Generalized Linear models

Gavin L. Simpson

January 17, 2018

Generalized linear models

Generalized linear models

Generalised linear models (GLMs) are a synthesis and extension of linear regression plus Poisson, logistic and other regression models

GLMs extend the types of data and error distributions that can be modelled beyond the Gaussian data of linear regression

With GLMs we can model count data, binary/presence absence data, and concentration data where the response variable is not continuous.

Such data have different mean-variance relationships and we would not expect errors to be Gaussian.

Typical uses of GLMs in ecology are

- Poisson GLM for count data
- Logistic GLM for presence absence data
- Gamma GLM for non-negative or positive continuous data

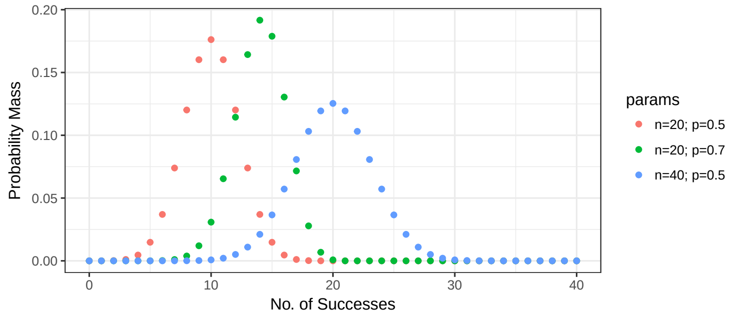
GLMs can handle many problems that appear non-linear

Not necessary to transform data as this is handled as part of the GLM process

Binomial distribution

- For a fixed number of trials (n),
- fixed probability of “success” (p), &
- two outcomes per trial (heads or tails)

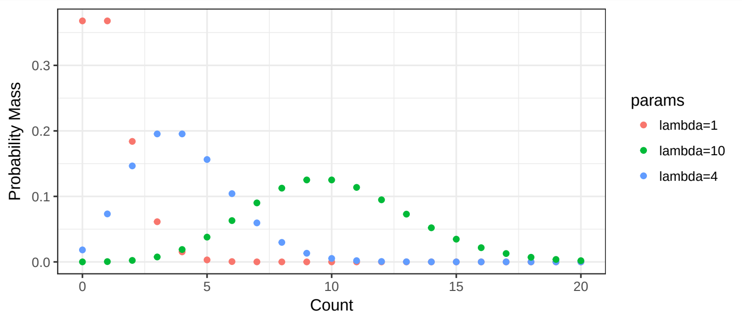
Flip a coin 10 times with p = 0.7, the probability of 7 heads is ~ $Bin\left(n = 10, p = 0.7\right)$, $\approx$ 0.27



Poisson distribution

The Poisson gives the distribution of the number of “things” (individuals, events, counts) in a given sampling interval/effort if each event is independent.

Has a single parameter λ the average density or arrival rate



The structure of a GLM

A GLM consists of three components, chosen/specified by the user

1. A random component, specifying the conditional distribution of of the response Y_i given the values of the explanatory data. Error Function
2. A Linear Predictor η — the linear function of regressors

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

The X_{ij} are prescribed functions of the explanatory variables and can be transformed variables, dummy variables, polynomial terms, interactions etc.

3. A smooth and invertible Link Function $g(\cdot)$, which transforms the expectation of the response $\mu_i \equiv E(Y_i)$ to the linear predictor

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

As $g(\cdot)$ is invertible, we can write

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

Conditional distribution of y_i

Originally GLMs were specified for error distribution functions belonging to the *exponential family* of probability distributions

- Continuous probability distributions
    - Gaussian (or normal distribution; used in linear regression)
    - Weibull
    - Gamma (data with constant coefficient of variation)
    - Exponential (time to death, survival analysis)
    - Chi-square
    - Inverse-Gaussian
- Discrete probability distributions
    - Poisson (count data)
    - Binomial (0/1 data, counts from a total)
    - Multinomial

Choice depends on range of Y_i and on the relationship between the variance and the expectation of Y_i — *mean-variance relationship*

Conditional distribution of y_i

Characteristics of common GLM probability distributions

	Canonical Link	Range of Y_i	Variance function
Gaussian	Identity	$(-\infty, +\infty)$	ϕ
Poisson	Log	$0, 1, 2, \ldots, \infty$	μ_i
Binomial	Logit	$\frac{0,1,\ldots,n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Gamma	Inverse	$(0, \infty)$	$\phi\mu_i^2$
Inverse-Gaussian	Inverse-square	$(0, \infty)$	$\phi\mu_i^3$

ϕ is the dispersion parameter; μ_i is the expectation of Y_i. In the binomial family, n_i is the number of trials

Ecologically-relevant probability distributions

Gaussian distribution is rarely adequate in (palaoe)ecology; GLMs offer ecologically meaningful alternatives

- Poisson — counts; integers, non-negative, variance increases with mean
- Binomial — observed proportions from a total; integers, non-negative, bounded at 0 and 1, variance largest at $\pi = 0.5$
- Binomial — presence absence data; discrete values, 0 and 1, models probability of success
- Gamma — concentrations; non-negative (strictly positive with log link) real values, variance increases with mean, many zero values and some high values

Notation

Old notation

Wrote linear model as

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_j x_{ij} + \varepsilon_i$$

And assumed

$$y_i | X \sim \text{Normal}(0, \sigma^2)$$

This doesn’t work out the same for GLMs — we don’t have residuals in the linear predictor

Sampling variation comes from the response distribution

New notation

Rewrite linear model as

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\eta_i &= \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_j x_{ij}
\end{aligned}
$$

This now matches the general form for the GLM

$$
\begin{aligned}
y_i &\sim \text{EF}(\mu_i, \boldsymbol{\theta}) \\
g(\mu_i) &= \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_j x_{ij}
\end{aligned}
$$

New notation

Binomial GLM

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n, p_i) \\
\text{logit}(p_i) &= \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_j x_{ij}
\end{aligned}
$$

Poisson GLM

$$
\begin{aligned}
y_i &\sim \text{Poisson}(\lambda_i) \\
\log(\lambda_i) &= \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_j x_{ij}
\end{aligned}
$$

Examples

Logistic regression — *Darlingtonia*

Timed censuses at 42 randomly-chosen leaves of the cobra lily (*Darlingtonia californica*)

- Recorded number of wasp visits at 10 of the 42 leaves
- Test hypothesis that the probability of visitation is related to leaf height
- Response is dichotomous variable (0/1)
- A suitable model is the logistic model

$$\pi = \frac{e^{\beta_0 + \beta_i X}}{1 + e^{\beta_0 + \beta_1 X_i}}$$

- The logit transformation produces

$$\log_e \left(\frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 X_i$$

- This is the logistic regression and it is a special case of the GLM, with a binomial error distribution and the logit link function

Logistic regression — *Darlingtonia*

$$\log_e\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_i$$

- β_0 is a type of intercept; determines the probability of success $(Y_i = 1)$ π where X = 0
- If $\beta_0 = 0$ then $\pi = 0.5$
- β_1 is similar to the slope and determines how steeply the fitted logistic curve rises to the maximum value of $\pi = 1$
- Together, β_0 and β_1 specify the range of the X variable over which most of the rise occurs and determine how quickly the probability rises from 0 to 1
- Estimate the model parameters using Maximum Likelihood; find parameter values that make the observed data most probable

Logistic regression — *Darlingtonia*

```
> mod <- glm(visited ~ leafHeight, data = wasp, family = binomial)
> mod
```

```

Call:  glm(formula = visited ~ leafHeight, family = binomial, data = wasp)

Coefficients:
(Intercept)   leafHeight  
    -7.2930       0.1154  

Degrees of Freedom: 41 Total (i.e. Null);  40 Residual
Null Deviance:      46.11 
Residual Deviance: 26.96    AIC: 30.96
```

```
> plogis(coef(mod))
```

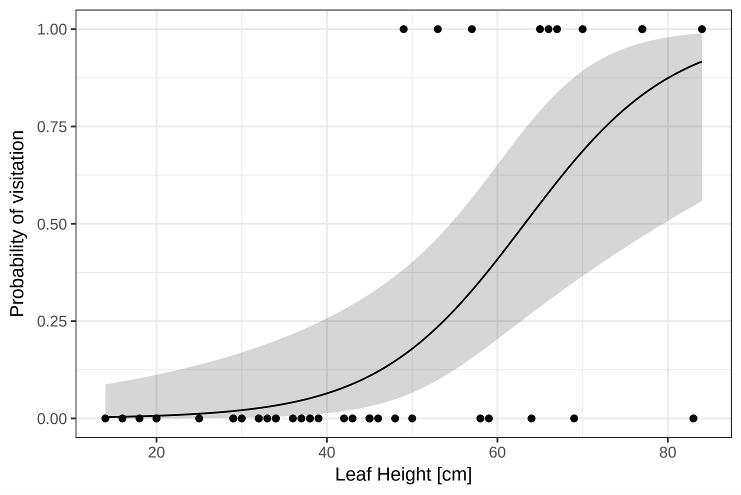
```
 (Intercept)   leafHeight 
0.0006798556 0.5288181121 
```

Logistic regression — *Darlingtonia*

```
> summary(mod)
```

```
Call:
glm(formula = visited ~ leafHeight, family = binomial, data = wasp)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.18274  -0.46820  -0.23897  -0.08519   1.90573

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.29295    2.16081  -3.375 0.000738 ***
leafHeight   0.11540    0.03655   3.158 0.001591 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.105  on 41  degrees of freedom
Residual deviance: 26.963  on 40  degrees of freedom
AIC: 30.963

Number of Fisher Scoring iterations: 6
```

Logistic regression — *Darlingtonia*



Wald statistics

z values are Wald statistics, which under the null hypothesis follow *assymptotically* a standard normal distribution

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.2930	2.1608	-3.3751	0.0007
leafHeight	0.1154	0.0365	3.1575	0.0016

Tests the null hypothesis that $\beta_i = 0$

$$z = \hat{\beta}_i / \mathrm{SE}(\hat{\beta}_i)$$

Deviance

- In least squares we have the residual sum of squares as the measure of lack of fitted
- In GLMs, deviance plays the same role
- Deviance is defined as twice the log likelihood of the observed data under the current model
- Deviance is defined relative to an arbitrary constant — only differences of deviances have any meaning
- Differences in deviances are also known as ratios of likelihoods
- An alternative to the Wald tests are deviance ratio or likelihood ratio tests

$$F = \frac{(D_a - D_b)/(\mathrm{df}_a - \mathrm{df}_b)}{D_b/\mathrm{df}_b}$$

- D_j deviance of model, where we test if model A is a significant improvement over model B; df_k are the degrees of freedom of the respective model

A Gamma GLM — simple age-depth modelling

Radiocarbon age estimates from depths within a peat bog (Brew & Maddy, 1995, QRA Technical Guide 5)

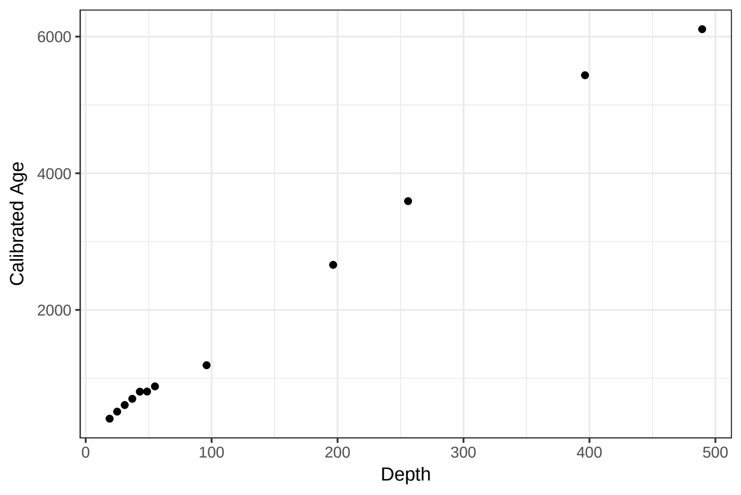
Estimate accumulation rate; assumption here is linear accumulation

Uncertainty or error is greater at depth; mean variance relationship

Fit mid-depth & mid-calibrated age points

Sample	upperDepth	lowerDepth	ageBP	ageError	calUpper	calLower	midDepth	calMid
SRR-4556	20	22.0	355	35	509	307	19.00	408.0
SRR-4557	26	28.0	465	35	542	480	25.00	511.0
SRR-4558	32	34.0	635	35	671	545	31.00	608.0
SRR-4559	38	40.0	740	35	732	666	37.00	699.0
SRR-4560	44	46.0	865	35	916	691	43.00	803.5
SRR-4561	50	52.5	870	35	918	692	48.75	805.0

A Gamma GLM — simple age-depth modelling

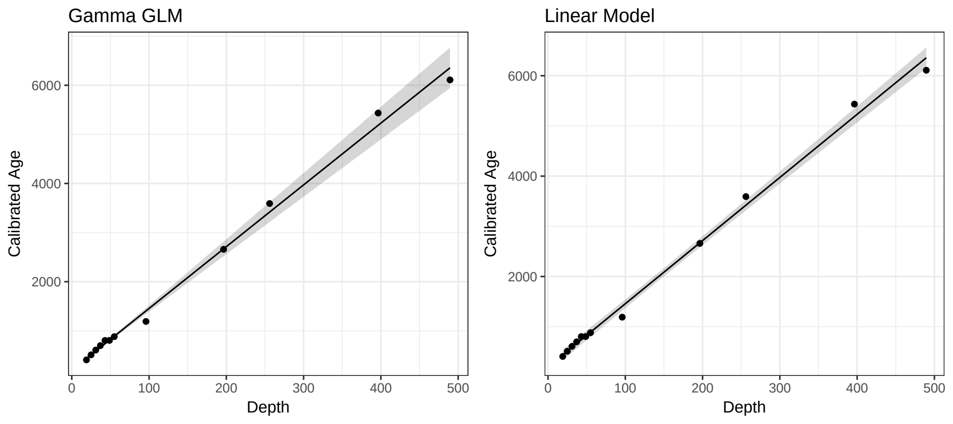


A Gamma GLM — simple age-depth modelling

```
> mod <- glm(calMid ~ midDepth, data = maddy, family = Gamma(link = "identity"))
> summary(mod)
```

```
Call:
glm(formula = calMid ~ midDepth, family = Gamma(link = "identity"), 
    data = maddy)

Deviance Residuals: 
      Min         1Q     Median         3Q        Max  
-0.161184  -0.016734  -0.002595   0.048033   0.085943  

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 197.2909    22.5603   8.745 5.35e-06 ***
midDepth     12.5799     0.4543  27.693 8.74e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.004612561)

    Null deviance: 10.439047  on 11  degrees of freedom
Residual deviance:  0.048316  on 10  degrees of freedom
AIC: 145.57

Number of Fisher Scoring iterations: 4
```

A Gamma GLM — simple age-depth modelling



Re-use