we need to link η and $E(Y_i)$ based on our distributional assumption.

What's the expected value of a Gaussian distribution? The mean, μ .

What's the expected value of the deterministic model? $\eta = X\beta$.

So here $\mu_i = \eta = X\beta$. We call this the identity link.



A GLM with a Gaussian distribution and an identity link is given by:

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$E(Y_i) = \mu_i \quad \text{and} \quad \text{var}(Y_i) = \sigma^2$$

$$\mu_i = \eta = X\beta$$

GLMs for count data



As biologists we often find ourselves sitting around counting things.



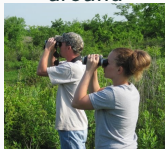
Source: Biocompare

As biologists we often find ourselves sitting around counting things.



Source: Biocompare

... or standing
around



Source: NPS

As biologists we often find ourselves sitting around counting things.



Source: Biocompare

... or standing
around



Source: NPS

... or kneeling
around



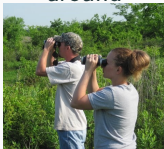
Source: Govt. of Western
Australia

As biologists we often find ourselves sitting around counting things.



Source: Biocompare

... or standing
around



Source: NPS

... or kneeling
around



Source: Govt. of Western
Australia

... or diving
around



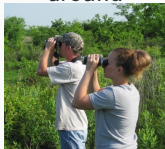
Source:
<http://educationcareerarticles.com>

As biologists we often find ourselves sitting around counting things.



Source: Biocompare

... or standing
around



Source: NPS

... or kneeling
around



Source: Govt. of Western
Australia

... or diving
around



Source:
<http://educationcareerarticles.com>

... we count a lot of things.



If we want to model count data in a GLM framework the first step is identify the right distribution.

If we want to model count data in a GLM framework the first step is identify the right distribution.

Count data usually range between 0 and ∞ .

If we want to model count data in a GLM framework the first step is identify the right distribution.

Count data usually range between 0 and ∞ .

They're also usually discrete integers because we don't count fractions of things (unless those things were very unlucky...).

If we want to model count data in a GLM framework the first step is identify the right distribution.

Count data usually range between 0 and ∞ .

They're also usually discrete integers because we don't count fractions of things (unless those things were very unlucky...).

The Gaussian distribution is continuous and has support between $-\infty, \infty$

If we want to model count data in a GLM framework the first step is identify the right distribution.

Count data usually range between 0 and ∞ .

They're also usually discrete integers because we don't count fractions of things (unless those things were very unlucky...).

The Gaussian distribution is continuous and has support between $-\infty, \infty$, so we can already tell it's probably not a good option.

If we want to model count data in a GLM framework the first step is identify the right distribution.

Count data usually range between 0 and ∞ .

They're also usually discrete integers because we don't count fractions of things (unless those things were very unlucky...).

The Gaussian distribution is continuous and has support between $-\infty, \infty$, so we can already tell it's probably not a good option.

What we're looking for is a discrete distribution with support between 0 and ∞ . Any ideas?

If we want to model count data in a GLM framework the first step is identify the right distribution.

Count data usually range between 0 and ∞ .

They're also usually discrete integers because we don't count fractions of things (unless those things were very unlucky...).

The Gaussian distribution is continuous and has support between $-\infty, \infty$, so we can already tell it's probably not a good option.

What we're looking for is a discrete distribution with support between 0 and ∞ . Any ideas? The Poisson distribution is a good candidate for modelling count data.

The Poisson distribution describes the probability of a given number of events occurring in a fixed interval of time or space.

Parameters: λ

Type: Discrete

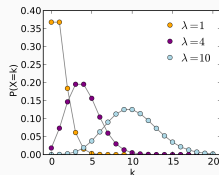
Biological scenarios: Counts of a species per unit time, the number of mutations on a strand of DNA per unit length, number of births/deaths per year in a given age group, prey caught per unit time.

PMF: $\Pr(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

Range: discrete $(0, \infty)$

Mean: λ

Variance: λ



Source: Wikipedia



So after step 1 we get:

$$Y_i \sim P(\lambda = \mu_i)$$

So after step 1 we get:

$$Y_i \sim P(\lambda = \mu_i)$$

The second step of a GLM is to specify deterministic part:

$$\eta = \beta_0 + \beta_1 X_{i1} \dots + \beta_n X_{in}$$

So after step 1 we get:

$$Y_i \sim P(\lambda = \mu_i)$$

The second step of a GLM is to specify deterministic part:

$$\eta = \beta_0 + \beta_1 X_{i1} \dots + \beta_n X_{in}$$

Note how this step hasn't changed.

So after step 1 we get:

$$Y_i \sim P(\lambda = \mu_i)$$

The second step of a GLM is to specify deterministic part:

$$\eta = \beta_0 + \beta_1 X_{i1} \dots + \beta_n X_{in}$$

Note how this step hasn't changed.

The last step is to link η and μ_i .

So after step 1 we get:

$$Y_i \sim P(\lambda = \mu_i)$$

The second step of a GLM is to specify deterministic part:

$$\eta = \beta_0 + \beta_1 X_{i1} \dots + \beta_n X_{in}$$

Note how this step hasn't changed.

The last step is to link η and μ_i . Because η can be positive or negative, we can't use an identity link.

So after step 1 we get:

$$Y_i \sim P(\lambda = \mu_i)$$

The second step of a GLM is to specify deterministic part:

$$\eta = \beta_0 + \beta_1 X_{i1} \dots + \beta_n X_{in}$$

Note how this step hasn't changed.

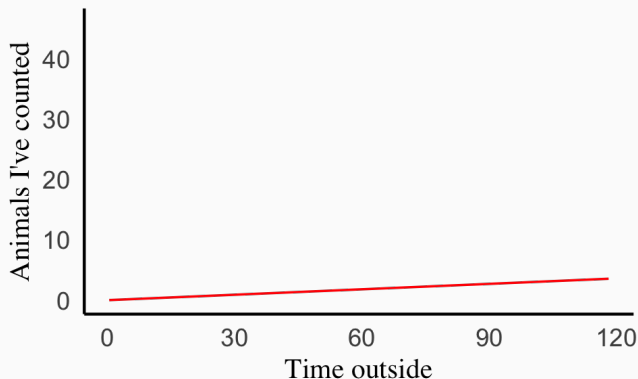
The last step is to link η and μ_i . Because η can be positive or negative, we can't use an identity link. Instead, we use a log-link to ensure the fitted values are always positive:

$$\log(\mu_i) = \eta \quad \text{or} \quad \mu_i = e^\eta$$



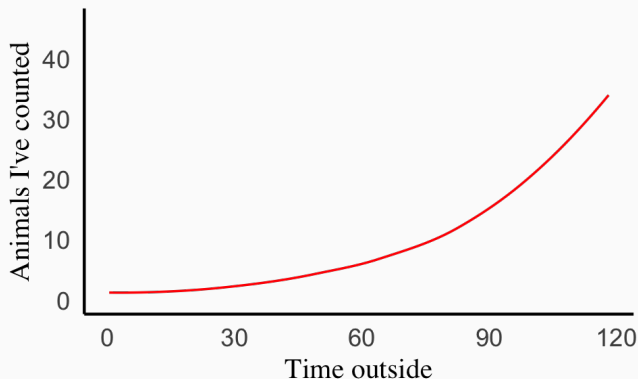
Let's say I have a model describing how many animals I see depending on how long I sit on my back porch:

$$\mu_i = 0.01 + 0.03 \times X_i$$

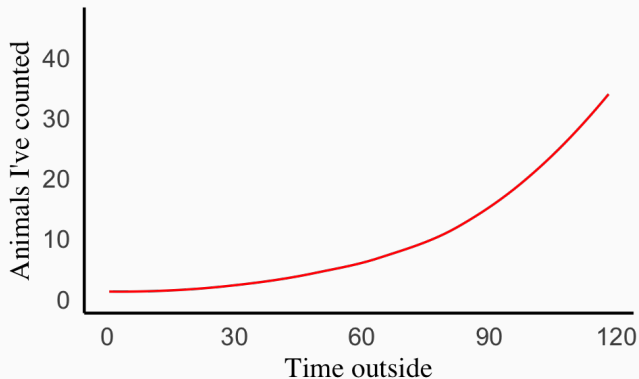


Let's say I have a model describing how many animals I see depending on how long I sit on my back porch:

$\mu_i = 0.01 + 0.03 \times X_i$ with a log link this becomes: $\mu_i = e^{0.01+0.03 \times X_i}$

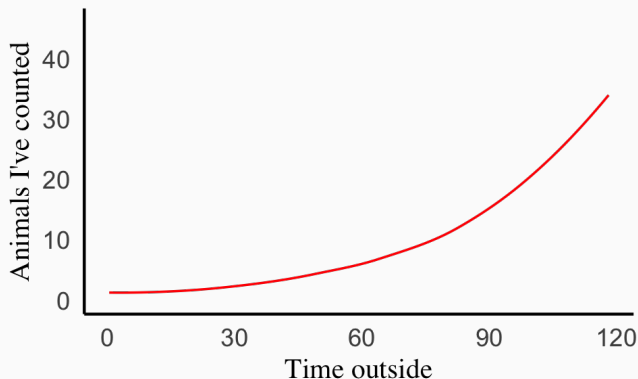


But there will probably be some variance from 'experiment' to 'experiment'



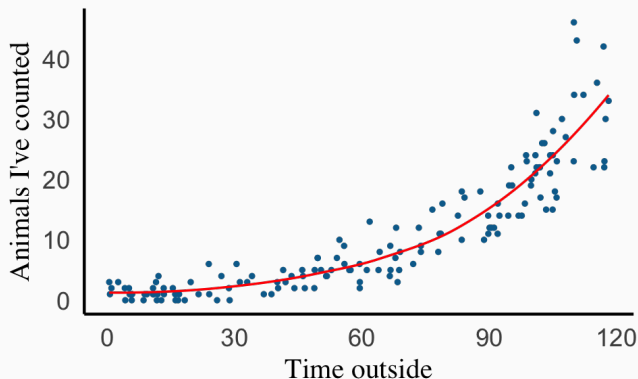
But there will probably be some variance from 'experiment' to 'experiment'

So: $Y_i \sim \text{Poisson}(\lambda = e^{0.01+0.03 \times X_i})$

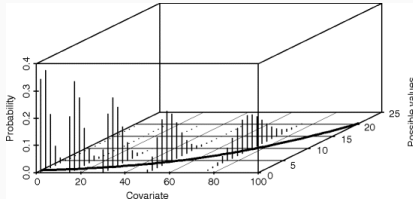
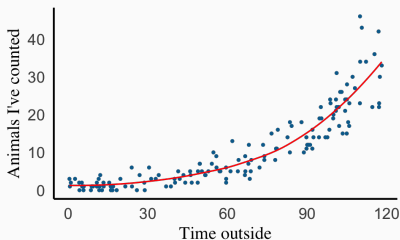


But there will probably be some variance from 'experiment' to 'experiment'

So: $Y_i \sim \text{Poisson}(\lambda = e^{0.01+0.03 \times X_i})$, giving us Poisson distributed errors



So when we fit a Poisson GLM we're fitting a curve of the form e^{η} with Poisson distributed errors at each μ_i

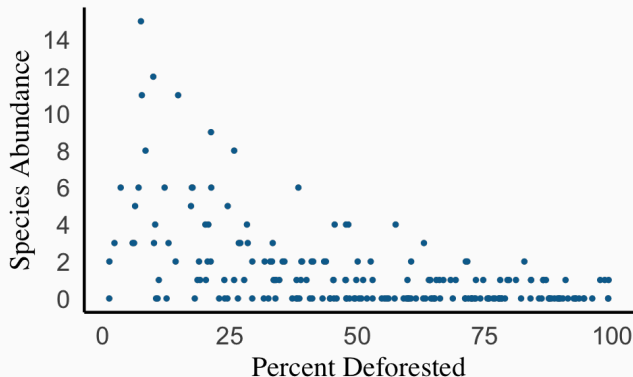


Source: Zuur et al. 2009

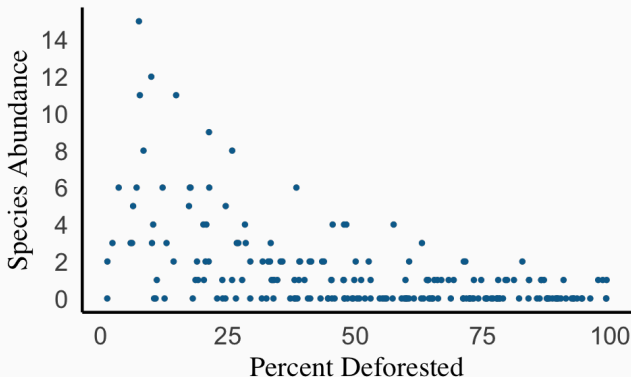
GLMs for count data in R



We'll work with a simulated dataset of species abundance as a function of deforestation.



We'll work with a simulated dataset of species abundance as a function of deforestation.



With these data we're interested in knowing what whether deforestation influence species abundance.





We already know a Gaussian model isn't a great choice, but let's see what that would look like.

We already know a Gaussian model isn't a great choice, but let's see what that would look like.

```
library(nlme)

FIT <- gls(Abundance ~ Deforestation, data = DATA,
          method = "ML")

summary(FIT)

Generalized least squares fit by maximum likelihood
Model: Abundance ~ Deforestation
Data: DATA
      AIC      BIC    logLik
849.5338 859.4287 -421.7669

Coefficients:
              Value Std.Error   t-value p-value
(Intercept)  3.797854 0.30782465 12.337720      0
Deforestation -0.045935 0.00527925 -8.700991      0
```

We already know a Gaussian model isn't a great choice, but let's see what that would look like.

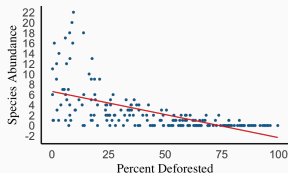
```
library(nlme)

FIT <- gls(Abundance ~ Deforestation, data = DATA,
           method = "ML")

summary(FIT)

Generalized least squares fit by maximum likelihood
Model: Abundance ~ Deforestation
Data: DATA
      AIC      BIC    logLik
849.5338 859.4287 -421.7669

Coefficients:
              Value Std.Error   t-value p-value
(Intercept)  3.797854 0.30782465 12.337720      0
Deforestation -0.045935 0.00527925 -8.700991      0
```



We already know a Gaussian model isn't a great choice, but let's see what that would look like.

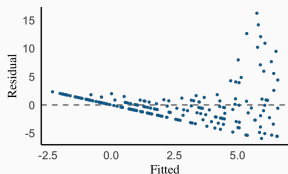
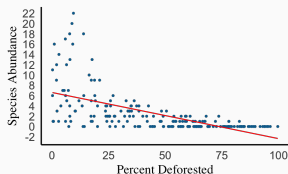
```
library(nlme)

FIT <- gls(Abundance ~ Deforestation, data = DATA,
          method = "ML")

summary(FIT)

Generalized least squares fit by maximum likelihood
Model: Abundance ~ Deforestation
Data: DATA
      AIC      BIC    logLik
849.5338 859.4287 -421.7669

Coefficients:
              Value Std. Error  t-value p-value
(Intercept)  3.797854  0.30782465  12.337720      0
Deforestation -0.045935  0.00527925  -8.700991      0
```



We already know a Gaussian model isn't a great choice, but let's see what that would look like.

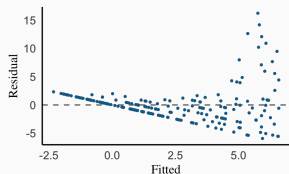
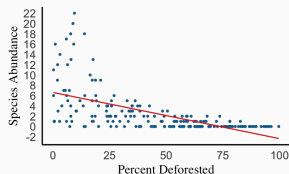
```
library(nlme)

FIT <- gls(Abundance ~ Deforestation, data = DATA,
          method = "ML")

summary(FIT)

Generalized least squares fit by maximum likelihood
Model: Abundance ~ Deforestation
Data: DATA
      AIC      BIC    logLik
849.5338 859.4287 -421.7669

Coefficients:
              Value Std. Error  t-value p-value
(Intercept)  3.797854  0.30782465  12.337720      0
Deforestation -0.045935  0.00527925  -8.700991      0
```



The residuals look terrible.



To improve this we can carry out the 3 steps of fitting a GLM:

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:**

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance.

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:**

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:** Specify η .

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:**

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η .

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
FIT2 <- glm(Abundance ~ Deforestation,  
            family = poisson(link = "log"),  
            data = DATA)
```

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
FIT2 <- glm(Abundance ~ Deforestation,
            family = poisson(link = "log"),
            data = DATA)

summary(FIT2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.805377   0.097964   18.43  <2e-16
Deforestation -0.037026   0.002738  -13.52  <2e-16
---
(Dispersion parameter for poisson family taken to be
 1)

Null deviance: 570.51  on 199  degrees of freedom
Residual deviance: 342.97  on 198  degrees of freedom
AIC: 618.5
```

To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
FIT2 <- glm(Abundance ~ Deforestation,  
            family = poisson(link = "log"),  
            data = DATA)
```

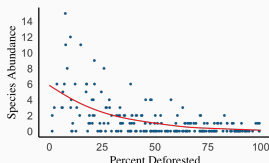
```
summary(FIT2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.805377	0.097964	18.43	<2e-16
Deforestation	-0.037026	0.002738	-13.52	<2e-16

(Dispersion parameter for poisson family taken to be
1)

Null deviance: 570.51 on 199 degrees of freedom
Residual deviance: 342.97 on 198 degrees of freedom
AIC: 618.5



To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
FIT2 <- glm(Abundance ~ Deforestation,
            family = poisson(link = "log"),
            data = DATA)
```

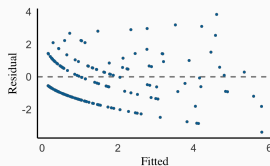
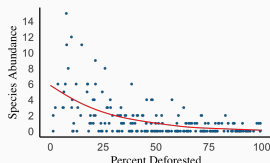
```
summary(FIT2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.805377	0.097964	18.43	<2e-16
Deforestation	-0.037026	0.002738	-13.52	<2e-16

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 570.51 on 199 degrees of freedom
Residual deviance: 342.97 on 198 degrees of freedom
AIC: 618.5



To improve this we can carry out the 3 steps of fitting a GLM:

1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
FIT2 <- glm(Abundance ~ Deforestation,  
            family = poisson(link = "log"),  
            data = DATA)
```

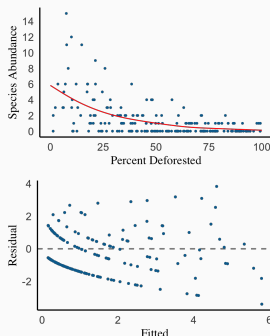
summary(FIT2)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.805377	0.097964	18.43	<2e-16
Deforestation	-0.037026	0.002738	-13.52	<2e-16

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 570.51 on 199 degrees of freedom
Residual deviance: 342.97 on 198 degrees of freedom
AIC: 618.5



Are the residuals normally distributed?

To improve this we can carry out the 3 steps of fitting a GLM:

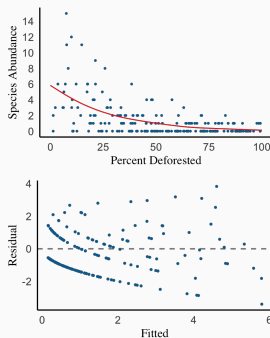
1. **Step 1:** Make a distributional assumption on abundance. Poisson.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
FIT2 <- glm(Abundance ~ Deforestation,
            family = poisson(link = "log"),
            data = DATA)

summary(FIT2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.805377   0.097964   18.43  <2e-16
Deforestation -0.037026   0.002738  -13.52  <2e-16
---
(Dispersion parameter for poisson family taken to be
 1)

Null deviance: 570.51  on 199  degrees of freedom
Residual deviance: 342.97  on 198  degrees of freedom
AIC: 618.5
```



Are the residuals normally distributed? Should they be?

```
FIT2 <- glm(Abundance ~ Deforestation,  
            family = poisson(link = "log"),  
            data = DATA)  
  
summary(FIT2)  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)  1.805377   0.097964   18.43  <2e-16  
Deforestation -0.037026   0.002738  -13.52  <2e-16  
---  
(Dispersion parameter for poisson family taken to be  
 1)  
  
Null deviance: 570.51  on 199  degrees of freedom  
Residual deviance: 342.97  on 198  degrees of freedom  
AIC: 618.5
```



```
FIT2 <- glm(Abundance ~ Deforestation,  
            family = poisson(link = "log"),  
            data = DATA)  
  
summary(FIT2)  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)   1.805377   0.097964   18.43  <2e-16  
Deforestation -0.037026   0.002738  -13.52  <2e-16  
---  
(Dispersion parameter for poisson family taken to be  
 1)  
  
Null deviance: 570.51  on 199  degrees of freedom  
Residual deviance: 342.97  on 198  degrees of freedom  
AIC: 618.5
```

Notice how we don't have
'residuals' anymore.

```
FIT2 <- glm(Abundance ~ Deforestation,  
            family = poisson(link = "log"),  
            data = DATA)  
  
summary(FIT2)  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)   1.805377   0.097964   18.43  <2e-16  
Deforestation -0.037026   0.002738  -13.52  <2e-16  
---  
(Dispersion parameter for poisson family taken to be  
 1)  
  
Null deviance: 570.51  on 199  degrees of freedom  
Residual deviance: 342.97  on 198  degrees of freedom  
AIC: 618.5
```

Notice how we don't have
'residuals' anymore. Instead we
have 'deviances'.

```
FIT2 <- glm(Abundance ~ Deforestation,  
            family = poisson(link = "log"),  
            data = DATA)  
  
summary(FIT2)  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)   1.805377   0.097964   18.43  <2e-16  
Deforestation -0.037026   0.002738  -13.52  <2e-16  
---  
(Dispersion parameter for poisson family taken to be  
 1)  
  
Null deviance: 570.51  on 199  degrees of freedom  
Residual deviance: 342.97  on 198  degrees of freedom  
AIC: 618.5
```

Notice how we don't have 'residuals' anymore. Instead we have 'deviances'.

Think of null and residual deviances as GLM equivalents of total and residual and sum of squares.

```
FIT2 <- glm(Abundance ~ Deforestation,  
            family = poisson(link = "log"),  
            data = DATA)  
  
summary(FIT2)  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)   1.805377   0.097964   18.43  <2e-16  
Deforestation -0.037026   0.002738  -13.52  <2e-16  
---  
(Dispersion parameter for poisson family taken to be  
 1)  
  
Null deviance: 570.51  on 199  degrees of freedom  
Residual deviance: 342.97  on 198  degrees of freedom  
AIC: 618.5
```

Notice how we don't have 'residuals' anymore. Instead we have 'deviances'.

Think of null and residual deviances as GLM equivalents of total and residual and sum of squares.

GLMs don't have an R^2 .

```
FIT2 <- glm(Abundance ~ Deforestation,  
            family = poisson(link = "log"),  
            data = DATA)  
  
summary(FIT2)  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)   1.805377   0.097964   18.43  <2e-16  
Deforestation -0.037026   0.002738  -13.52  <2e-16  
---  
(Dispersion parameter for poisson family taken to be  
1)  
  
Null deviance: 570.51  on 199  degrees of freedom  
Residual deviance: 342.97  on 198  degrees of freedom  
AIC: 618.5
```

Notice how we don't have 'residuals' anymore. Instead we have 'deviances'.

Think of null and residual deviances as GLM equivalents of total and residual and sum of squares.

GLMs don't have an R^2 . An approximation for this would be:

$$\frac{\text{null deviance} - \text{resid deviance}}{\text{null deviance}} \times 100$$

```
FIT2 <- glm(Abundance ~ Deforestation,  
            family = poisson(link = "log"),  
            data = DATA)  
  
summary(FIT2)  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)  1.805377   0.097964   18.43  <2e-16  
Deforestation -0.037026   0.002738  -13.52  <2e-16  
---  
(Dispersion parameter for poisson family taken to be  
 1)  
  
Null deviance: 570.51  on 199  degrees of freedom  
Residual deviance: 342.97  on 198  degrees of freedom  
AIC: 618.5
```

Notice how we don't have 'residuals' anymore. Instead we have 'deviances'.

Think of null and residual deviances as GLM equivalents of total and residual and sum of squares.

GLMs don't have an R^2 . An approximation for this would be:

$$\frac{\text{null deviance} - \text{resid deviance}}{\text{null deviance}} \times 100$$
$$\frac{570.51 - 342.97}{570.51} \times 100 = 39.9\%$$



So far we've been using residuals to assess a model's fit

So far we've been using residuals to assess a model's fit, but we just saw that GLMs don't have 'residuals' in an OLS sense...

So far we've been using residuals to assess a model's fit, but we just saw that GLMs don't have 'residuals' in an OLS sense...

Residuals are just observed - expected ($y_i - \mu_i$), which we can calculate.

So far we've been using residuals to assess a model's fit, but we just saw that GLMs don't have 'residuals' in an OLS sense...

Residuals are just observed - expected ($y_i - \mu_i$), which we can calculate.

So if our GLM model is performing well the spread in our predictions should be even across the full range of deforestation values right?

So far we've been using residuals to assess a model's fit, but we just saw that GLMs don't have 'residuals' in an OLS sense...

Residuals are just observed - expected ($y_i - \mu_i$), which we can calculate.

So if our GLM model is performing well the spread in our predictions should be even across the full range of deforestation values right?

The mean and variance of the Poisson distribution are the same, so the spread will change for different values of μ_i .

So far we've been using residuals to assess a model's fit, but we just saw that GLMs don't have 'residuals' in an OLS sense...

Residuals are just observed - expected ($y_i - \mu_i$), which we can calculate.

So if our GLM model is performing well the spread in our predictions should be even across the full range of deforestation values right?

The mean and variance of the Poisson distribution are the same, so the spread will change for different values of μ_i . This makes Poisson GLM (and all GLM) residuals very difficult to interpret.

So far we've been using residuals to assess a model's fit, but we just saw that GLMs don't have 'residuals' in an OLS sense...

Residuals are just observed - expected ($y_i - \mu_i$), which we can calculate.

So if our GLM model is performing well the spread in our predictions should be even across the full range of deforestation values right?

The mean and variance of the Poisson distribution are the same, so the spread will change for different values of μ_i . This makes Poisson GLM (and all GLM) residuals very difficult to interpret.

The most important thing to look for are patterns and a lack of fit.

So far we've been using residuals to assess a model's fit, but we just saw that GLMs don't have 'residuals' in an OLS sense...

Residuals are just observed - expected ($y_i - \mu_i$), which we can calculate.

So if our GLM model is performing well the spread in our predictions should be even across the full range of deforestation values right?

The mean and variance of the Poisson distribution are the same, so the spread will change for different values of μ_i . This makes Poisson GLM (and all GLM) residuals very difficult to interpret.

The most important thing to look for are patterns and a lack of fit.

See Zuur et al. 2009 Section 9.8 for a detailed discussion of GLM residuals



Switching from a Gaussian distribution to a Poisson distribution is often a good fix for modelling count data

Switching from a Gaussian distribution to a Poisson distribution is often a good fix for modelling count data, but it's not always the most appropriate dist. for count data.

Switching from a Gaussian distribution to a Poisson distribution is often a good fix for modelling count data, but it's not always the most appropriate dist. for count data.

One of the primary reasons why a Poisson won't work very well on count data is over-dispersion (because the variance is tied to the mean and therefore less flexible).

Switching from a Gaussian distribution to a Poisson distribution is often a good fix for modelling count data, but it's not always the most appropriate dist. for count data.

One of the primary reasons why a Poisson won't work very well on count data is over-dispersion (because the variance is tied to the mean and therefore less flexible).

Can you think of another option worth considering?

The negative binomial distribution describes the number of *failures* in a sequence of independent and identically distributed trials.

Parameters: p Probability per trial,
 k Overdispersion parameter

Type: Discrete

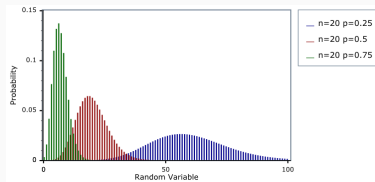
Biological scenarios: Same as the Poisson distribution, but allowing for more heterogeneity because variance \neq mean.

$$\text{PMF: } \frac{\Gamma(k+r)}{k! \cdot \Gamma(r)} p^k (1-p)^r$$

Range: discrete ($x \geq 0$)

$$\text{Mean: } \frac{pr}{1-p}$$

$$\text{Variance: } \frac{pr}{(1-p)^2}$$



Negative binomial GLM on count data



THE UNIVERSITY OF BRITISH COLUMBIA
Okanagan Campus



1. Step 1:



1. **Step 1:** Distributional assumption.

1. **Step 1:** Distributional assumption. Negative Binomial.

1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:**

1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η .

1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$

1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:**

1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η .

1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
library(MASS)
```


1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
library(MASS)
```

```
FIT3 <- glm.nb(Abundance ~ Deforestation,  
              link = "log",  
              data = DATA)
```

1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
library(MASS)

FIT3 <- glm.nb(Abundance ~ Deforestation,
               link = "log",
               data = DATA)

summary(FIT3)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.767992   0.166842  10.597  <2e-16
Deforestation -0.036120   0.003747  -9.639  <2e-16
---
(Dispersion parameter for Negative Binomial(1.6398)
 family taken to be 1)

Null deviance: 312.80  on 199  degrees of freedom
Residual deviance: 194.39  on 198  degrees of freedom
AIC: 566.42
```

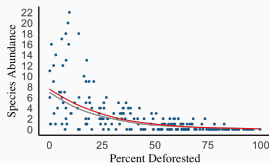
1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
library(MASS)

FIT3 <- glm.nb(Abundance ~ Deforestation,
               link = "log",
               data = DATA)

summary(FIT3)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.767992   0.166842  10.597 <2e-16
Deforestation -0.036120   0.003747  -9.639 <2e-16
---
(Dispersion parameter for Negative Binomial(1.6398)
 family taken to be 1)

Null deviance: 312.80  on 199  degrees of freedom
Residual deviance: 194.39  on 198  degrees of freedom
AIC: 566.42
```



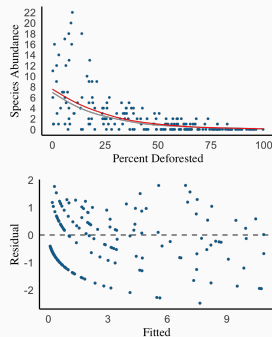
- Step 1:** Distributional assumption. Negative Binomial.
- Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
- Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
library(MASS)

FIT3 <- glm.nb(Abundance ~ Deforestation,
               link = "log",
               data = DATA)

summary(FIT3)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.767992   0.166842  10.597 <2e-16
Deforestation -0.036120   0.003747  -9.639 <2e-16
---
(Dispersion parameter for Negative Binomial(1.6398)
 family taken to be 1)

Null deviance: 312.80  on 199  degrees of freedom
Residual deviance: 194.39  on 198  degrees of freedom
AIC: 566.42
```



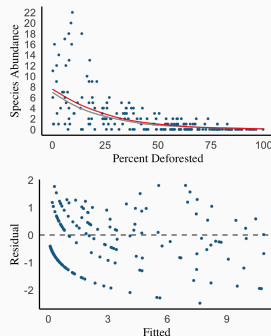
1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
library(MASS)

FIT3 <- glm.nb(Abundance ~ Deforestation,
              link = "log",
              data = DATA)

summary(FIT3)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.767992   0.166842  10.597 <2e-16
Deforestation -0.036120   0.003747  -9.639 <2e-16
---
(Dispersion parameter for Negative Binomial(1.6398)
 family taken to be 1)

Null deviance: 312.80  on 199  degrees of freedom
Residual deviance: 194.39  on 198  degrees of freedom
AIC: 566.42
```



ΔAIC of ~ 52 suggests a big improvement over Poisson. Grey line is Poisson GLM, do you see a big difference?

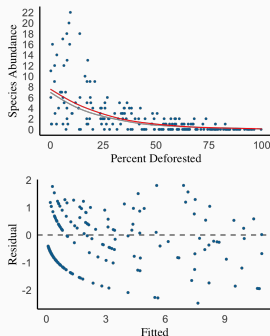
1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
library(MASS)

FIT3 <- glm.nb(Abundance ~ Deforestation,
              link = "log",
              data = DATA)

summary(FIT3)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.767992   0.166842  10.597 <2e-16
Deforestation -0.036120   0.003747  -9.639 <2e-16
---
(Dispersion parameter for Negative Binomial(1.6398)
 family taken to be 1)

Null deviance: 312.80  on 199  degrees of freedom
Residual deviance: 194.39  on 198  degrees of freedom
AIC: 566.42
```



ΔAIC of ~ 52 suggests a big improvement over Poisson. Grey line is Poisson GLM, do you see a big difference? Where is the benefit coming from?

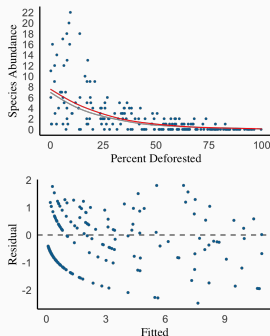
1. **Step 1:** Distributional assumption. Negative Binomial.
2. **Step 2:** Specify η . $\eta = \beta_0 + \beta_1 \times \text{Deforestation}$
3. **Step 3:** Specify the 'link' between the expect. of abund. and η . e^η

```
library(MASS)

FIT3 <- glm.nb(Abundance ~ Deforestation,
              link = "log",
              data = DATA)

summary(FIT3)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.767992   0.166842  10.597 <2e-16
Deforestation -0.036120   0.003747  -9.639 <2e-16
---
(Dispersion parameter for Negative Binomial(1.6398)
 family taken to be 1)

Null deviance: 312.80  on 199  degrees of freedom
Residual deviance: 194.39  on 198  degrees of freedom
AIC: 566.42
```



ΔAIC of ~ 52 suggests a big improvement over Poisson. Grey line is Poisson GLM, do you see a big difference? Where is the benefit coming from? A better description of the system's stochastic component.

References

- Nelder, J.A. & Wedderburn, R.W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135, 370–384.
- Zuur et al. (2009) Chapters 8 & 9