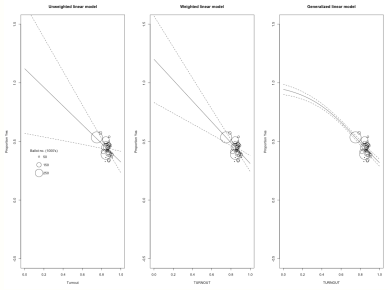
Advancing in R
Module 6: Generalized models



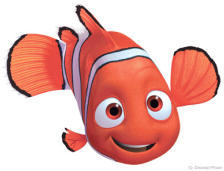
Case study

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2012) **279**, 1883–1888
doi:10.1098/rspb.2011.2041
Published online 7 December 2011

Probability of successful larval dispersal declines fivefold over 1 km in a coral reef fish

Peter M. Buston[1,*], Geoffrey P. Jones[2], Serge Planes[3] and Simon R. Thorrold[4]



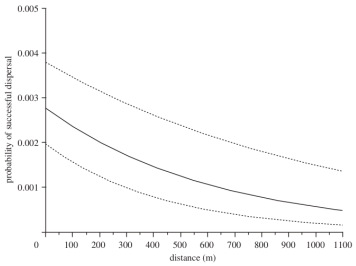
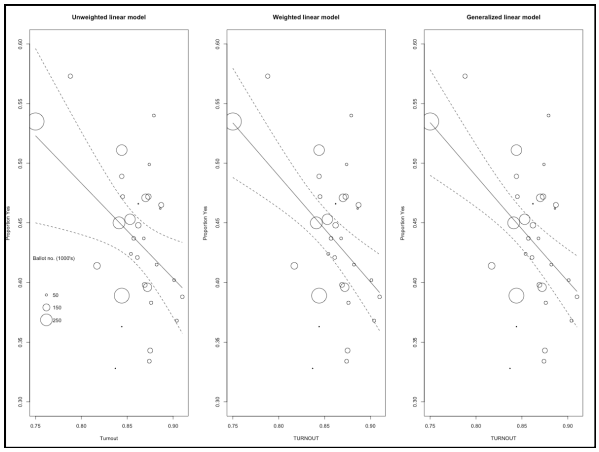
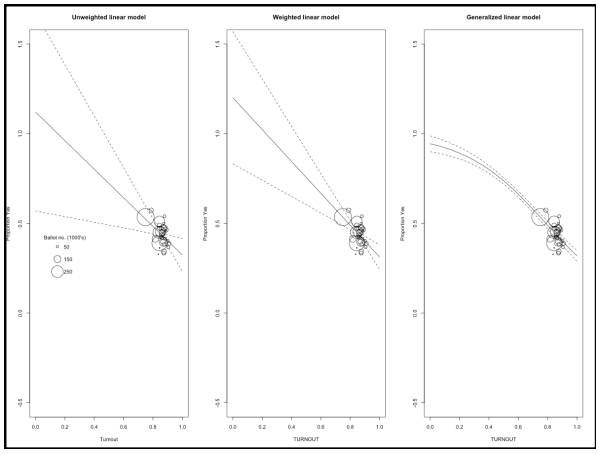


Figure 3. Probability of successful dispersal between populations as a function of distance between populations, within a metapopulation of the clown anemonefish *Amphiprion percula*. Solid line represents the relationship between probability of successful dispersal and distance between populations estimated from a logistic model (table 1). Dashed lines represent the 95% confidence intervals (CIs) around this estimated relationship (table 1). Reported probability of successful dispersal is the probability of successful dispersal over approximately eight months.

Table 1. Probability of successful dispersal between populations in relation to multiple independent variables. Summary of the results of a stepwise logistic model that investigated the effects of distance, direction, depth change and all interactions.

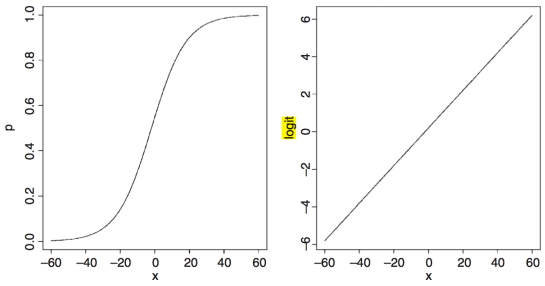
parameter	estimate	lower 95%	upper 95%	χ^2	prob > χ^2
intercept	−5.8880	−6.2263	−5.5717	1246.00	<0.0001
depth	−0.0839	−0.1374	−0.0299	9.36	0.0022
distance	−0.0016	−0.0023	−0.0009	21.31	<0.0001





Logit transform linearizes binomial probability function

$$\ln\left(\frac{p}{q}\right) = a + bx.$$



Outline

- Generalizing the linear model
- 4 common situations exhibiting:
  - Non-constant / non-homogeneous variance
  - Non-normal error distribution
- Linear predictors and link functions
- Overdispersion

"Difficult" response variables

- May not have constant variance
- Errors not normally distributed
- Logan text covers four situations:
  - Count data (no proportions)
  - Proportion data
  - Binary responses
  - "Time-to-event" data

GLMs (pronounced glims): Generalized linear models

- Not to be confused with general linear models (also sometimes called GLMs)
- Have three properties
  - Error structure
  - Linear predictor
  - Link function

Error structure

- We previously assumed normal errors
- Actual errors can have properties that violate this assumption in several ways:
  - Strong skew
  - Kurtosis
  - Strict bounds
    - e.g., predicted values must be between 0 & 1, or predicted values never below zero

Error structure

- Four common error structures:
  - Poisson errors, for count data
  - Binomial errors, for proportion data
  - Exponential errors, for time to event
  - Gamma errors, for data with constant CV

Linear predictor

- The sum of linear effects of 1 or more explanatory variables
- GLM compares *transformed* value from linear predictor to observations
  - Transformation specified by the link function
  - Fitted value = predicted value * reciprocal of link function

Link function

- Relates the mean of y to linear predictor
- Model prediction is not y except in special case we have used until now: the identity link
- 4 canonical (default) link functions

Error	Canonical link
normal	*identity*
poisson	*log*
binomial	*logit*
Gamma	*reciprocal*

From Logan:

Table 17.1 Common generalized linear models and associated canonical link-distribution pairs.

Model	Response variable	Predictor variable(s)	Residual distribution	Link
Linear regression[a]	Continuous	Continuous/ Categorical	Gaussian (normal)	Identity $g(\mu) = \mu$
Logistic regression	Binary	Continuous/ Categorical	Binomial	Logit $g(\mu) = log_e \frac{\mu}{1-\mu}$
Log-linear models	Counts	Categorical	Poisson	Log $g(\mu) = log_e \mu$

[a] Includes the standard ANOVA and ANCOVA designs.

Count data

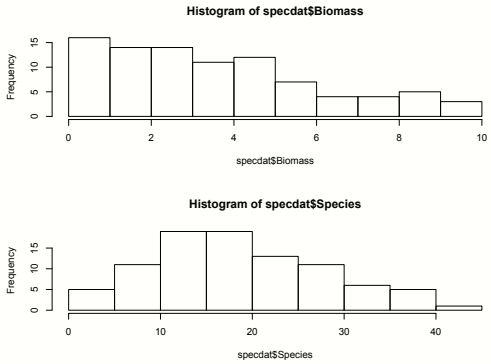
- These are data on frequencies rather than proportions
- Count data are bounded below (can't have counts <0)
- Variance not constant (increases with the mean)
- Errors not normally distributed
- Data are whole numbers (integers), which affects the error distribution

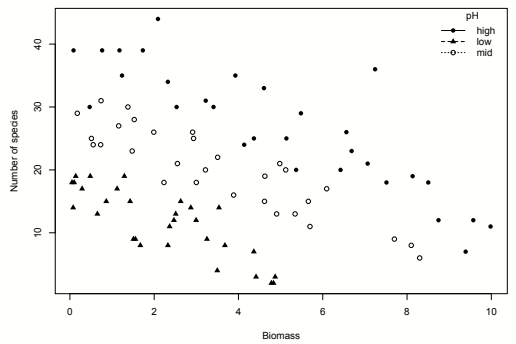
Count data

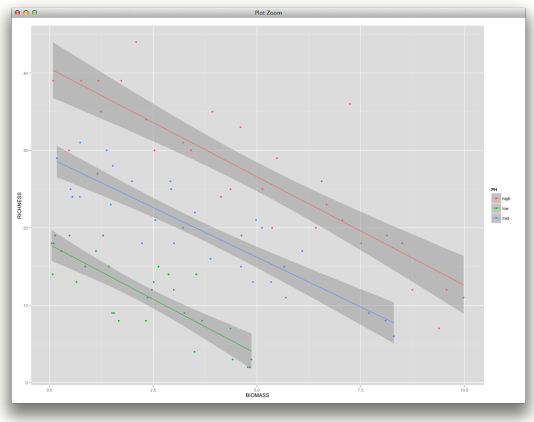
- Use generalized linear model
  - with log link (to ensure fitted values bounded below)
  - and family=poisson to specify appropriate error variance
  - Can specify family=quasipoisson if data are overdispersed, but this is quite conservative
  - Alternative error distributions available (e.g., negative binomial distributions), but you need to read about these yourself!

Data concern the number of species in a plot as a function of Biomass and pH

```
> head(specdat)
    pH   Biomass Species
1 high 0.4692972      30
2 high 1.7308704      39
3 high 2.0897785      44
4 high 3.9257871      35
5 high 4.3667927      25
6 high 5.4819747      29
> str(specdat)
'data.frame':  90 obs. of  3 variables:
 $ pH     : Factor w/ 3 levels "high","low","mid": 1 1 1 1 1 1 1
   1 1 1 ...
 $ Biomass: num  0.469 1.731 2.09 3.926 4.367 ...
 $ Species: int  30 39 44 35 25 29 23 18 19 12 ...
```







```
> summary(model1)
Call:
glm(formula = Species ~ Biomass * pH, family = poisson, data =
   specdat)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4978  -0.7485  -0.0402   0.5575   3.2297

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.76812    0.06153  61.240  < 2e-16 ***
Biomass       -0.10713    0.01249  -8.577  < 2e-16 ***
pHlow         -0.81557    0.10284  -7.931 2.18e-15 ***
pHmid         -0.33146    0.09217  -3.596 0.000323 ***
Biomass:pHlow -0.15503    0.04003  -3.873 0.000108 ***
Biomass:pHmid -0.03189    0.02308  -1.382 0.166954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 452.346  on 89  degrees of freedom
Residual deviance:  83.201  on 84  degrees of freedom
AIC: 514.39
Number of Fisher Scoring iterations: 4
```

Overdispersion

- Variance of Poisson or binomial models assumed to relate to mean or sample size, respectively
- Dispersion (variance) parameter set to 1
- Often can get more (or less) variance than expected
- If residual deviance/df < 0.5 OR > 2, can use quasibinomial or quasipoisson to model the dispersion (but this will be conservative)
- Or try other error distributions:
  - *e.g.*, negative binomial (negbin) for count data; betabinomial (betabin) for binomial data (see {aod})
- Or try other models (e.g., zero-altered models)


```
> summary(model1)
Call:
glm(formula = Species ~ Biomass * pH, family = poisson, data =
   specdat)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4978  -0.7485  -0.0402   0.5575   3.2297

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.76812    0.06153  61.240  < 2e-16 ***
Biomass       -0.10713    0.01249  -8.577  < 2e-16 ***
pHlow         -0.81557    0.10284  -7.931 2.18e-15 ***
pHmid         -0.33146    0.09217  -3.596 0.000323 ***
Biomass:pHlow -0.15503    0.04003  -3.873 0.000108 ***
Biomass:pHmid -0.03189    0.02308  -1.382 0.166954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 452.346  on 89  degrees of freedom
Residual deviance:  83.201  on 84  degrees of freedom
AIC: 514.39
Number of Fisher Scoring iterations: 4
```

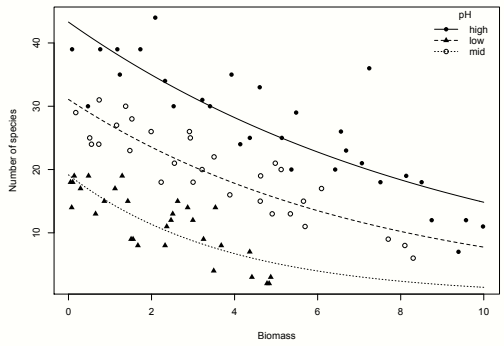

Can't do an F-test to compare generalized models

```
> model1<-glm(Species~Biomass*pH,data=specdat,poisson)
>
> model2<-glm(Species~Biomass+pH,data=specdat,poisson)
> anova(model1,model2,test="Chi")
Analysis of Deviance Table

Model 1: Species ~ Biomass * pH
Model 2: Species ~ Biomass + pH
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        84     83.201
2        86     99.242 -2   -16.04 0.0003288 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(model1)
Call:
glm(formula = Species ~ Biomass * pH, family = poisson, data =
   specdat)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4978  -0.7485  -0.0402   0.5575   3.2297

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.76812    0.06153  61.240  < 2e-16 ***
Biomass      -0.10713    0.01249  -8.577  < 2e-16 ***
pHlow        -0.81557    0.10284  -7.931 2.18e-15 ***
pHmid        -0.33146    0.09217  -3.596 0.000323 ***
Biomass:pHlow -0.15503    0.04003  -3.873 0.000108 ***
Biomass:pHmid -0.03189    0.02308  -1.382 0.166954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 452.346  on 89  degrees of freedom
Residual deviance:  83.201  on 84  degrees of freedom
AIC: 514.39
Number of Fisher Scoring iterations: 4
```



Binomial data

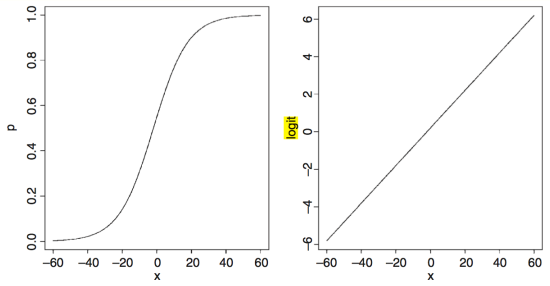
- These are data on proportions or binary outcomes
- Errors not normally distributed
- Variance not constant
- Response is bounded (by 1 above and by 0 below)
- Calculating a percentage and transforming, loses information of the size of the sample from which the proportion was estimated.

Binomial data

- Use generalized linear model
  - with logit link (to ensure fitted values bounded both above and below)
  - and family=binomial to specify appropriate error variance
  - If data as counts of two outcomes, bind columns to create 2-vector response
  - If data are binary outcome, leave as is

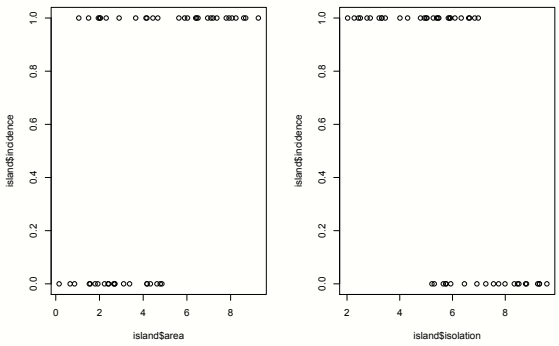
Logit transform linearizes binomial probability function

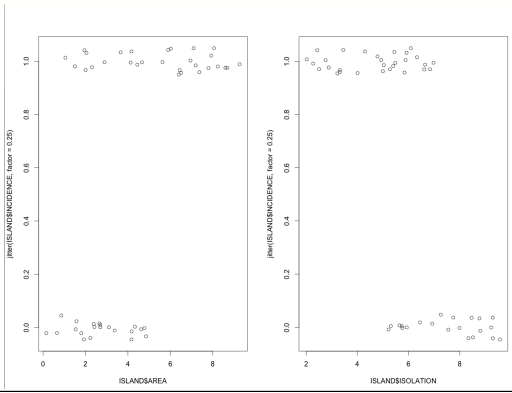
$$\ln\left(\frac{p}{q}\right) = a + bx.$$



Data concern the incidence of breeding birds on islands (yes or no) as a function of area and isolation (distance from mainland)

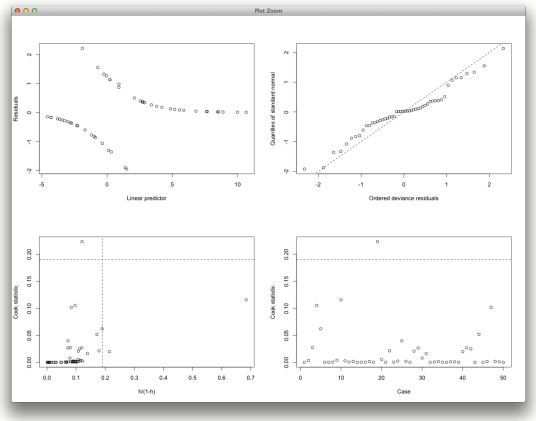
```
> str(island)
'data.frame':  50 obs. of  3 variables:
 $ incidence: int  1 0 1 0 0 1 1 0 1 1 ...
 $ area     : num  7.93 1.93 2.04 4.78 1.54 ...
 $ isolation: num  3.32 7.55 5.88 5.93 5.31 ...
> head(island)
  incidence  area isolation
1         1 7.928     3.317
2         0 1.925     7.554
3         1 2.045     5.883
4         0 4.781     5.932
5         0 1.536     5.308
6         1 7.369     4.934
```





```
> model1<-glm(incidence~area*isolation,data=island,
   family=binomial)
>
> model2<-glm(incidence~area+isolation,data=island,
   family=binomial)
>
> anova(model1,model2,test="Chi")
Analysis of Deviance Table

Model 1: incidence ~ area * isolation
Model 2: incidence ~ area + isolation
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         46     28.252
2         47     28.402 -1 -0.15043    0.6981
```




```
> summary(model2)
Call:
glm(formula = incidence ~ area + isolation, family = binomial,
    data = island)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8189  -0.3089   0.0490   0.3635   2.1192

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.6417     2.9218   2.273  0.02302 *
area          0.5807     0.2478   2.344  0.01909 *
isolation    -1.3719     0.4769  -2.877  0.00401 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.029  on 49  degrees of freedom
Residual deviance: 28.402  on 47  degrees of freedom
AIC: 34.402

Number of Fisher Scoring iterations: 6
```

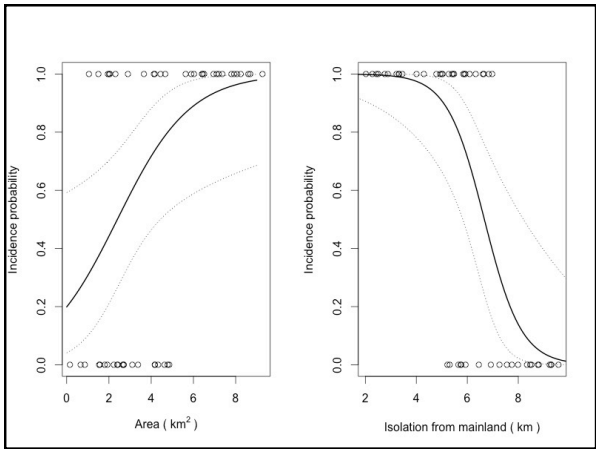
```
> summary(model2)
Call:
glm(formula = incidence ~ area + isolation, family = binomial,
    data = island)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8189  -0.3089   0.0490   0.3635   2.1192

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.6417     2.9218   2.273  0.02302 *
area          0.5807     0.2478   2.344  0.01909 *
isolation    -1.3719     0.4769  -2.877  0.00401 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.029  on 49  degrees of freedom
Residual deviance: 28.402  on 47  degrees of freedom
AIC: 34.402

Number of Fisher Scoring iterations: 6
```



Suggested reading:

- Ch. 17 in Logan
- Chs. 13, 14, 16 in Crawley