

LINEAR REGRESSION II

GENERALISED LINEAR MODELS

Alejandro Ordonez

Assistant Professor - Department of Bioscience

Section for Ecoinformatics & Biodiversity

Center for Biodiversity Dynamics in a Changing World (BIOCHANGE)

GENERALISED LINEAR REGRESSIONS

WHAT WE WILL TALK ABOUT TODAY

Difference between GLM and GLMMs.

Parts of a GLMM.

Logistic regressions.

- Principles, assessing model performance, model selection

Poisson regressions.

- Principles, assessing model performance, model selection


Any questions?
Ready to start?

GENERALISED LINEAR MODELS (GLMM)

A core assumption of linear models is that the error terms (ε_i) from the fitted models we are normally distributed.

How to deal when this is not the case

- **Transformations** → Useful way to overcome problems with non-normal error terms
- **The best way** → Use a technique for modelling that allows other types of distributions besides normal



Generalised
Linear Models

GENERALISED LINEAR MODELS (GLMM)

TWO IMPORTANT POINTS

GLMM are linear models by definition

- A linear combination of predictors describes the relation between predictors.
- Many elements of simple regressions can be translated to GLMM.

GLMM are parametric models!

- **We assume that error terms follow a specific distribution**
 - That distribution is NOT NORMAL!!!
 - These distributions are of the exponential family.

GENERALISED LINEAR MODELS (GLMM)

ITS' PARTS

Random component

defines the probability distribution of the response variable Y

Systematic component

represent the predictors X

$$y_i = g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon_i$$

Link function

the expression linking the expected value of the response variable Y to predictors X

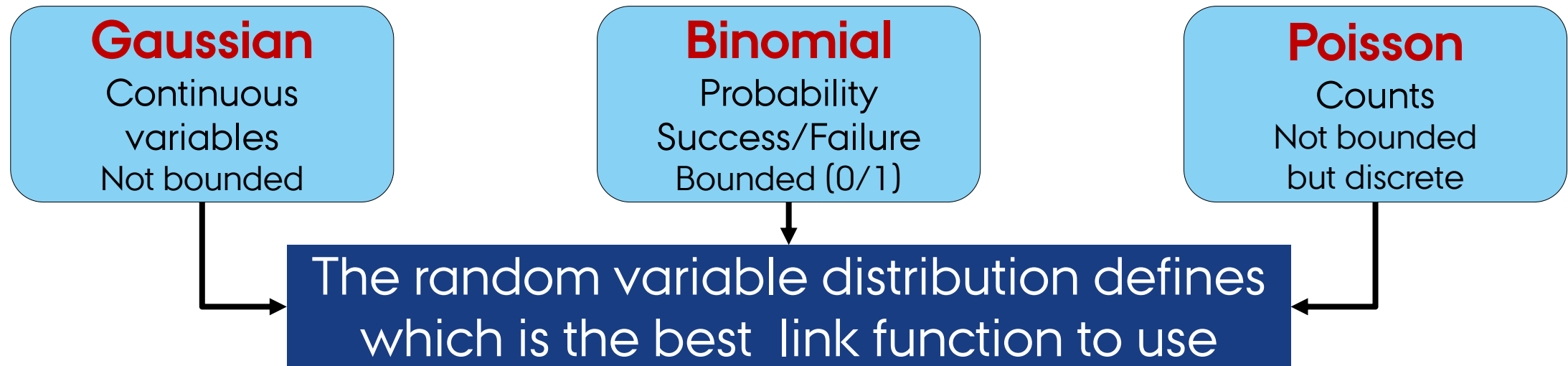
Error

GENERALISED LINEAR MODELS (GLMM)

THE RANDOM PART

As in simple linear regressions, Y is the **random variable** → hence the random part refers to Y

As a **random variable**, it can be described by a probability distribution



GENERALISED LINEAR MODELS (GLMM)

THE LINK FUNCTION

The **Link function** connects the random and systematic component

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

- $g(\mu)$: Link function $\beta_1, \beta_2, \dots, \beta_n$: Parameters to be estimated

The link function is the translator that links the predictors and the response variable

GENERALISED LINEAR MODELS (GLMM)

THE LINK FUNCTION

Link function transform the relation between Y and X so that it can be evaluated as a linear model

continuous
responses

Identity link

$g(\mu) = \mu$
Traditional Linear
model

binary or
probability

Logit link

$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
Logistic regression

count
responses

Log-link

$g(\mu) = \log(\mu)$
Poisson regression

GENERALISED LINEAR MODELS

TYPES WE WILL DISCUSS

Logistic regression

Response variable

- Binary \rightarrow Pres/Abs or Probability

What is modelled:

- Probability that $Y=1$ given X

Error model:

- Binomial

Poisson regression

Response variable

- Counts \rightarrow Discrete numbers

What is modelled:

- Mean Count of Y given X

Error model:

- Poisson

So far so good?

Any questions?

Ready to continue?

GENERALIZED LINEAR MODELS

LOGISTIC REGRESSIONS

We start by calculating the odds that an event occurs as:

$$\frac{\pi(x)}{1 - \pi(x)}$$

$\pi(x)$ is the probability that $y_i = 1$
 $1 - \pi(x)$ is the probability that $y_i = 0$

The Logit transformation

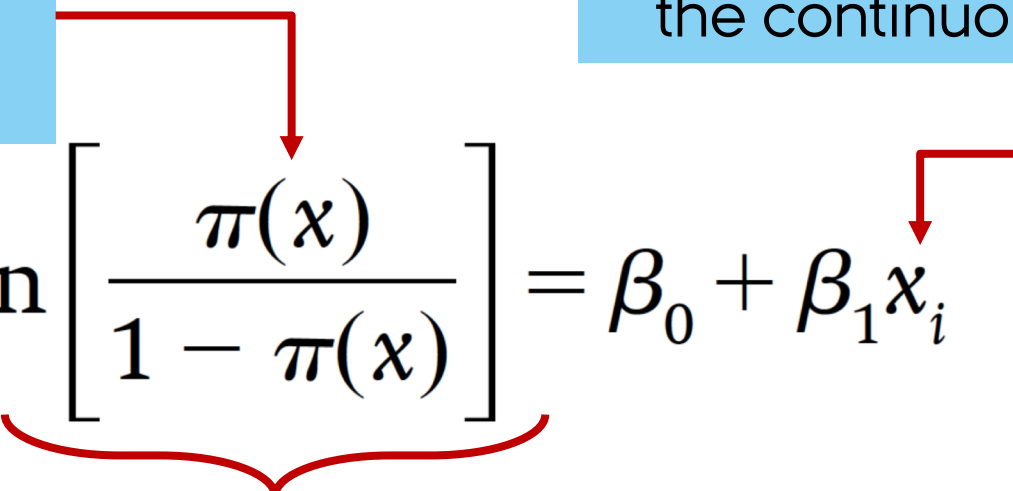
$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right]$$

This is how can we make proportions or probabilities to vary between $-\infty$ and ∞ ?

LOGISTIC REGRESSIONS - ITS' PARTS

The random component is $\pi(x)$ with a binomial probability distribution

The systematic component is the continuous predictor x_i

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x_i$$


The logit link function links the expected value of $\pi(x)$ to the predictor(s)

GENERALIZED LINEAR MODELS

POISSON REGRESSIONS

In biology we usually have discrete units that is the number of something!

Problem-1: Counts follow a **Poisson distribution**

- Which is a violation of normal regressions assumptions.

Problem-2: in a Poisson distribution the mean ($\mu=\lambda$) and variance ($\sigma=\lambda$) are the same

- Which is a violation of independence of variance as the variance increases with the mean!!

POISSON REGRESSIONS - ITS' PARTS

A solution to these two (2) problems is to use a GLM with a Poisson error term and a log link function

The random component is μ with a Poisson distribution

The systematic component is the continuous predictor x_i

$$\log(\mu) = \beta_0 + \beta_1 x_i$$

The log link function links the expected value of μ to the predictor(s)

So far so good?

Any questions?

Ready to continue?

LOGISTIC REGRESSIONS - ASSUMPTIONS

The same as a linear model

- Independence of observations
- Little or no multicollinearity among the independent variables.
- Linearity of independent variables and the link function...
 - However, this means No real Linearity → response is a logistic function of the predictors [which is an exponential function]

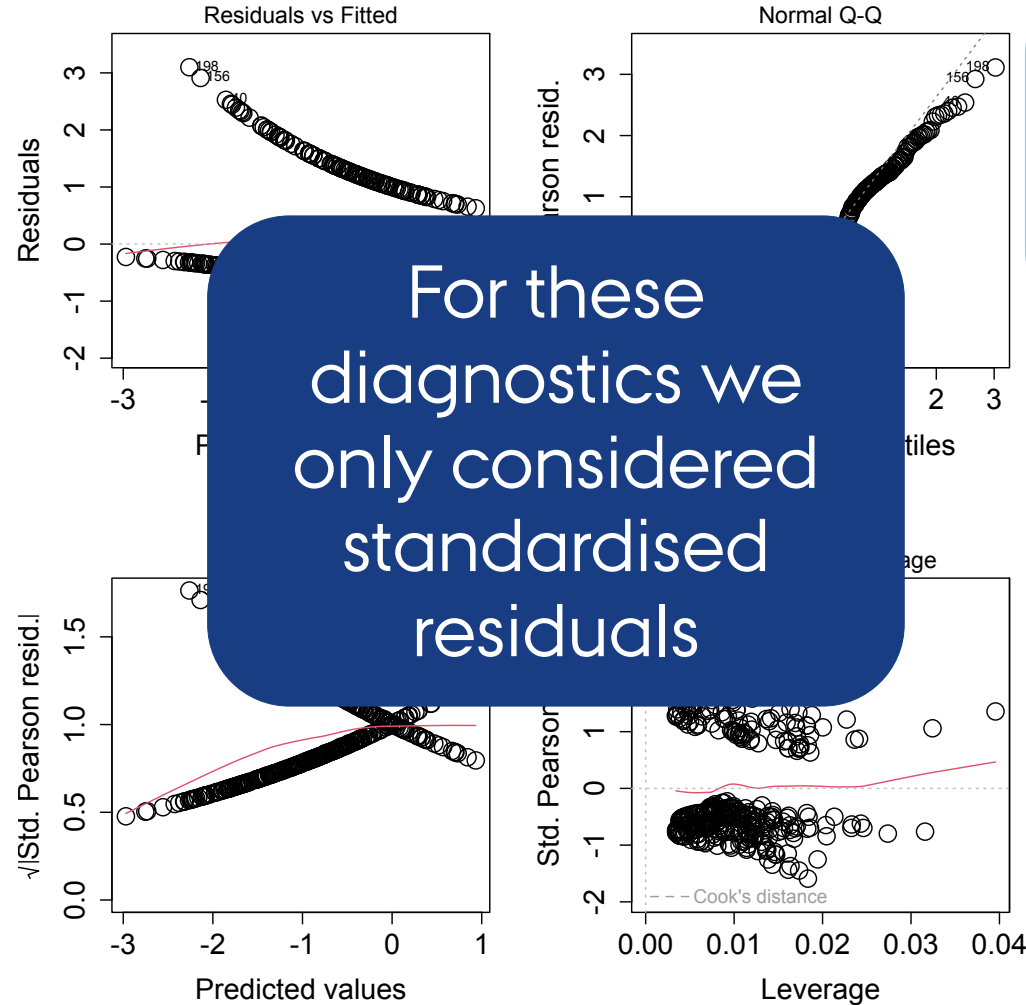
but...

- Errors are **binomial distributed and non-homogenous!**
 - Probability distribution that adequately describes the random component is **binomial**.

LOGISTIC REGRESSIONS - ASSUMPTIONS

Homoscedasticity
Equal Variance.
Don't bother! as it is done in
the raw residuals

Linearity.



Normal Distribution
of Stdz. Residuals.
Don't bother!

Leverage &
outliers.

POISSON REGRESSIONS - ASSUMPTIONS

The same as a linear model

- Linearity → additive predictors
- Independence of observations
- Little or no multicollinearity among the independent variables.

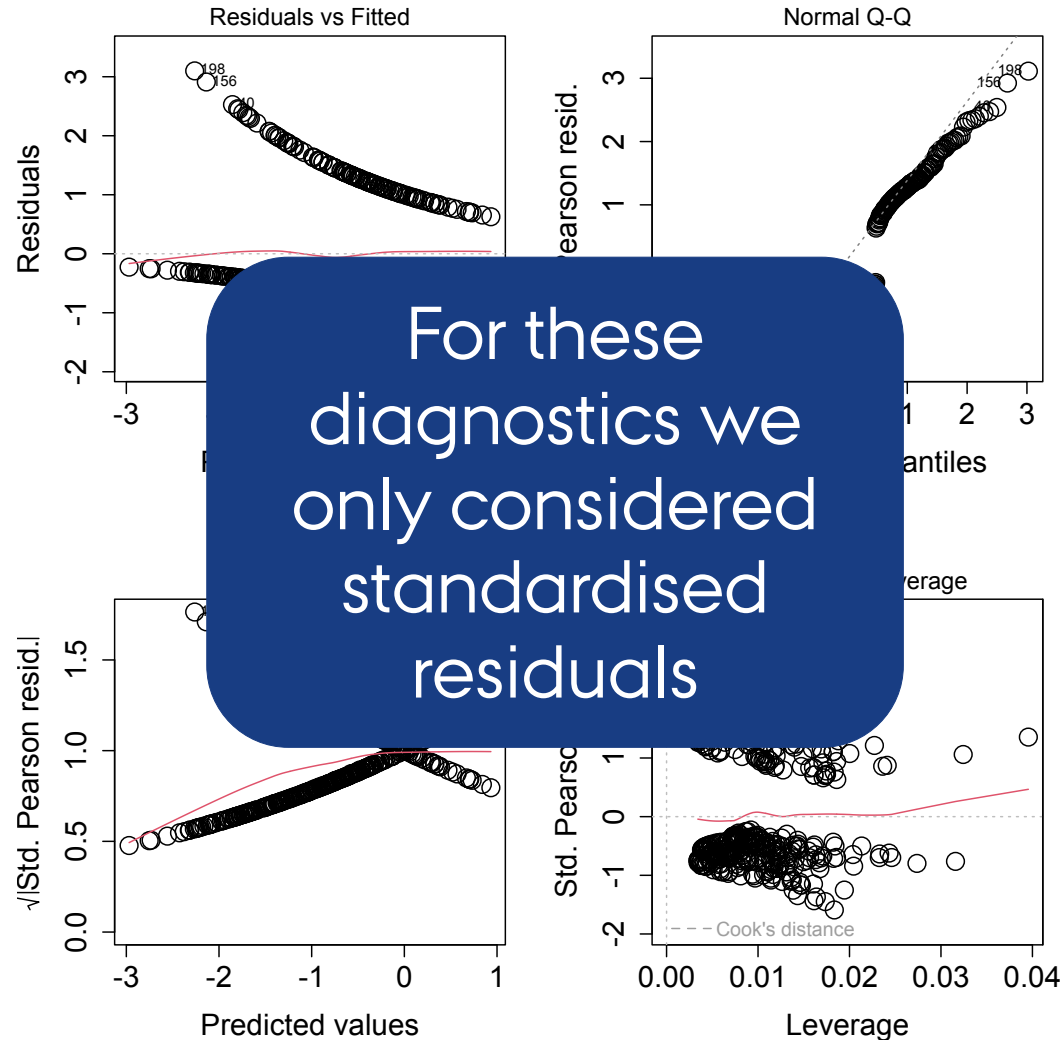
but...

- There is now homogeneity of variances
 - In a **Poisson** model the mean $\mu = \sigma$
- Errors are **Poisson distributed**.
 - Probability distribution that adequately describes the random component is **Poisson**.

POISSON REGRESSIONS - ASSUMPTIONS

Homoscedasticity
Equal Variance.
Don't bother! as it is done in
the raw residuals

Linearity.



Normal Distribution
of Stdz. Residuals.
Don't bother!

Leverage &
outliers.

STANDARDISED RESIDUALS

What are Raw residuals?

- Provides an indication of the **absolute error** of the prediction.
- Difference between the observed and predicted value.
- Some residuals will be negative and some will be positive.

What are Standardised residuals?

- Measure of the **strength** of the difference between observed and expected values.
- Raw residual divided by an estimate of the standard deviation of the residuals.
 - Standardizing as we discussed in the data exploration lecture.

STANDARDISED RESIDUALS

Pearson Residuals.

Scaling by the variance.

$$\hat{\varepsilon}_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(Y_i)}}$$

Avoid problems with changing variances

Deviance residuals.

Scaling by the deviance.

$$\hat{\varepsilon}_i^D = \frac{ABS(y_i - \mu_i)}{d_i}$$

Better for model checking as these behave similar to residuals from a Gaussian linear regression

So far so good?

Any questions?

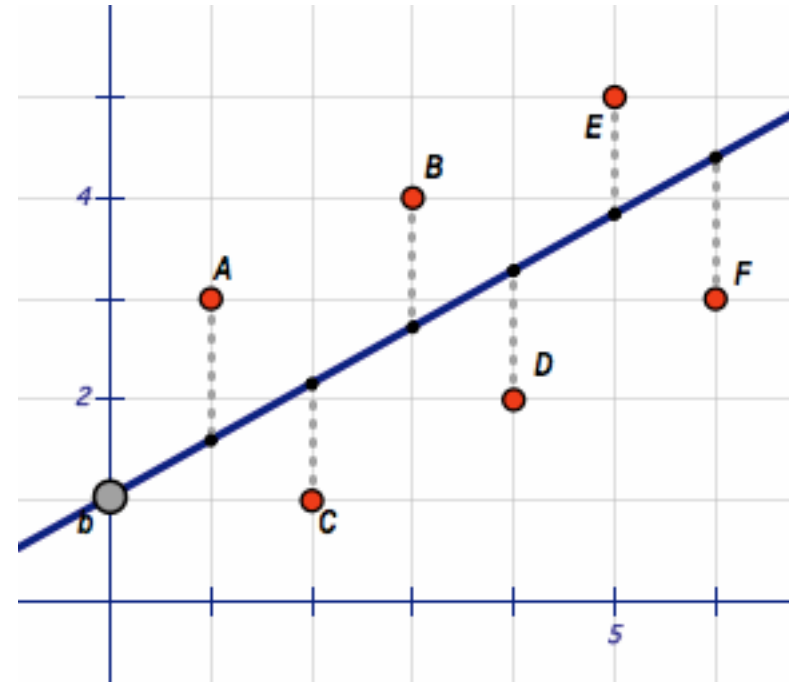
Ready to continue?

HOW THE PARAMETERS ARE ESTIMATED

GLMMS USE MAXIMUM LIKELIHOOD!

Linear regressions:

- Parameter estimates are determined using an Ordinary Least Squares method (OLS)
- The goal in OLS is find a parameter that minimises the sum of squared residuals (**Sum of squares**)



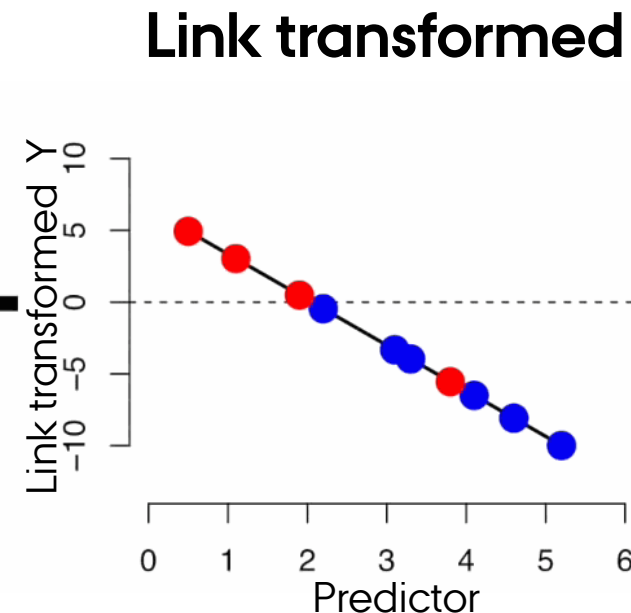
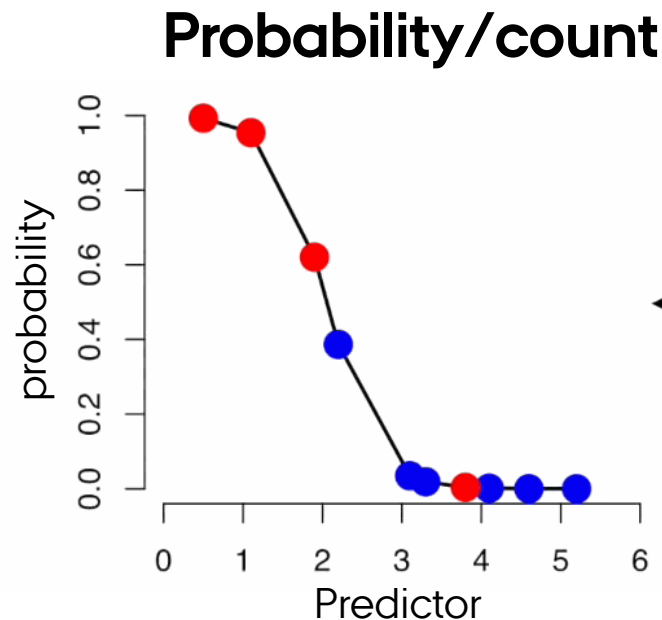
HOW THE PARAMETERS ARE ESTIMATED

GLMMS USE MAXIMUM LIKELIHOOD!

Generalized Linear models:

- Parameter estimates are determined using an Maximum Likelihood
- *Why?* Because here residuals can be $-\infty$ or ∞ so **OLS does not work!**

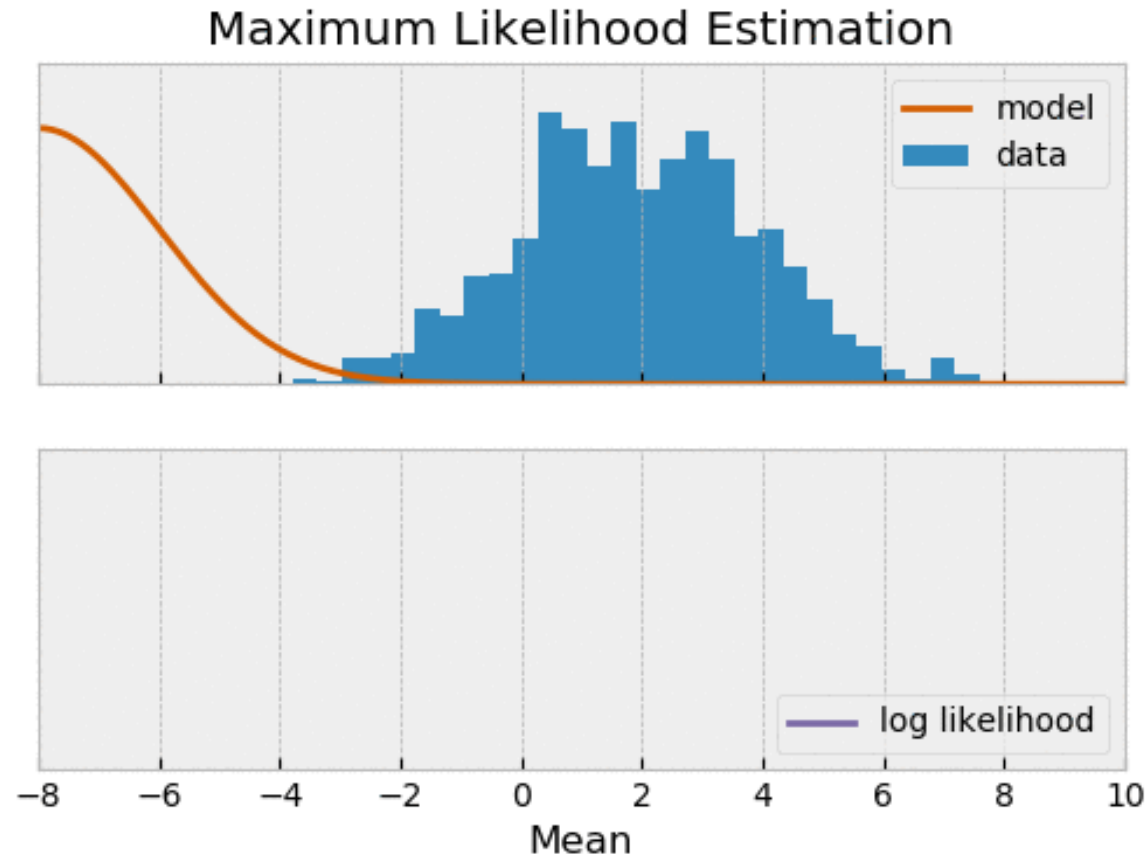
The likelihood is the multiplication of the estimated probabilities for a given reset of parameters



What is changing here is the β_i , which makes the line change.

HOW THE PARAMETERS ARE ESTIMATED

GLMMS USE MAXIMUM LIKELIHOOD!



It is simpler to think of this in finding the best parameters for the likelihood function

For a logistic regression:

$$L(\beta_0, \dots, \beta_n) = \prod_{j=1}^n p(x_j)^{y_j} * (1 - p(x_j))^{1-y_j}$$

For a Poisson regression:

$$L(\lambda, \beta_0, \dots, \beta_n) = \prod_{j=1}^n \frac{\lambda^{x_j} * e^{-\lambda}}{x_j!}$$

So far so good?

Any questions?

Ready to continue?

LOGISTIC REGRESSIONS

ODDS RATIO AND REG. COEFFICIENTS

Odds ratios show the changes in the odds of an outcome for an increase in one unit of X.

$$\text{odds ratio} = \frac{\ln \left[\frac{\pi(x_i)}{1 - \pi(x_i)} \right]}{\ln \left[\frac{\pi(x_{i+1})}{1 - \pi(x_{i+1})} \right]} = e^{\beta_1}$$

Odds ratio measure of how the changes of having a success changes with a one unit change the predictor.

As these β_i have a range of variation (90%-CI) the Odd ratio also does

If one is included in this interval
The variable has no effect!

LOGISTIC/POISSON REGRESSIONS

IMPORTANCE OF PREDICTORS

The H_0 of interest when assessing the Importance of a predictor in a logistic regression model is that $\beta_j=0$ (**NO EFFECT**)

How to test this H_0 :

- **The Wald statistic:** a version of a t -test that works for GLMMs
 - Only use if you have large sample size (above 50 obs).
- **G^2 tests:** this is also called the likelihood ratio χ^2 statistic.

LOGISTIC/POISSON REGRESSIONS

IMPORTANCE OF PREDICTORS

Here you can ask two questions:

- Is the coefficient different from zero (0) $\rightarrow \beta_j=0$ (NO EFFECT)
 - **The Wald statistic gives the answer to this question**
- Does the model perform better by having this variable?
 - **G^2 tests gives the answer to this question**

LOGISTIC/POISSON REGRESSIONS

THE WALD-T STATISTIC

This is a ML version of a **t-test**, based on the relation between
Parameter estimate (b_1) vs Parameter standard error (S_{b_1})

$$Wald_t = \frac{b_1}{S_{b_1}}$$

The Wald statistic is traditionally compared to the standard normal z-distribution.

LOGISTIC/POISSON REGRESSIONS

G^2 TESTS

The significance of β_1 based on the difference in likelihood of two models

A full model

$$g(x) = \beta_0 + \beta_1 x_i$$

A reduced model

$$g(x) = \beta_0$$

G^2 tests is the contrast of likelihood of the two models
It is a likelihood ratio statistic (Λ)

Methodologically like a F-ratio test for linear models

LOGISTIC/POISSON REGRESSIONS

G² TESTS

The likelihood ratio statistic (Λ) is the contrasts the likelihood of two models

A full model

$$g(x) = \beta_0 + \beta_1 x_i$$

A reduced model

$$g(x) = \beta_0$$

The closer Λ is to one, β_1 contributes little to the full model fit

Note these are Likelihood but R gives you Log-likelihoods!

$$\Lambda = \frac{\text{Likelihood}_{\text{Reduced}}}{\text{Likelihood}_{\text{Full}}}$$

$$G^2 = -2\ln(\Lambda) \sim \chi^2_{df=1}$$

LOGISTIC/POISSON REGRESSIONS

HOW GOOD IS MY MODEL?

GLMM have no R^2 or adjusted R^2 → **How to define model Fit?**

- **Deviance** measures the **unexplained variation** for a given model and therefore is a measure of goodness-of-fit

```
> summary(fit)
Call:
glm(formula = num_awards ~ prog + math, family = "poisson",
data = mydata)

...
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 287.67 on 199 degrees of freedom
Residual deviance: 203.45 on 197 degrees of freedom
AIC: 385.51
Number of Fisher Scoring iterations: 6
```

Comparing the
Residual deviance
[current vs saturated model]
and
Null deviance
[current vs no predictor model]
Provides a measurement of
goodness-of-fit similar to an R^2 .

So far so good?

Any questions?

Ready to continue?

POISSON REGRESSIONS

THE OVER DISPERSION PROBLEM

A key assumption of when modeling count data

The mean equals the variance.

BUT, In many cases variance > mean

This is called overdispersion.

Why can this happen?:

Omission of important factors

Why this is a problem?:

- Regression coefficients too small
- Inflated Type I error probabilities

POISSON REGRESSIONS

THE OVER DISPERSION PROBLEM

How to deal with the overdispersion problem:

- Correct the parameters standard errors

multiply by $\sqrt{\chi^2 / df}$ → **the easy way out**

Quasi-models are already
implemented in R
`family="quasipoisson"`
`family="quasibinomial"`

- Use models of the quasi-likelihood families

Dispersion parameter is estimated from the data rather than restricted by a Poisson distribution → **the best approach**

POISSON REGRESSIONS

ZERO INFLATION

How to deal with the overdispersion problem?:

- Use a **negative binomial model** → loosens the restrictive assumption that the variance is equal to the mean.
- Poisson regressions are a generalization of **negative binomial models**.

BEYOND POISSON REGRESSIONS

DEALING WITH TOO MANY ZEROS

Zero Inflated Poisson models assume that

- Some zeros occurred by a Poisson process → **think true absences.**
- Others not eligible to have the event occur → **think sampling error.**

This means there are **two processes at work!** → The tricky part is that either process can result in a 0 count

Zero Inflated models estimate "true zeros" and "excess zeros" simultaneously.

So far so good?

Any questions?

Ready to end?

SUMMARY

$$y_i = g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon_i$$

GLMM's are linear models by definition

Consider the right model:

- **Probabilities** [Logistic] vs. **Counts** [Poisson] vs. **Many zeros** [Negative binomial]

Always **test the assumptions** to define

- If the linear model you build is “statistically correct”
- GLMM's assumptions are the same as GLM → diff is the distribution of residuals

Remember that the results of a GLM need to be “**back transformed**” → they are given in the loglink function units

Always **check for overdispersion** by asking : *is the residual deviance much greater than the residual degrees of freedom?*



AARHUS
UNIVERSITY