Generalized linear models (GLMs)

Outline

GLM – what and why?

GLM in three steps

General Assumptions of GLMs

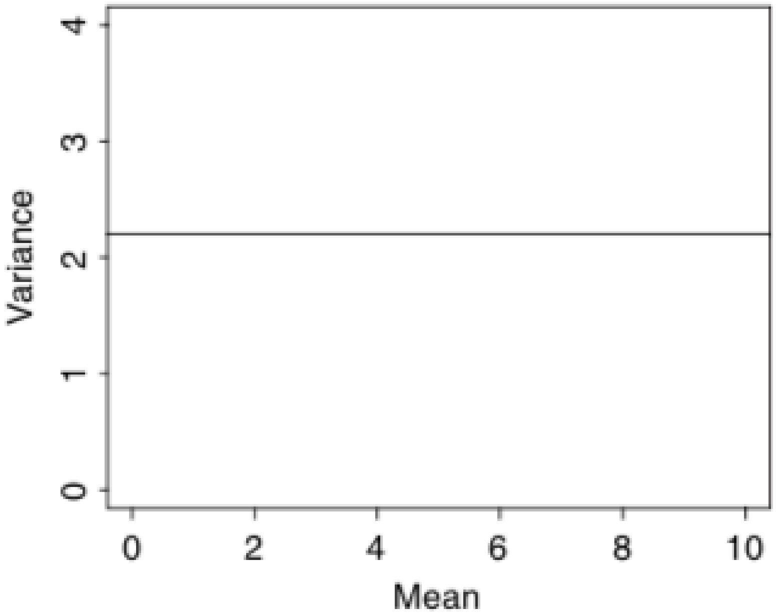
Examples

What is a GLM and why use it?

Why use GLMs

With linear modelling (lm or lme), central assumption is that variance is constant (flat line)

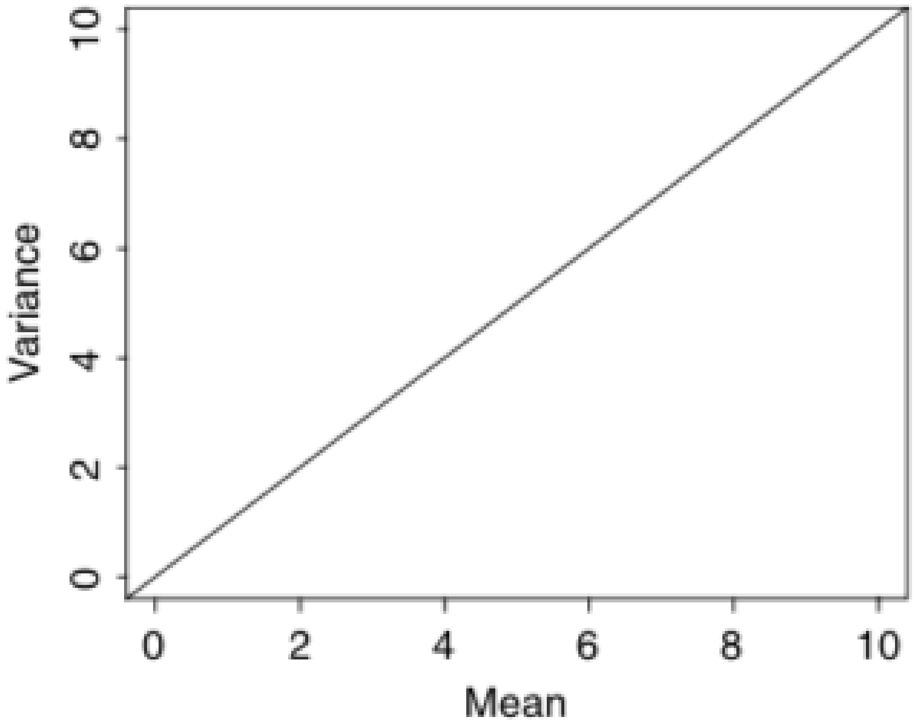
Homogeneity of variance=variance is constant



But in many practical applications, variance is **not** constant, so this assumption is invalid

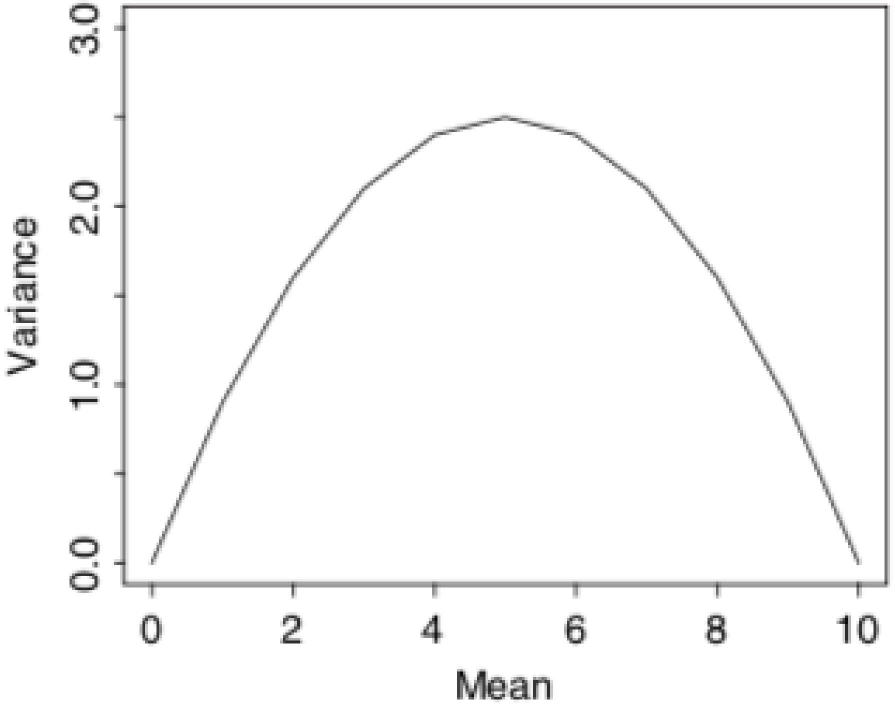
But…with **count data** (often zero-inflated)

Variance often increases
linearly with the mean



But…with **proportion data** (e.g. success vs failure),

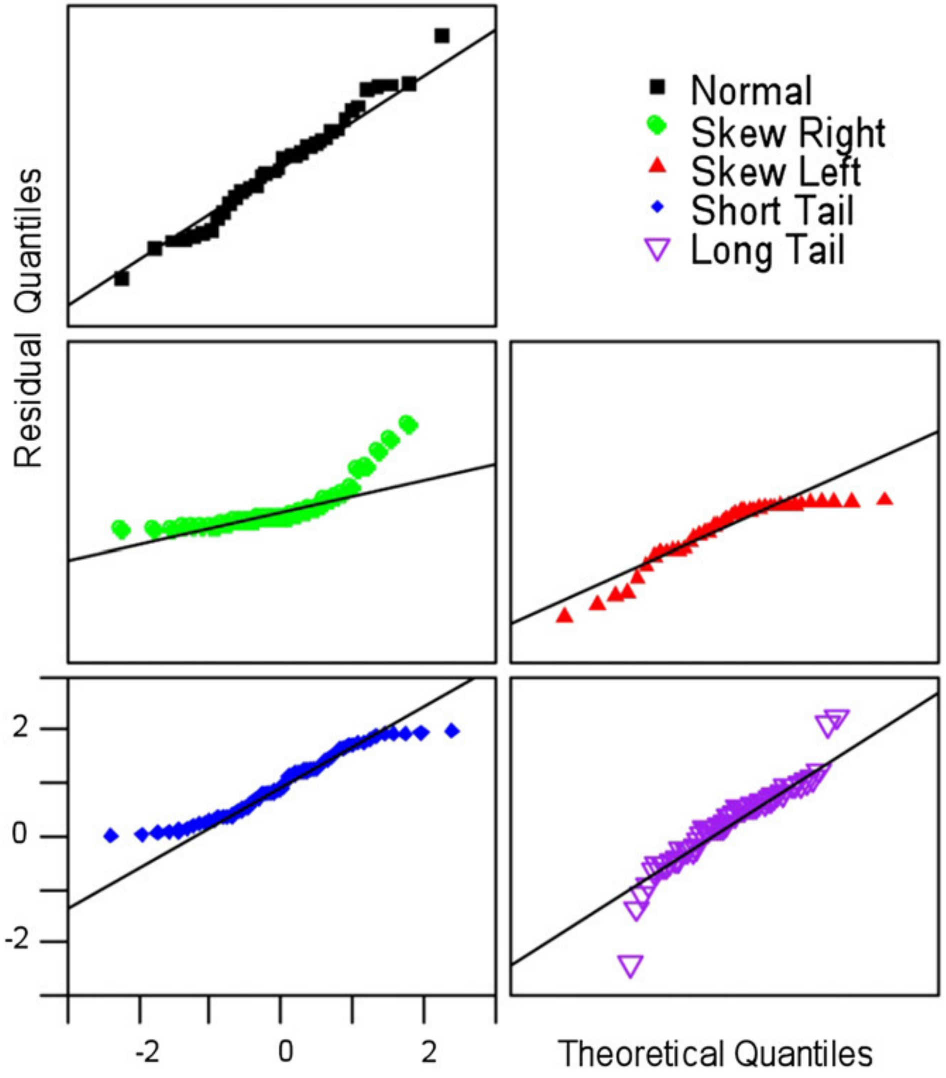
variance can be an inverted U-shaped function of the mean



Homogenity

Often fails with Count data & Proportion data

…..and often the distribution of residuals is not normal



GLMs are an extension of a regular LM

- GLMs extend the linear modeling framework to variables that are not normally distributed and don’t meet homogeneity of variance
- GLMS can have response variables with any distribution in the “exponential family”
- **What does Generalized mean?**
  - A glm is a flexible generalization of ordinary linear regression
  - Still assuming a linear relationship
  - If not linear—try a GAM instead.

You have already done a GLM!

- Linear regression is just a special case of a GLM
- **Probability distribution:** Gaussian (Normal).
- **Link function:** identity

- **GLMs** are just extensions of linear modeling = a non-Gaussian distribution for the response variable is used, and the relationship (or link) between the response variable and the explanatory variables may be different.

When to use a GLM instead of LM

- Response variable has a distribution other than the normal (Gaussian) distribution
- Transformation of the data is undesirable or impossible.
- Examples:
  - Binary response data (1 or 0, dead or alive)
  - Data that are counts (number of offspring, or leaves).
  - Continuous data with non-normal distribution (biomass of plants)

How does a GLM work?

You transform the r'ship & not the response-variable.

- The model still includes a *linear predictor* **But** the **predicted *Y*-values are not modeled directly**
- Instead, it models the relationship between the predictors and the expected values through the transformed linear predictor (**link function**).
- Non-normal distributions of errors and unequal error variances are ok because specified by the link function
- Predicted values are derived using the inverse of the link function.
- Uses maximum likelihood to estimate parameters
- Uses log-likelihood ratio tests to test parameters
- **R Code:** fit models using **glm()**

GLM: three main steps

Random component: choosing a distribution for the response variable

type of data your response variable is

Systematic component: defining the systematic part in terms of covariates (we will skip this step!)

Link function: specifying the relationship (or: link) between the expected value of the response variable (random component) and the predictors (systematic component)

how the 2 pieces are connected together.

Step 1: Choosing a distribution

There are many probability distributions to describe different types of data.

A distribution provides the probability of observing each possible outcome of an experiment or survey

Distributions can be **discrete** (only include integers), **continuous** (including fractions) and be limited to a given range or domain (*e.g.*, 0 to 1, 0 to +∞, −∞ to +∞).

All distributions can be broken down into **parameters** that dictate their shapes (*e.g.*, μ and σ2 for the Normal)

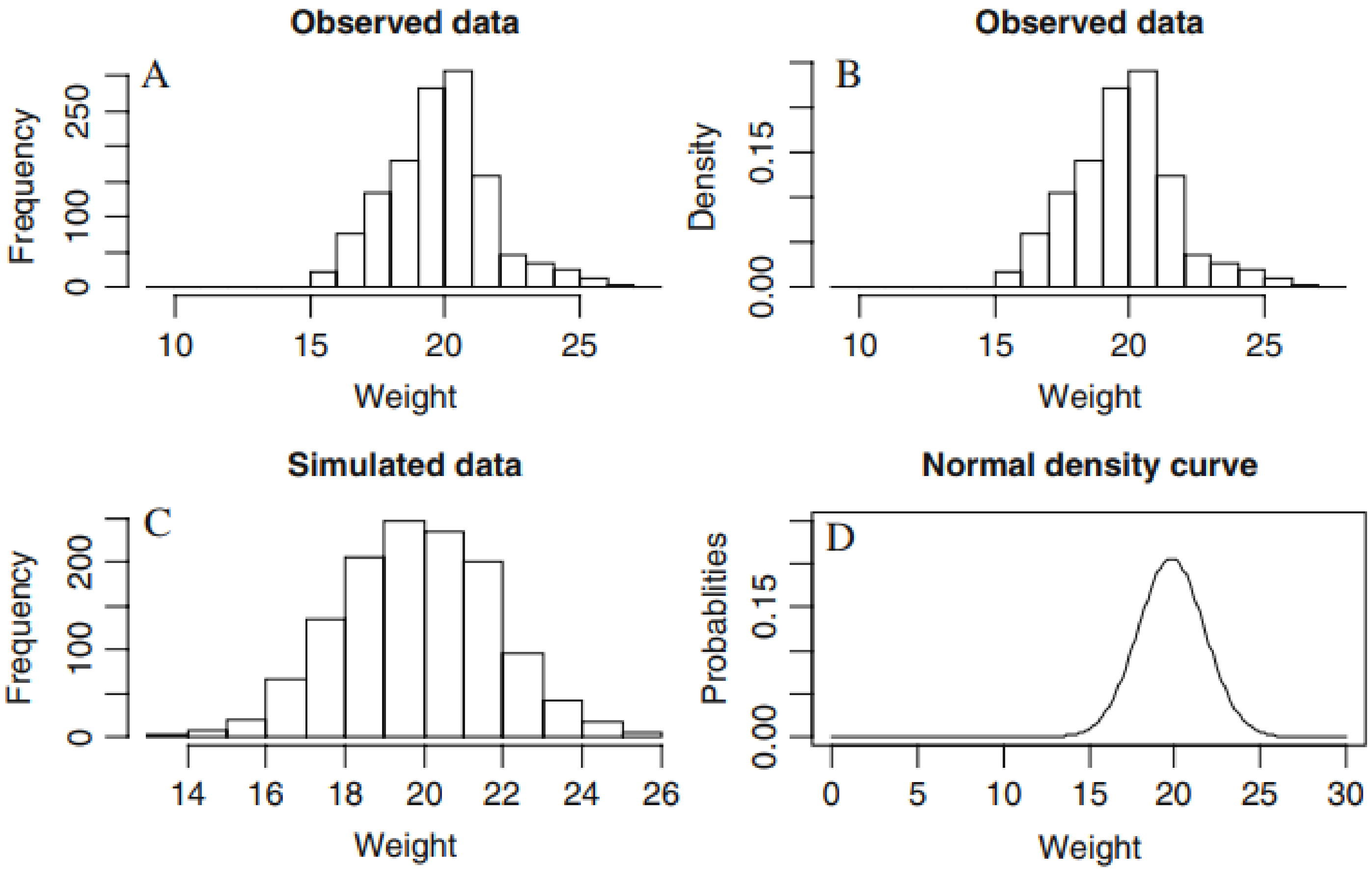
Step 1: Choosing a distribution

- Normal (Gaussian)
- Binomial/Bernoulli
- Poisson
- Negative binomial
- Gamma
- Tweedie family

Choose the distribution B4 you start.

Normal (gaussian) distribution

- Use for continuous data (not restricted to integers)
- range from −∞ to ∞ (your data might not range this much!)
- one of the most common distributions
- 2 parameters: μ = mean and σ= variance



Bernoulli and binomial distributions

takes 2 possible outcome.

If you observed more than one, use binomial distribution

- Use for discrete integer data in which one observation includes the number of successes out of a total number of trials (maximum number of successes)
- 2 parameters: p = success probability and n = trial sample size; variance= np(1-p)
- n is the upper limit, which differentiates the Binomial from the Poisson distribution
- n=1 Bernoulli distribution

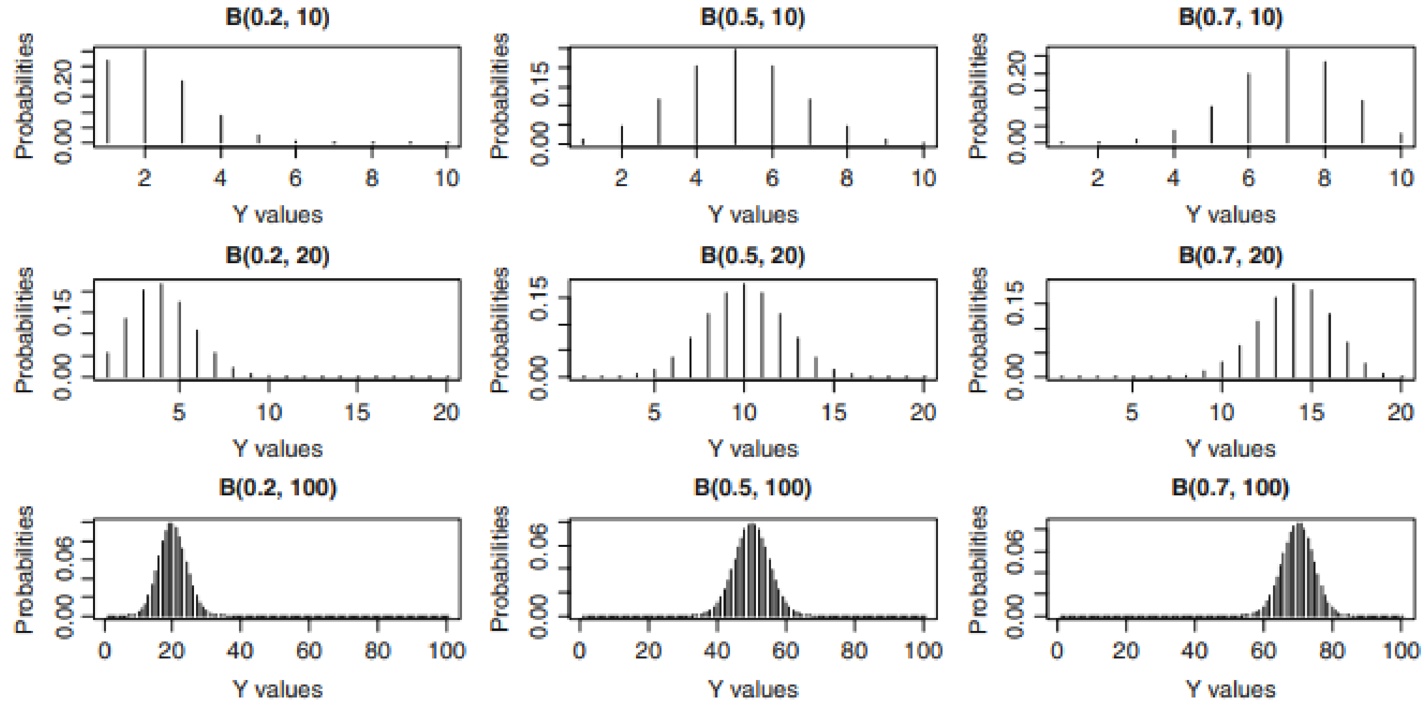


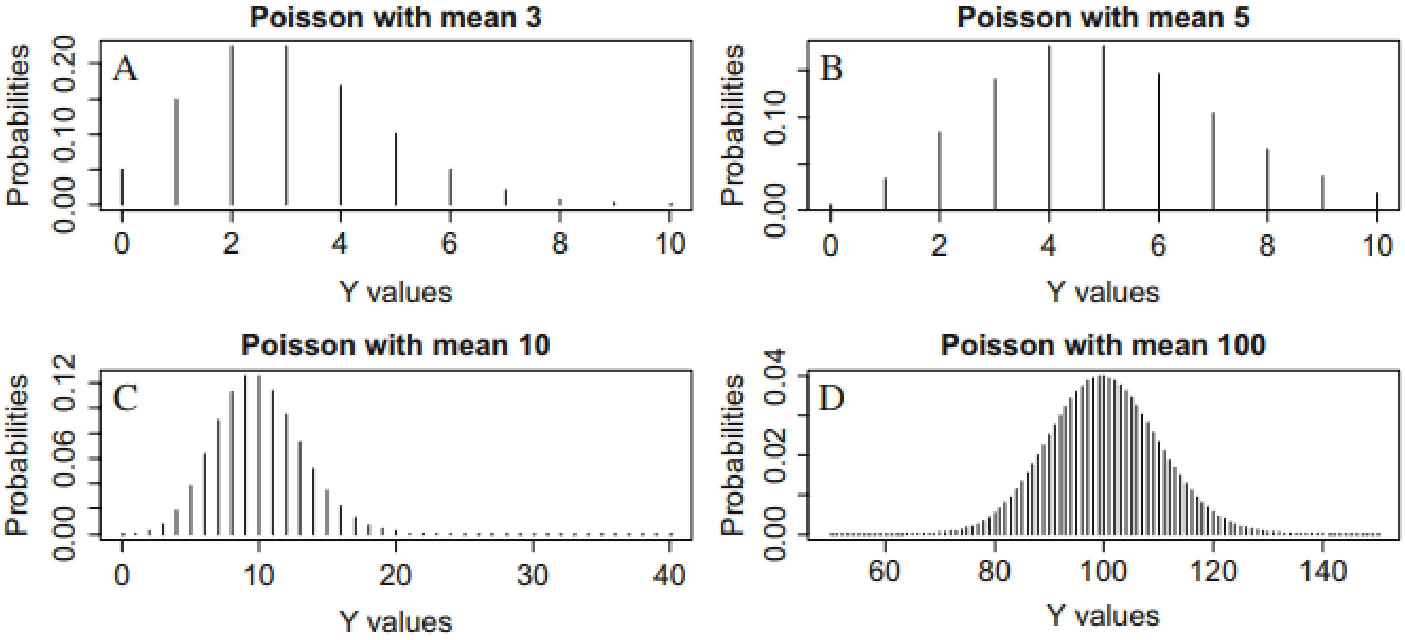
Fig. 8.5 Binomial density curves B(π, N) for various values of π (namely 0.2, 0.5, and 0.7) and N (namely 10, 20, and 100). R code to create this graph is on the book website

how many success do we have out of our trials.

Poisson Distribution

Use it for Count data. λ is the only parameters - As the mean Increases, the variance Increases.

- Classic distribution for counts; i.e., integers ≥0 to ∞ (no upper limit)
- 1 parameter: λ = mean = variance
- Approximate to Normal when λ is large
- Modifications exist for zero-inflation and over-dispersion
- Overdispersion: variance is larger than the mean; more extreme values than the Poisson distribution

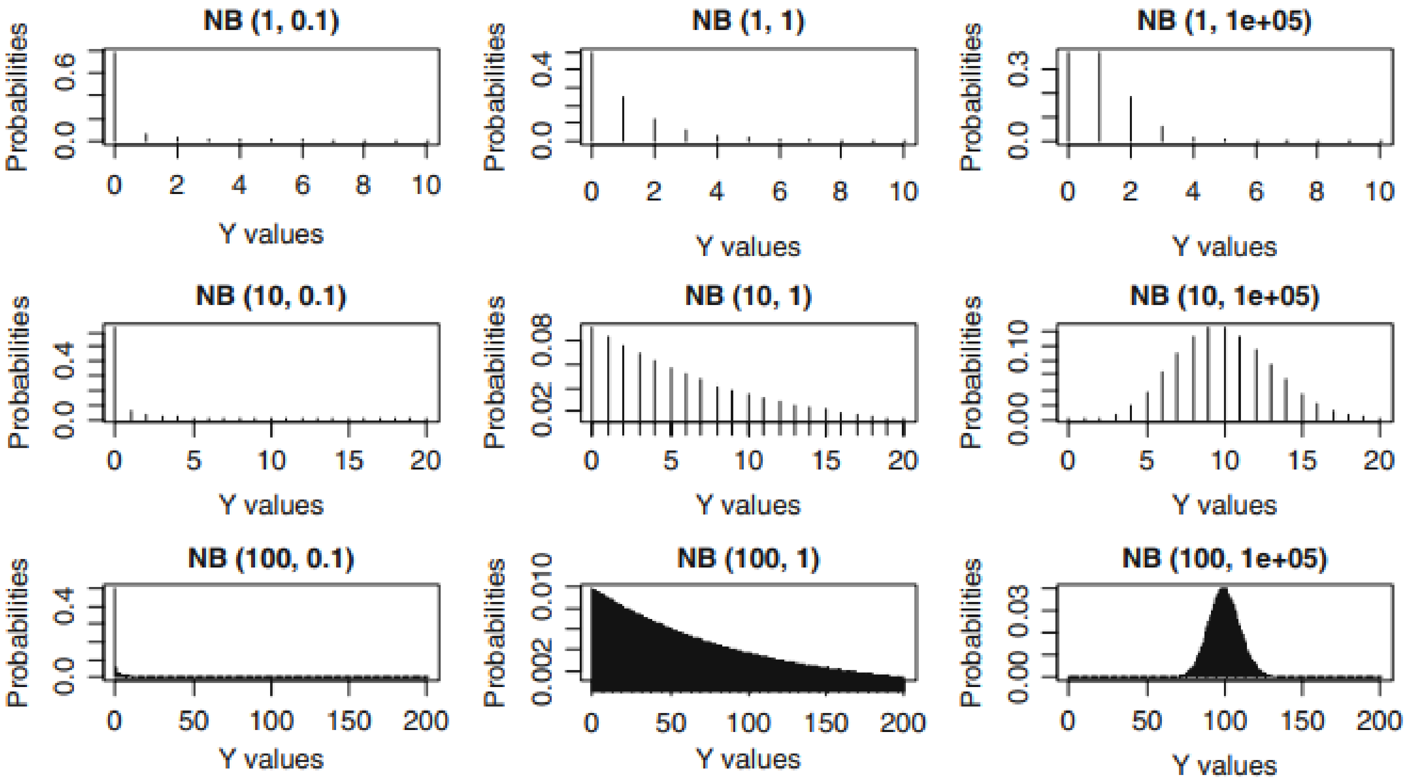


λ = λ

mean = variance.

Negative binomial distribution

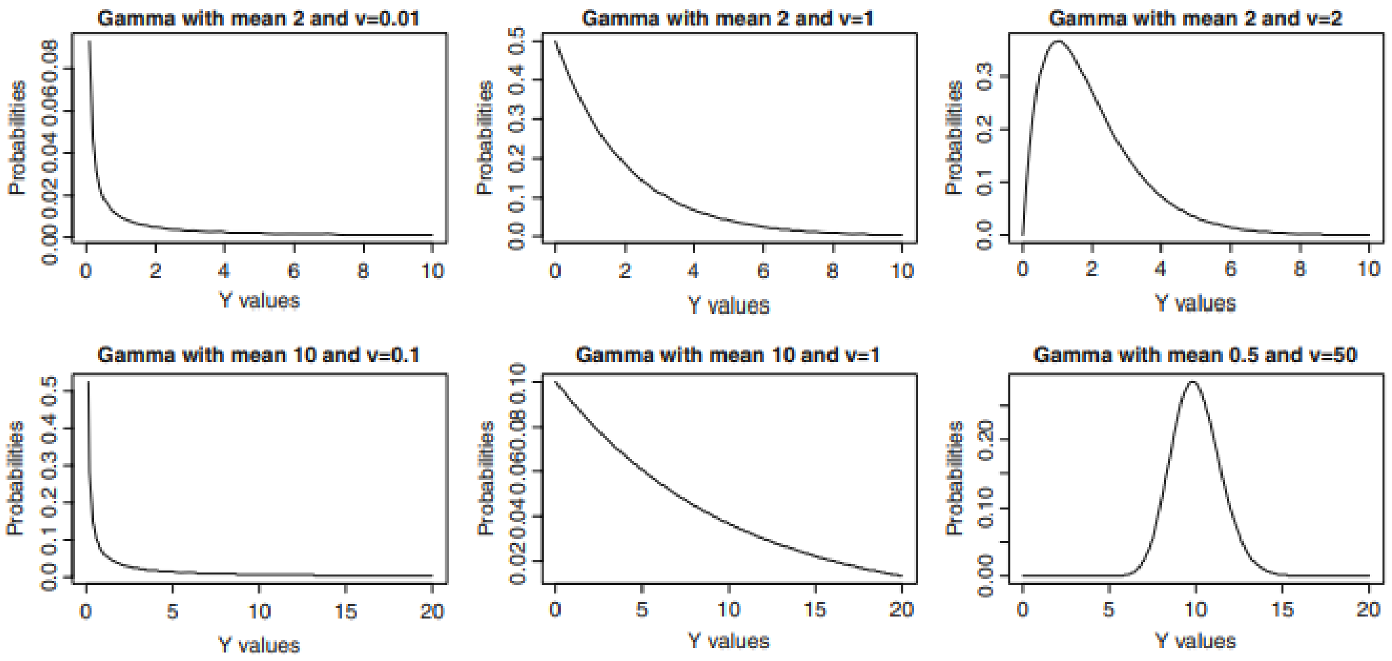
- For discrete (integers) and non-negative data
- Two parameters: μ (mean) and k (dispersion parameter)
- The smaller the k, the larger the overdispersion
- the variance is $\mu + \mu^2/k$.
- If k is large, the term $\mu^2/k = 0$, and the variance of Y is μ (poisson)



No 0 response - eg Plant biomass,

Gamma distribution

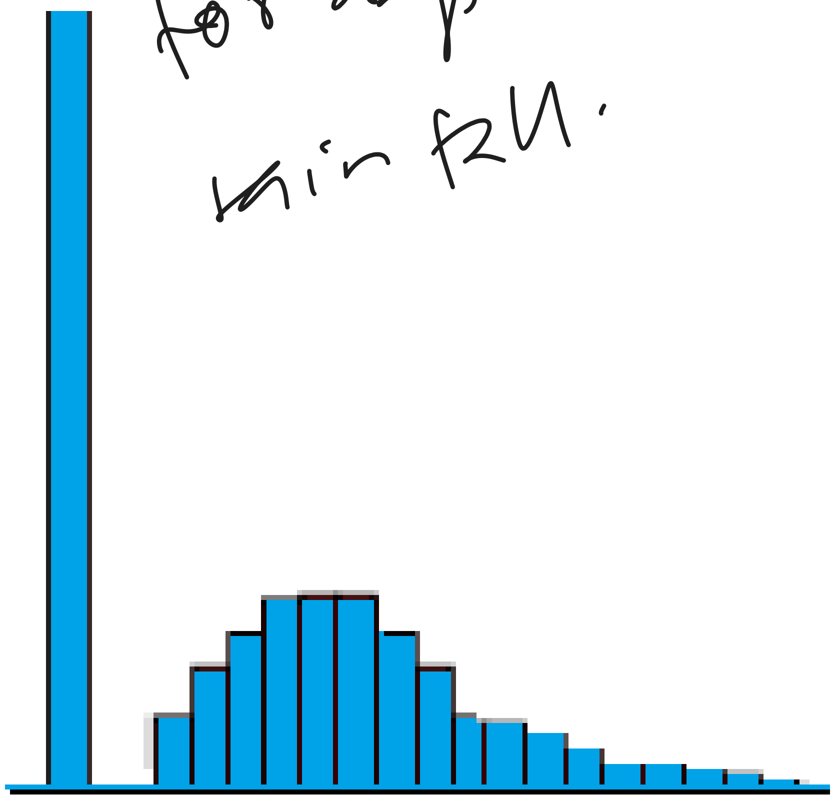
- continuous response variable *Y* that has positive values ($Y > 0$)
- Two parameters: μ (mean) and v (dispersion parameter)
- the variance is μ^2/v.
- Used for continuous data with right skew that can't be transformed



Tweedie distribution

rain fall, eg because zero can exist for days without rain fall.

- A family of distributions that are a subset of Exponential Dispersion Models (EDMs)
- useful for data with a mix of zeros and positive values (not necessarily counts).
- Three parameters: μ (mean), v (dispersion parameter), and power (*p*)
- the variance is $v\mu^{p}$
- *p*=0: normal; *p*=1: quasi-poisson; *p*=2: gamma; *p*=3: inverse gaussian
- 1<p<2: Compound Poisson-Gamma distribution



Power (P) determines the shape of distribution

What distribution should I use?

Decision should be made a priori based on the response variable

Distribution	Type of Data
Normal	Continuous
Poisson	Counts (integers) and density
Negative binomial	Overdispersed counts and density
Gamma	Continuous
Binomial	Proportional data
Bernoulli	Presence absence
Tweedie	Zero and continuous

What distribution would you choose?

the success rate of seed germination under different soil conditions

Gamma

the size of fish caught relative to depth at which they are caught (lots of small fish, few large fish)

Possion ?

the number of plant species across a climate gradient

Tweedie.

the biomass of understory story plants (with many zeros) relative to canopy cover

GLM: three main steps

Random component: choosing a distribution for the response variable

Systematic component: defining the systematic part in terms of covariates (we will skip this step!)

Link function: specifying the relationship (or: link) between the expected value of the response variable (random component) and the predictors (systematic component)

allows you to model
no normal model a normal distributed model.

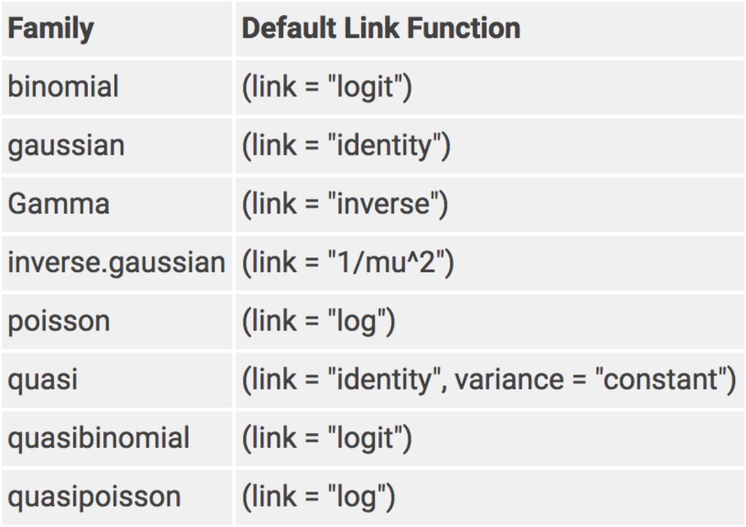
Step 3: Identify the Link function

The link function “links” or connects the linear combination of predictor variables (the linear predictor) to the expected value of the response variable by transforming the mean into the appropriate scale for the specified distribution

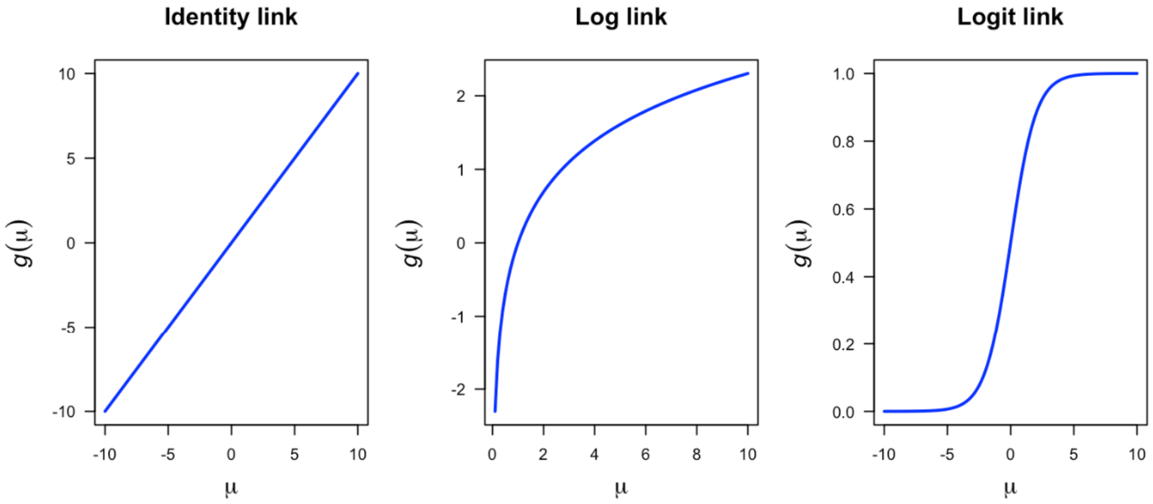
Common link functions

- Identity = ordinary regression, ANOVA, etc. (basically = no link function)
- Log link = Y is non-negative; yields a log-linear model Poisson, Count data.
- Logit link = Y is zero or one; models the log of the odds
- Inverse: Y changes at a rate inversely proportional to the predictors

Common probability distributions and their link functions; not the only options



Family	Default Link Function
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")



Assumptions of GLMS

Assumptions

- Observations are independent
- A linear relationship between the transformed expected response in terms of the link function and the explanatory variables (i.e. appropriate link function)
- The dependent variable follows the specified distribution
- The homogeneity of variance does NOT need to be satisfied. In fact, given the model structure, it is not even possible in many cases.
- BUT still need to check residual (deviance) plots (resid vs. predictors, fitted, time/space) to confirm that the variances of the residuals correspond to that expected from the distribution and link function

The R coding for GLMs is similar to lm

- The **R Code to fit a model is similar to lm(),** except that now you have to also specify an **error distribution** and **link function** using the family argument.

- The outputs are a bit different because you are not modeling the response variable directly (need to do inverse of link function to get back to the original response), but lots of overlap with lm()

Examples using the mite data

- **Question:** *Could the <u>abundance</u> or <u>occurrence</u> of Galumna sp. be predicted by environmental features?*

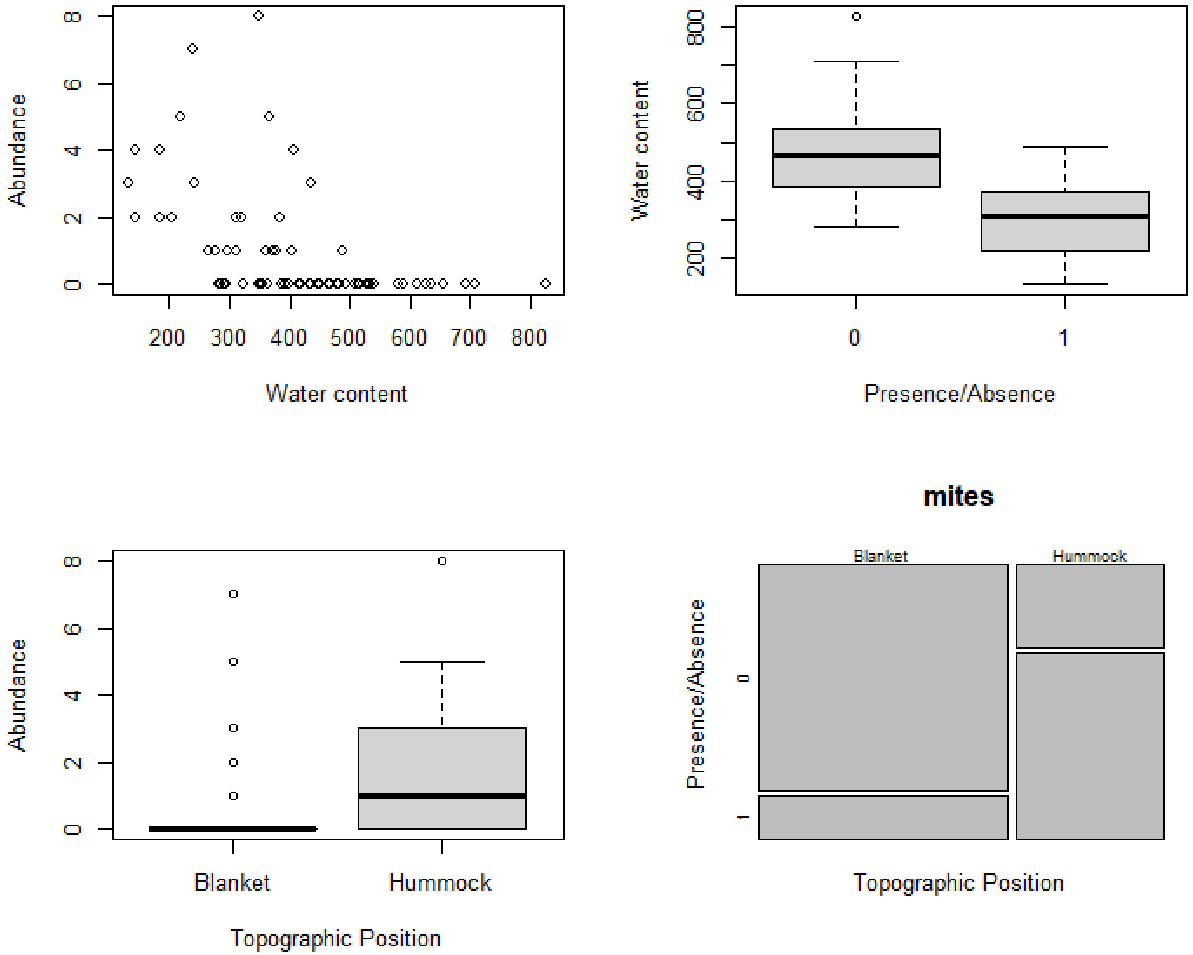
Response variables:

1. Occurrence: **`pa`**
2. Abundance: **`Galumna`**

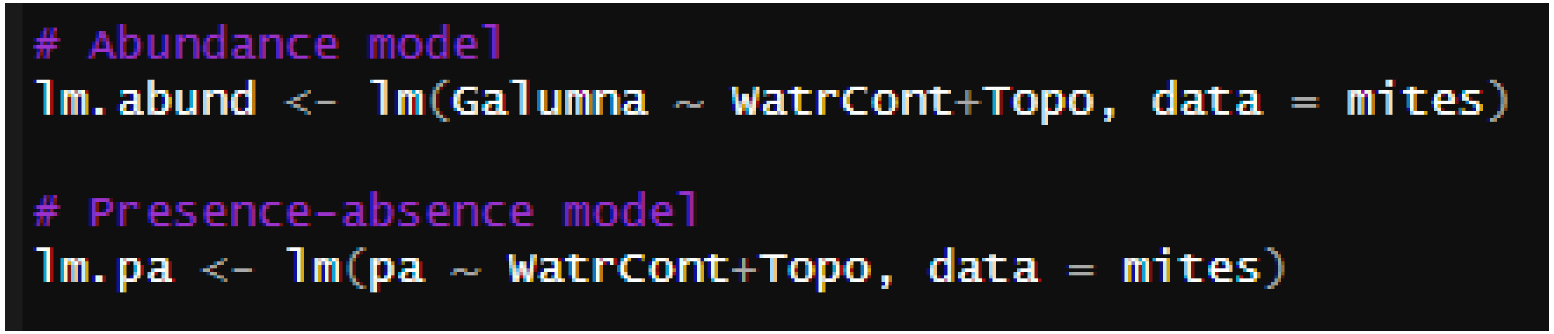
Predictor variables:

1. Substract Density: **`SubsDens`**
2. Water Content: **`WatrCont`**
3. Substrate: **`Substrate`**
4. Shrubs Nearby: **`Shrub`**
5. Topography: **`Topo`**

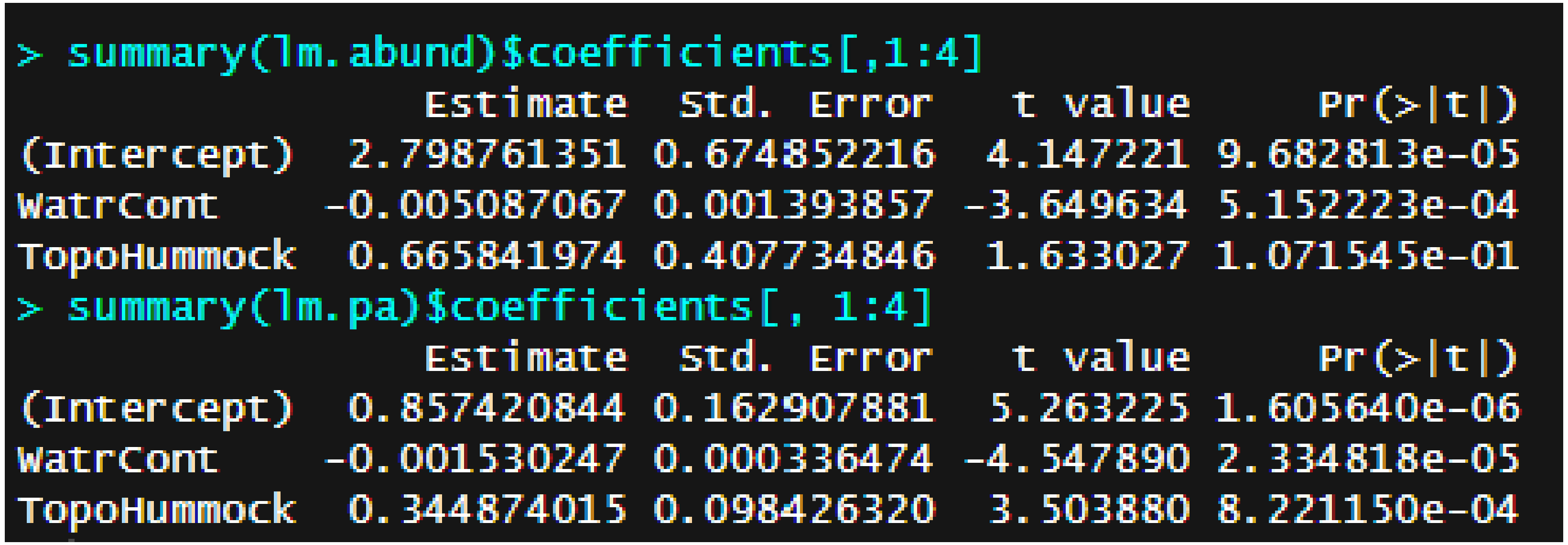
Does the composition of Galumna's communities (abundance, occurrence) vary as a function of water content and topographic position (hummock versus blanket)?



Let us use (general) linear models to test whether abundance (Galumna) and occurrence (pa) vary as a function of Water Content and Topographic position (hummock versus blanket) using the lm() function:

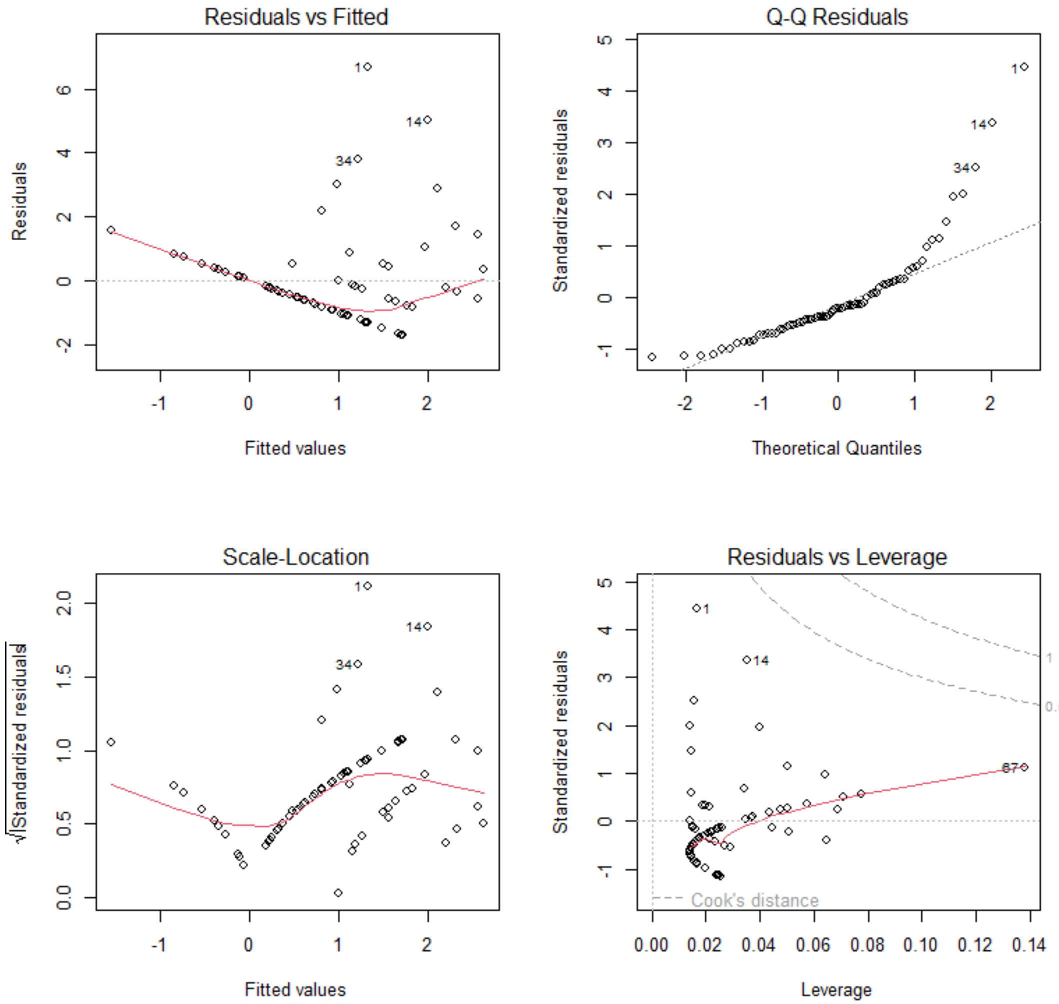


```
# Abundance model
lm.abund <- lm(Galumna ~ WatrCont+Topo, data = mites)

# Presence-absence model
lm.pa <- lm(pa ~ WatrCont+Topo, data = mites)
```

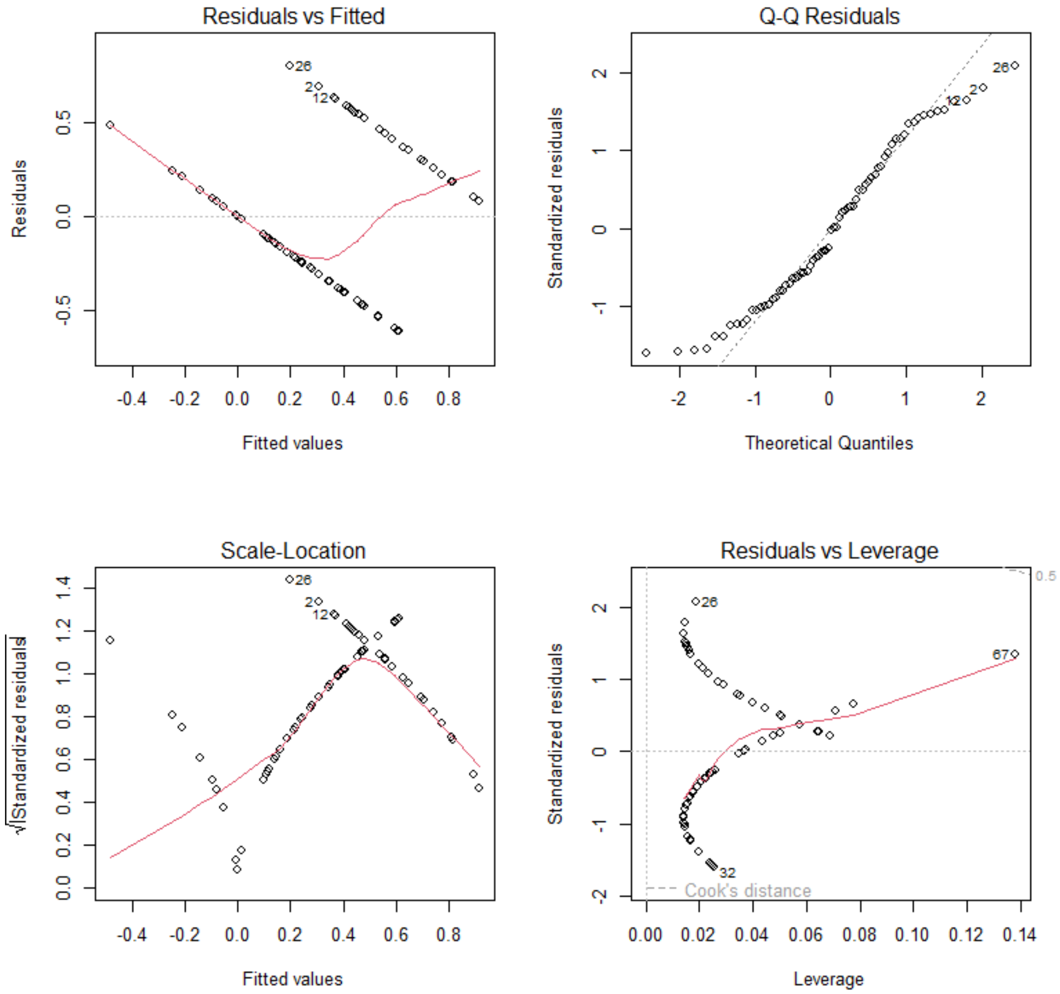


```
> summary(lm.abund)$coefficients[,1:4]
                Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  2.798761351 0.674852216  4.147221 9.682813e-05
WatrCont    -0.005087067 0.001393857 -3.649634 5.152223e-04
TopoHummock  0.665841974 0.407734846  1.633027 1.071545e-01
> summary(lm.pa)$coefficients[, 1:4]
                Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  0.857420844 0.162907881  5.263225 1.605640e-06
WatrCont    -0.001530247 0.000336474 -4.547890 2.334818e-05
TopoHummock  0.344874015 0.098426320  3.503880 8.221150e-04
```

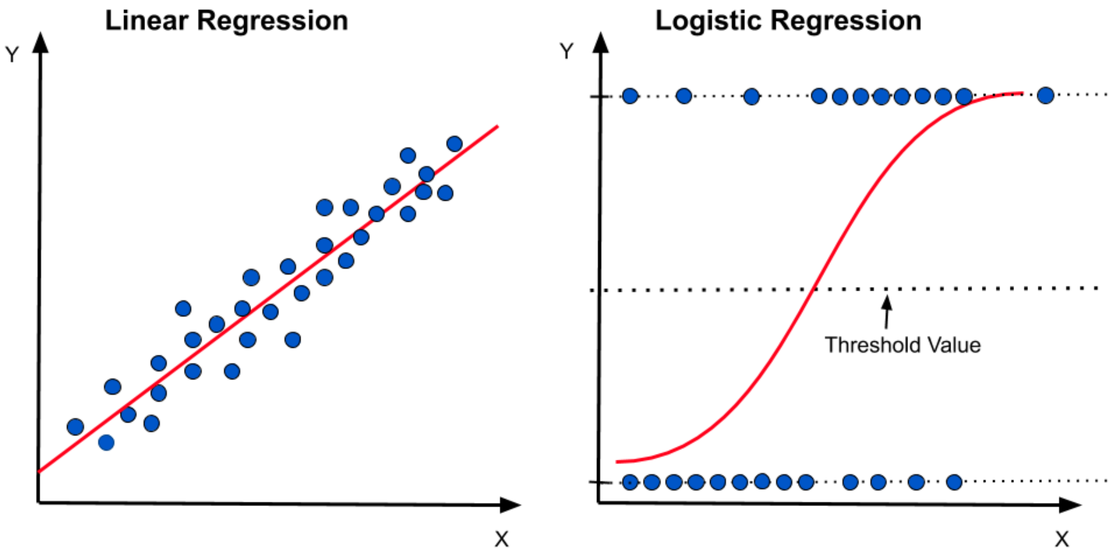
Abundance



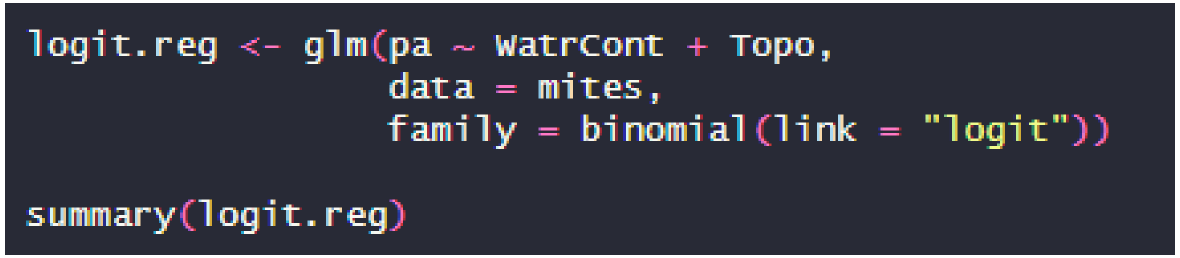
Occurrence



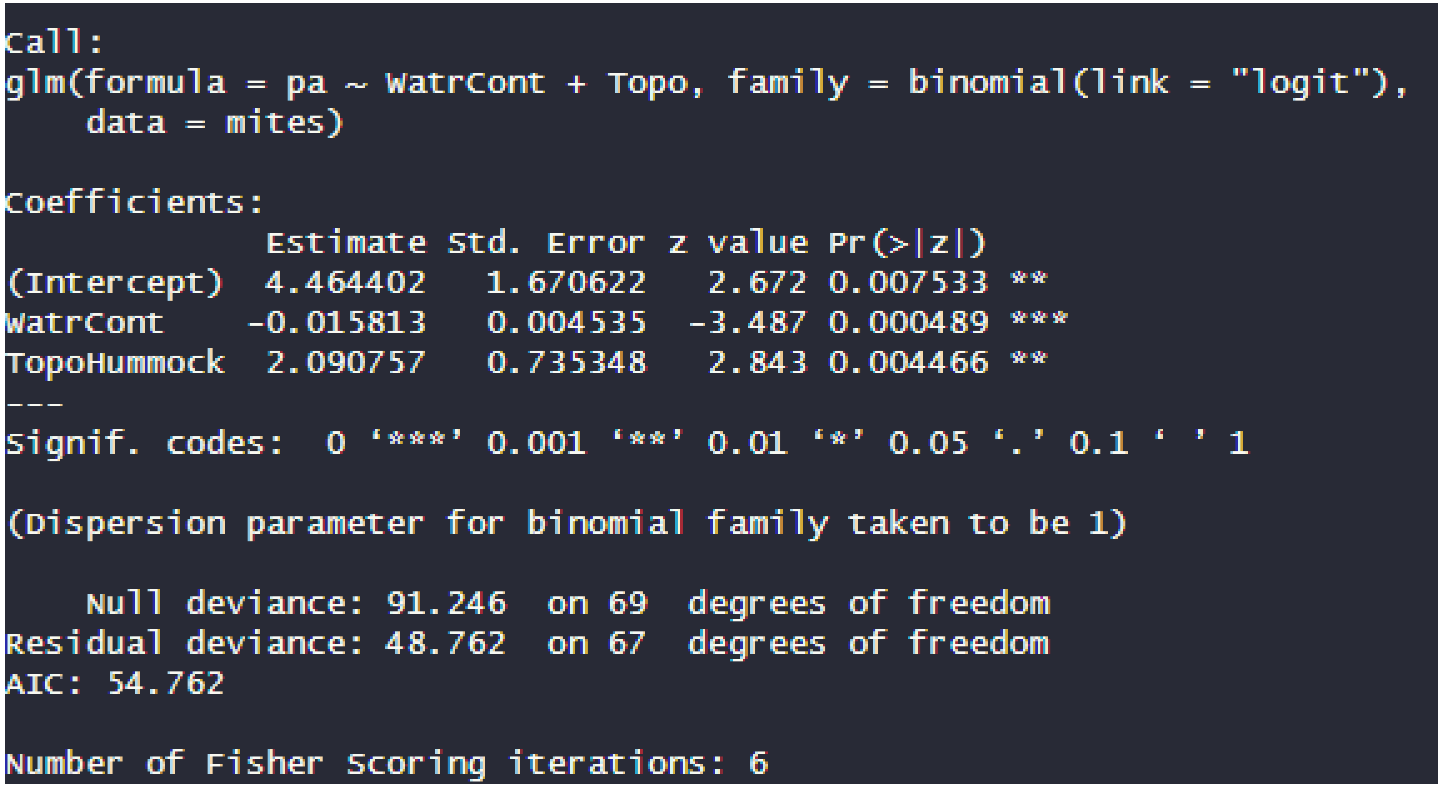
Binomial GLM – binary response



Let us build a model of the presence of Galumna sp. as a function of water content and topography.



```
logit.reg <- glm(pa ~ WatrCont + Topo,
                 data = mites,
                 family = binomial(link = "logit"))

summary(logit.reg)
```



```
Call:
glm(formula = pa ~ WatrCont + Topo, family = binomial(link = "logit"),
    data = mites)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.464402   1.670622   2.672 0.007533 **
WatrCont    -0.015813   0.004535  -3.487 0.000489 ***
TopoHummock  2.090757   0.735348   2.843 0.004466 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 91.246  on 69  degrees of freedom
Residual deviance: 48.762  on 67  degrees of freedom
AIC: 54.762

Number of Fisher Scoring iterations: 6
```

Interpreting the summary output

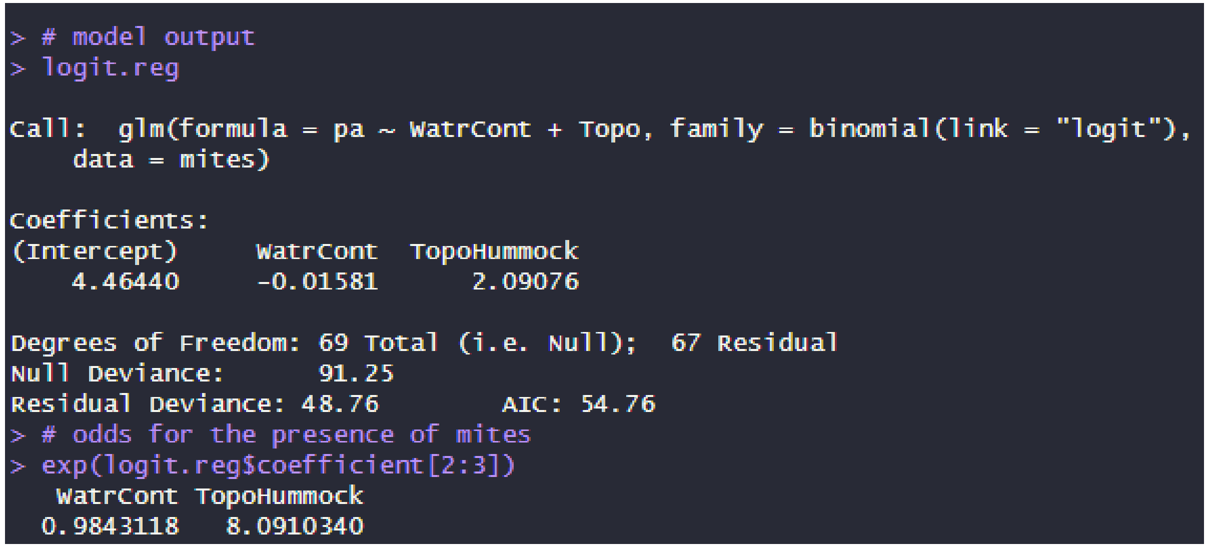
Coefficients:

- The intercept is the log-odds of the outcome when all predictors are at 0 or their reference level.
- The regression coefficient is the log of the odds ratio comparing individuals who differ in that predictor by one unit, holding the other predictors fixed.
- Use the exponential function ($e^{\beta 0}$) to convert to odds.

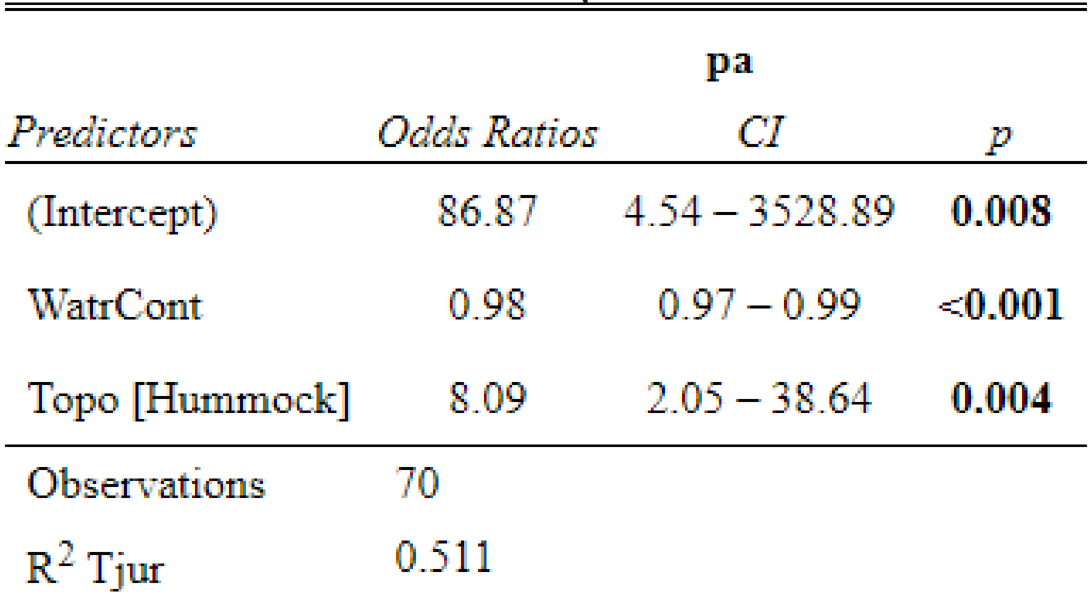
Use an Inverse function to properly
Interpret logit outcome.

From log odds to odds

exp( ) to get
inverse of logit



```
> # model output
> logit.reg

Call:  glm(formula = pa ~ WatrCont + Topo, family = binomial(link = "logit"),
    data = mites)

Coefficients:
(Intercept)     WatrCont  TopoHummock
    4.46440     -0.01581      2.09076

Degrees of Freedom: 69 Total (i.e. Null);  67 Residual
Null Deviance:      91.25
Residual Deviance: 48.76        AIC: 54.76
> # odds for the presence of mites
> exp(logit.reg$coefficient[2:3])
   WatrCont TopoHummock
  0.9843118   8.0910340
```



	pa		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	86.87	4.54 – 3528.89	**0.008**
WatrCont	0.98	0.97 – 0.99	**<0.001**
Topo [Hummock]	8.09	2.05 – 38.64	**0.004**
Observations	70		
R^2 Tjur	0.511		

- When an estimated parameter is between 0 and 1 on the odds' scale, the relationship between the response variable and the predictors is **negative**.
- If it is greater than 1, the relationship will be **positive**.
- If the confidence interval includes 1 on the odds scale, the relationship is *not significant*.

If odd ratio is less than 1, it is inverse r/ship.
If odd ratio is greater than 1, it is a direct r/ship.

From odds to probability (%)

- When odd is larger than 1:

  % = odds/(odds +1) * 100

- When the odds value is smaller than 1:

  we have to take the inverse value (*i.e.* 1 divided by the odds) to facilitate interpretation.

  % = (1/odds) - 1 * 100

  The interpretation is then how **LESS** likely it is to observe the event of interest.

Odds value is smaller than 1:

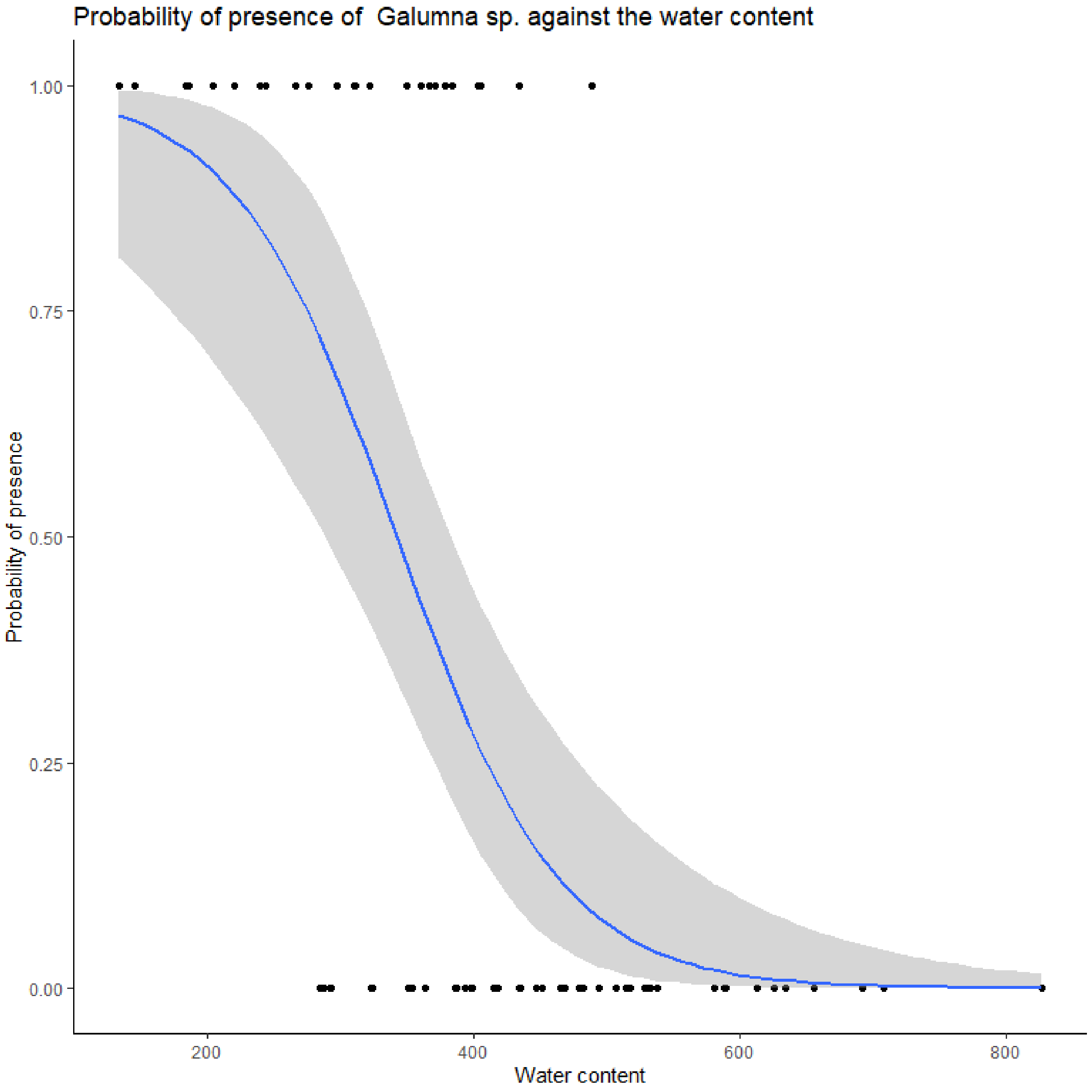
For water content, the odds is 0.984. The inverse is: 1/0.984=1.016

This means that a one-unit increase in water content decreases the likelihood of observing *Galumna* sp. by 1.016.

We can also subtract 1 from the odds value to obtain a percentage:

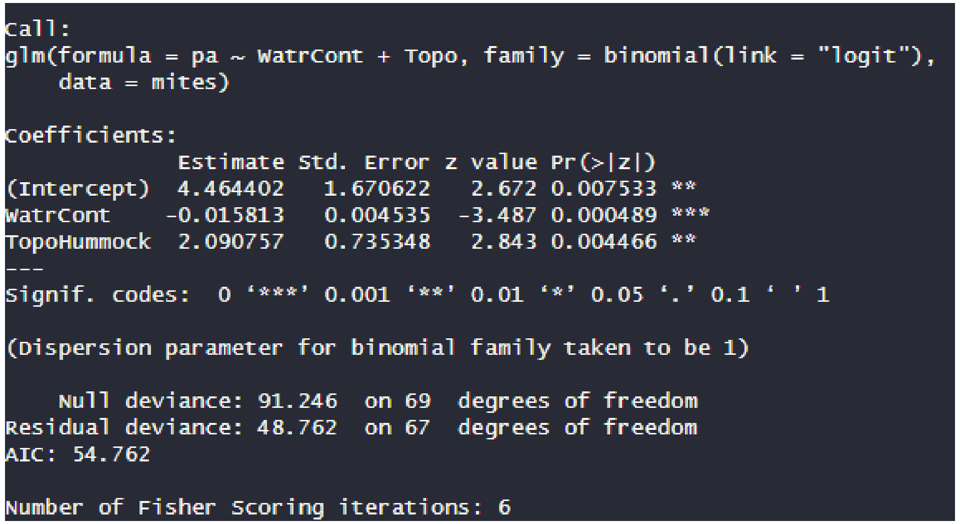
$$(1.016-1)*100=1.6$$

So there is a 1.6% decrease in the probability of observing *Galumna* sp. with a one-unit increase in water content.



Interpreting the summary output

- **Dispersion parameter for binomial family taken to be 1**: You'll only see this for Poisson and binomial (logistic) regression. It's letting you know that there has been an additional scaling parameter added to help fit the model. But, need test for overdispersion!
- **Null deviance:** The null deviance tells us how well we can predict our output only using the intercept. Smaller is better.
- **Residual deviance:** The residual deviance tells us how well we can predict our output using the intercept and our inputs. Smaller is better.



```
Call:
glm(formula = pa ~ WatrCont + Topo, family = binomial(link = "logit"), 
    data = mites)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  4.464402   1.670622   2.672 0.007533 ** 
WatrCont    -0.015813   0.004535  -3.487 0.000489 ***
TopoHummock  2.090757   0.735348   2.843 0.004466 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 91.246  on 69  degrees of freedom
Residual deviance: 48.762  on 67  degrees of freedom
AIC: 54.762

Number of Fisher Scoring iterations: 6
```

Interpreting the summary output

- We can obtain a pseudo-R^2, the analogue of the R^2 for models fitted by maximum likelihood.

- McFadden's (1973) pseudo-R^2:

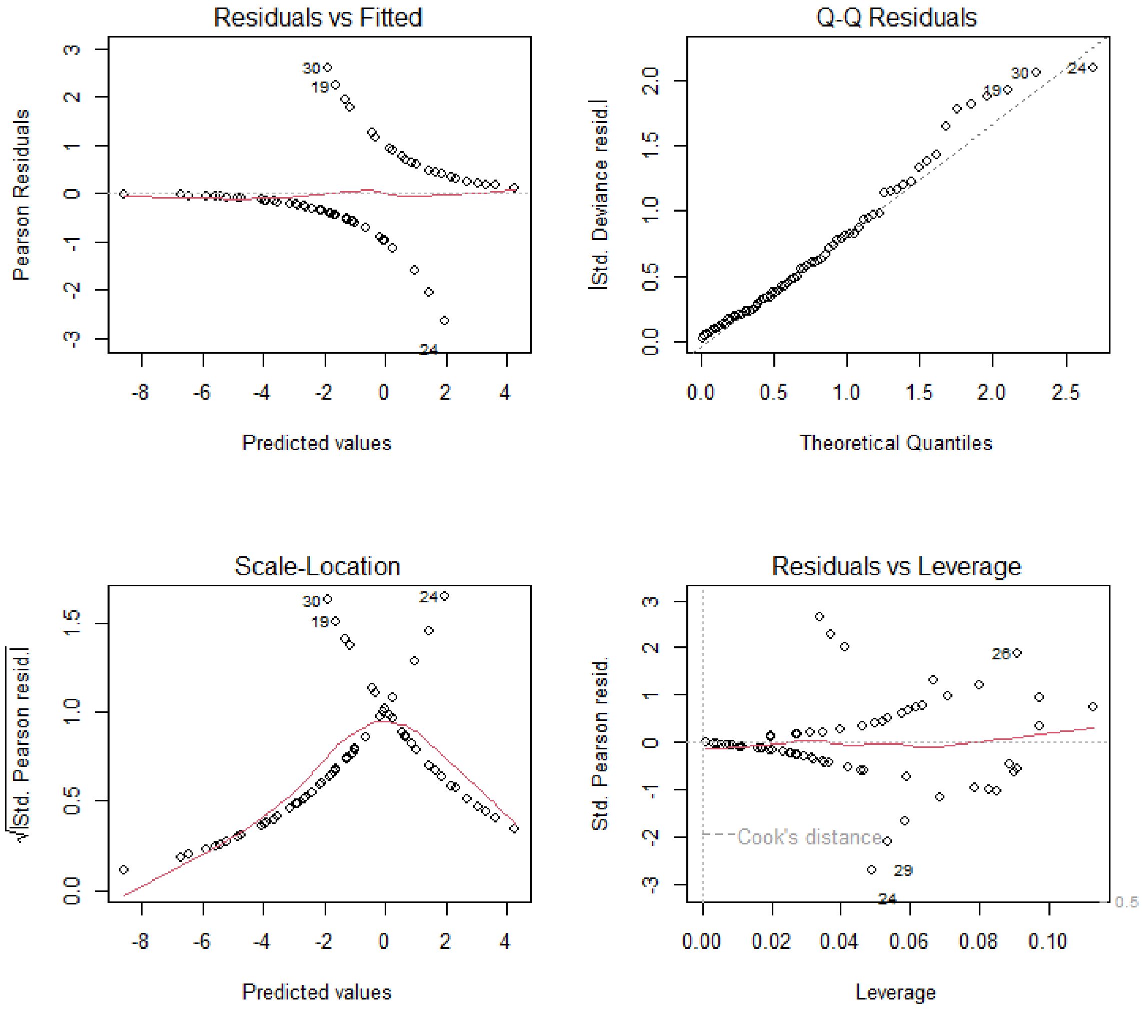
  pseudo-R^2=(null deviance - residual deviance)/null deviance

  pseudo-R^2 =0.4655

  The model explains 46.6% of the variability in the data

- Performance package has an r2() function that will give you the most “appropriate” R^2

What about model assumptions?



Take the deviance residuals and plot them against:

(i) the fitted values

(ii) each explanatory variable in the model

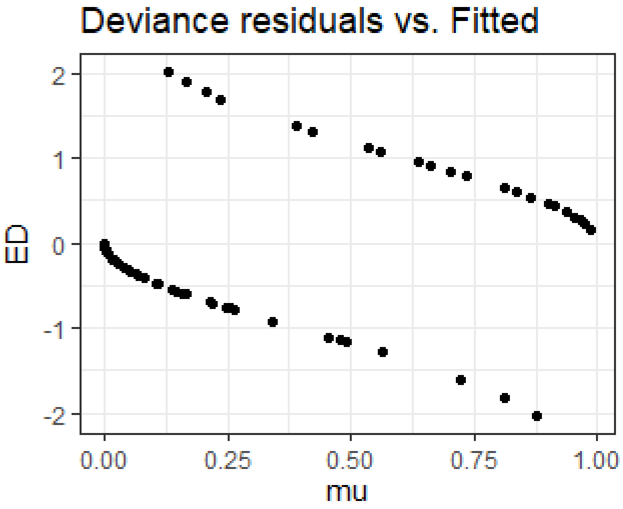
(iii) each explanatory variable not in the model (the ones not used in the model, or the ones dropped during the model selection procedure)

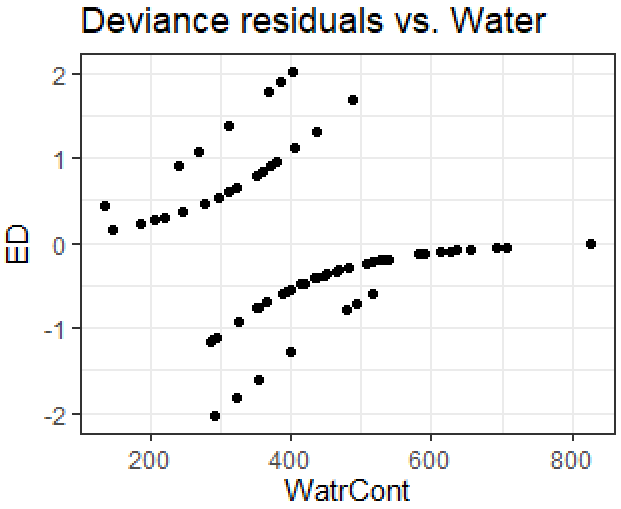
(iv) against time

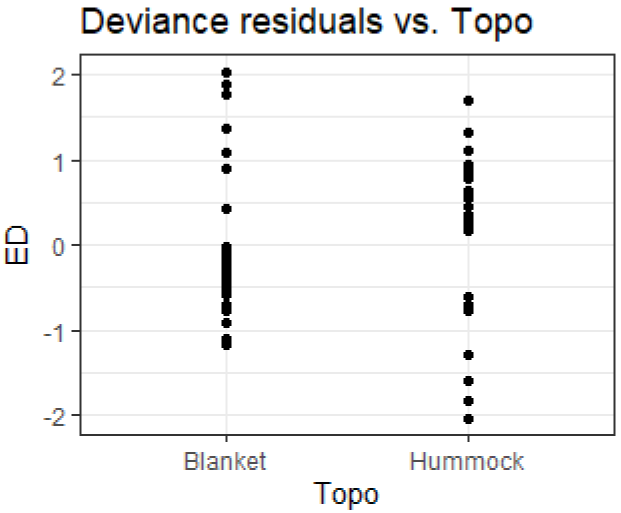
(v) against spatial coordinates, if relevant.

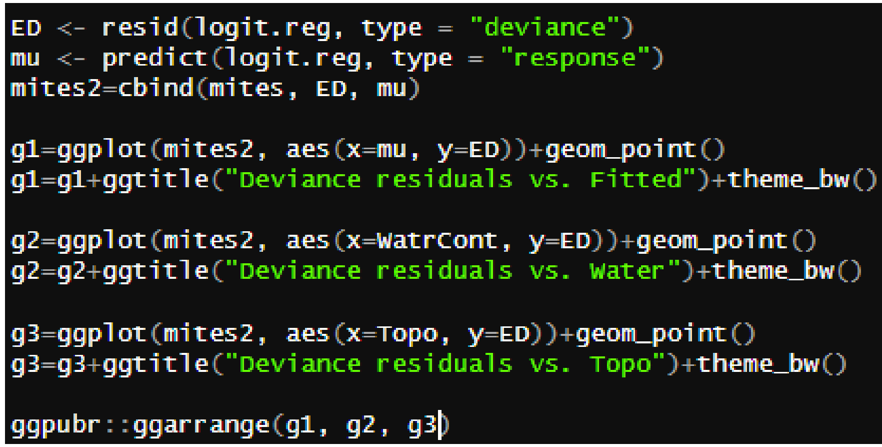
Deviance residuals

- Help measure how much each data point contributes to the model fit
- Larger deviance residuals indicate a poor fit for that particular observation.
- They should be symmetrically distributed around zero for a good model fit.


Deviance residuals vs. Fitted


Deviance residuals vs. Water


Deviance residuals vs. Topo



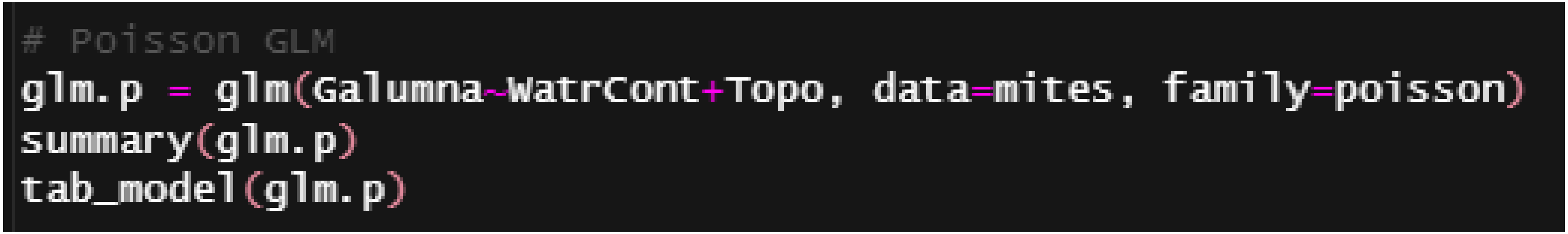
```
ED <- resid(logit.reg, type = "deviance")
mu <- predict(logit.reg, type = "response")
mites2=cbind(mites, ED, mu)

g1=ggplot(mites2, aes(x=mu, y=ED))+geom_point()
g1=g1+ggtitle("Deviance residuals vs. Fitted")+theme_bw()

g2=ggplot(mites2, aes(x=WatrCont, y=ED))+geom_point()
g2=g2+ggtitle("Deviance residuals vs. Water")+theme_bw()

g3=ggplot(mites2, aes(x=Topo, y=ED))+geom_point()
g3=g3+ggtitle("Deviance residuals vs. Topo")+theme_bw()

ggpubr::ggarrange(g1, g2, g3)
```

Count data

Count data is characterized by:

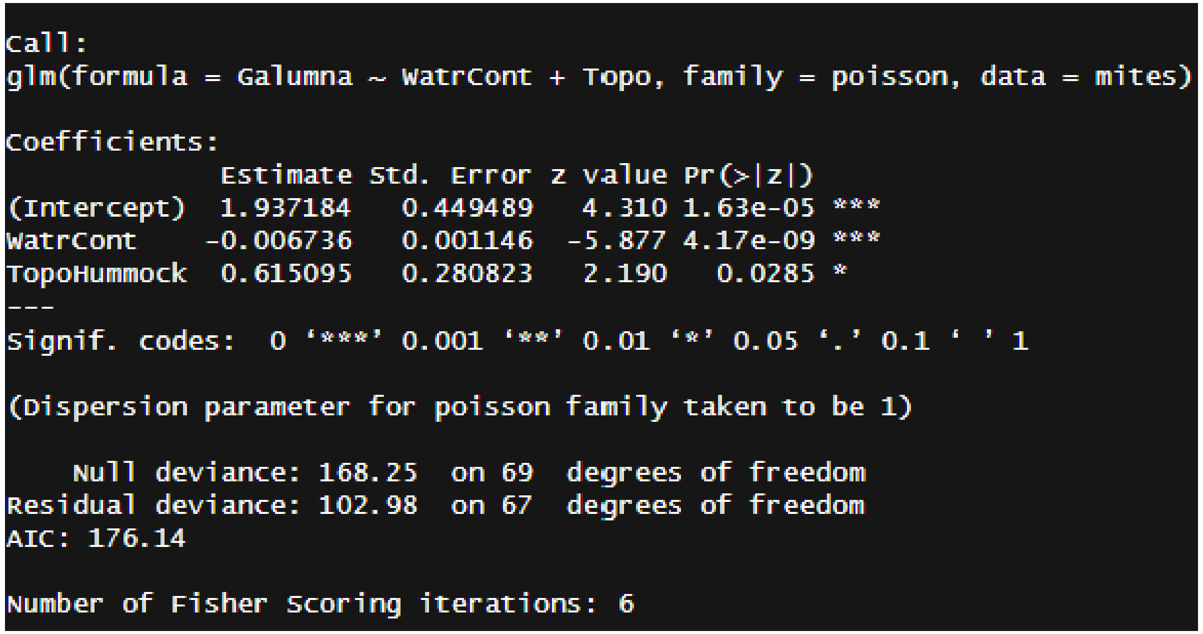
- Positive values: you do not count -7 individuals
- Integer values: you do not count 7.56 individuals
- Exhibits larger variance for large values (variance increases with mean)

Let us build a model of the abundance of *Galumna* sp. as a function of water content and topography.

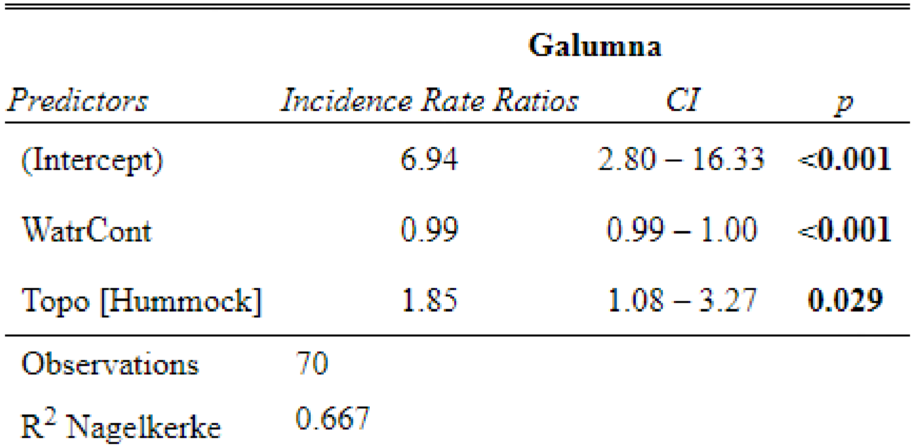


```
# Poisson GLM
glm.p = glm(Galumna~WatrCont+Topo, data=mites, family=poisson)
summary(glm.p)
tab_model(glm.p)
```

Interpreting the summary output



```
Call:
glm(formula = Galumna ~ WatrCont + Topo, family = poisson, data = mites)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.937184   0.449489   4.310 1.63e-05 ***
WatrCont    -0.006736   0.001146  -5.877 4.17e-09 ***
TopoHummock  0.615095   0.280823   2.190   0.0285 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 168.25  on 69  degrees of freedom
Residual deviance: 102.98  on 67  degrees of freedom
AIC: 176.14

Number of Fisher Scoring iterations: 6
```



	Galumna		
Predictors	*Incidence Rate Ratios*	*CI*	*p*
(Intercept)	6.94	2.80 – 16.33	**<0.001**
WatrCont	0.99	0.99 – 1.00	**<0.001**
Topo [Hummock]	1.85	1.08 – 3.27	**0.029**
Observations	70		
R^2 Nagelkerke	0.667		

For every one-unit increase in water content, the rate of Galumna occurrence decreases by a factor of 0.99, or 1%.

- IRR = exp(parameter)
- quantifies the association between predictor variables and the rate (or incidence) of an event occurring
- Percent change = (IRR-1)*100

Interpreting the summary output

- In a Poisson GLM, the residual deviance should be close to the residual degrees of freedom. However, our residual deviance is much higher than the degrees of freedom of our model!
  - 102>>67 - this indicates that the model is overdispersed.

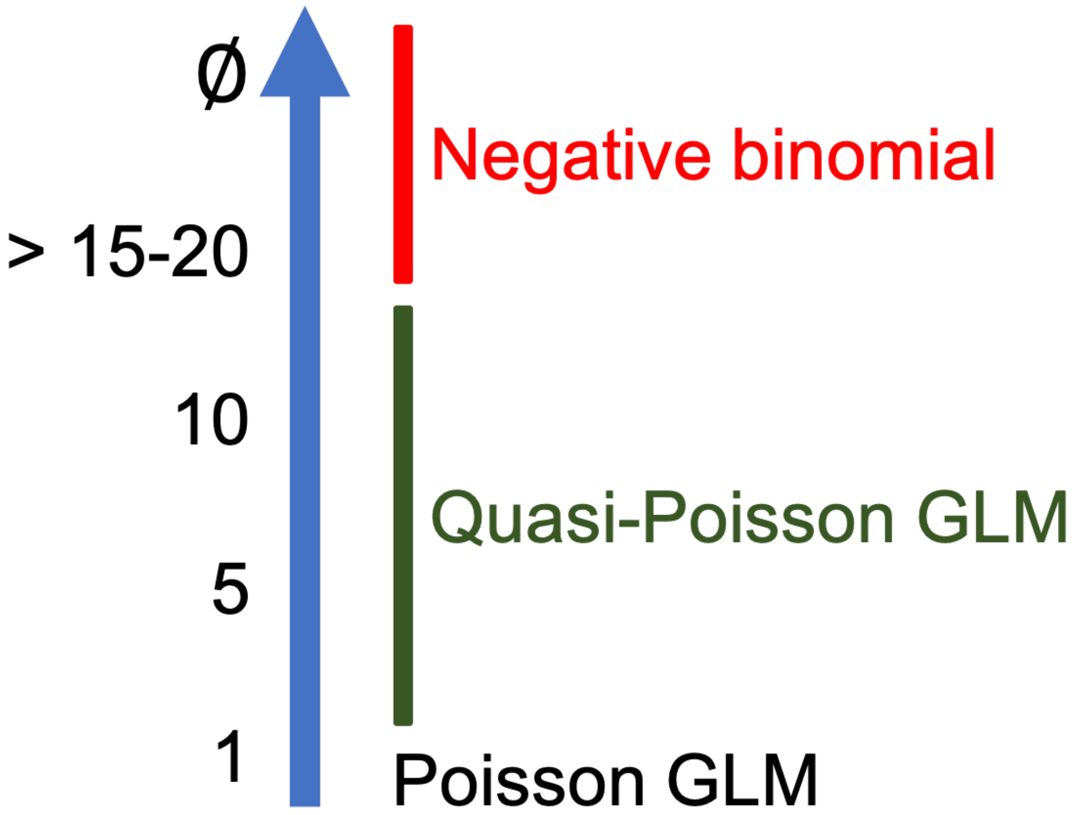
- Dispersion parameter should be close to 1
  - Dispersion parameter= 102/67 = 1.52

Overdispersion – why does it happen

- APPARENT:
  - missing covariates
  - missing interaction terms
  - Non-linear effects of covariates
  - wrong choice of link function
  - Presence of strong outliers
- REAL:
  - variance of the data is higher than expected from the Poisson distribution.
  - data includes many zeros and/or many very high values.
  - clustering or correlation of observations

Overdispersion
why should we care and what do we do?

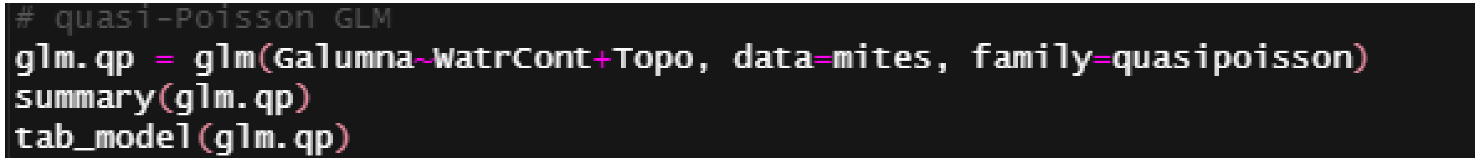
Tests on the explanatory variables will generally appear to be more significant and confidence intervals for the parameters will be narrower than warranted by the data!



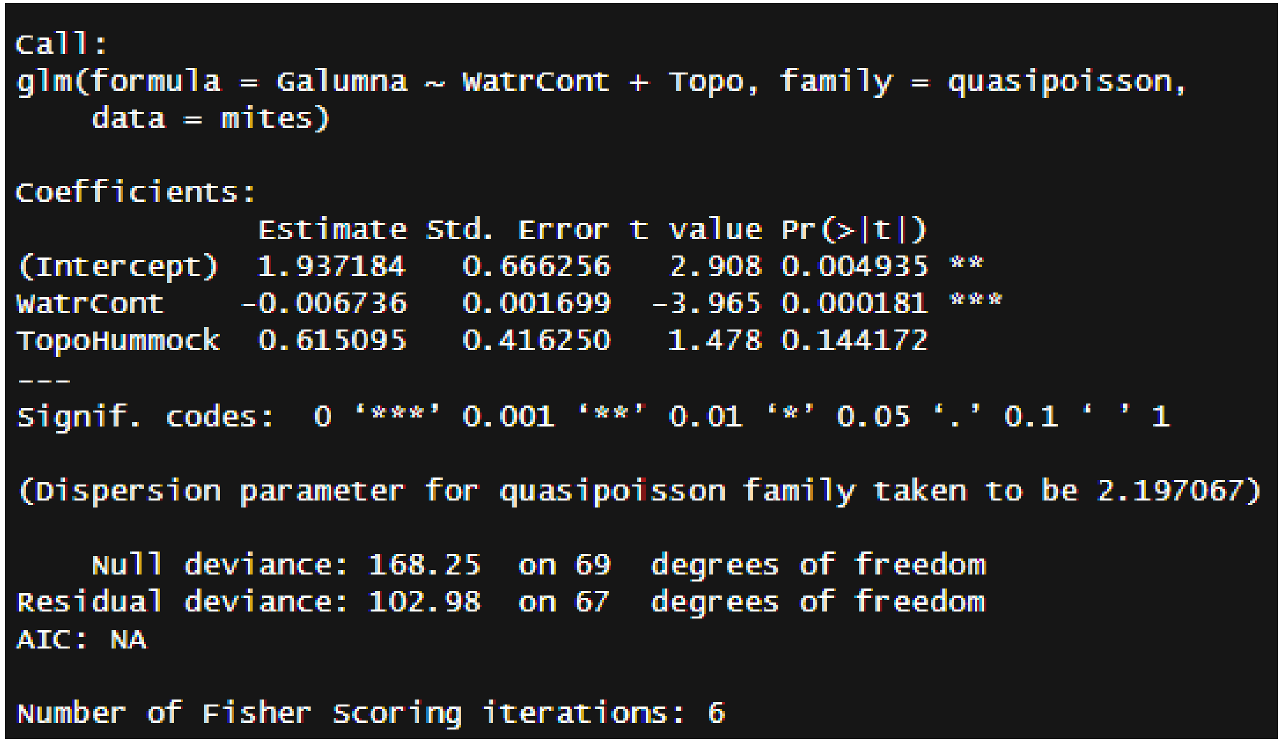
What is a quasi-poisson GLM?

- relaxes the assumption of equi-dispersion by allowing the variance to be a function of the mean
- It introduces an additional dispersion parameter, often denoted as ϕ (phi), which scales the variance of the response variable
- specified in a similar way to the standard Poisson model within the framework of GLMs.
  - It utilizes a log link function and assumes a Poisson distribution for the response variable.
  - instead of assuming that the variance is equal to the mean, it models the variance as a function of the mean, allowing for greater flexibility in handling overdispersion

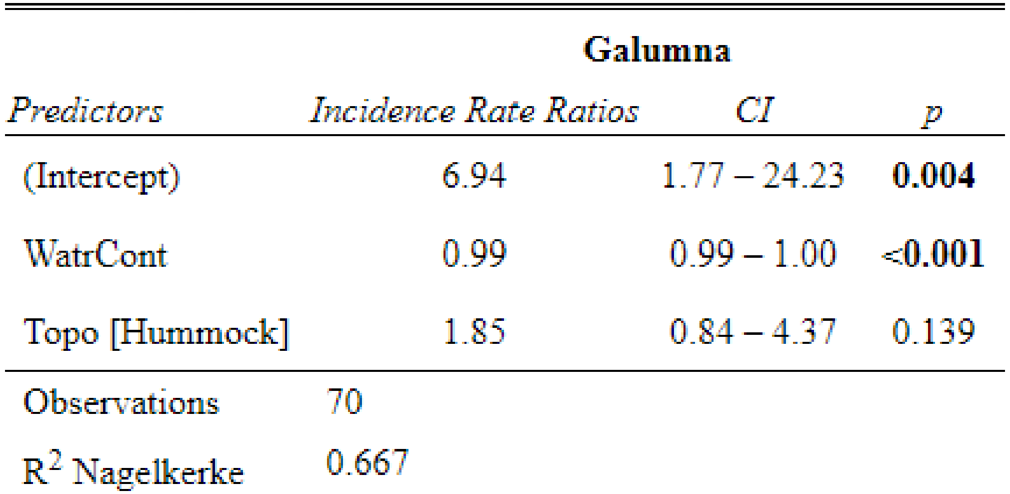
Quassi-poisson GLM



```
# quasi-Poisson GLM
glm.qp = glm(Galumna~WatrCont+Topo, data=mites, family=quasipoisson)
summary(glm.qp)
tab_model(glm.qp)
```



```
Call:
glm(formula = Galumna ~ WatrCont + Topo, family = quasipoisson,
    data = mites)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.937184   0.666256   2.908 0.004935 ** 
WatrCont    -0.006736   0.001699  -3.965 0.000181 ***
TopoHummock  0.615095   0.416250   1.478 0.144172    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 2.197067)

    Null deviance: 168.25  on 69  degrees of freedom
Residual deviance: 102.98  on 67  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```



	Galumna		
Predictors	*Incidence Rate Ratios*	*CI*	*p*
(Intercept)	6.94	1.77 – 24.23	**0.004**
WatrCont	0.99	0.99 – 1.00	**<0.001**
Topo [Hummock]	1.85	0.84 – 4.37	0.139
Observations	70		
R^2 Nagelkerke	0.667		

GLM key-points

GLMs extend linear models by allowing for response variables that follow different distributions, not just normal distributions.

Three main components: a random component (distribution of the response), a systematic component (linear predictor), and a link function

The link function defines how the linear predictor relates to the expected value of the response variable. Different link functions correspond to different distributions

The models are fitted via maximum likelihood estimation, so model reduction and selection can proceed the same as for lme or gls (exception: quasi-models)

Use deviance residuals for checking model assumptions (simulated residuals next week)