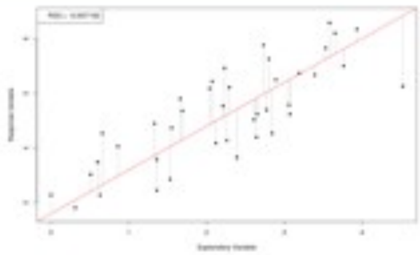
[illegible] - Planning and evaluation of biological experiments

Generalised Linear Models (GLMs)

Owen Jones
jones@biology.sdu.dk

Linear models

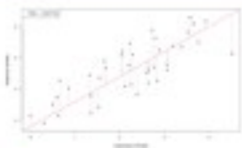


http://spark.rstudio.com/jonesor/leastsquares/

Linear models

$$y = a + bx$$

intercept   slope



$$y = \alpha + \beta x + \epsilon$$
$$y = \beta_0 + \beta_1 x + \epsilon$$
$$y = \mathbf{X}\beta + \epsilon$$

Textbooks differ in how this is represented: these are identical.

Linear models

$$y = \beta_0 + \beta_1 x + \epsilon$$

Response

Parameters (slope, intercept)

Explanatory variable

Error (variance in Y that is not explained by X)

Note: (1) error does not change with explanatory variable
(2) change in EV results in linear change in y

Linear models

$$y = \mathbf{X}\beta + \epsilon$$

Response

Explanatory variable(s)

Parameters (e.g. slope, intercept)

Error (variance in Y that is not explained by X)

Note: (1) error does not change with explanatory variable
(2) change in EV results in linear change in y

Linear models

Remember - we are modeling the *expected* response:

$$E(\mathbf{Y}) = \mathbf{X}\beta + \epsilon$$

Expected response

and since, $E(Y) \equiv \mu$

we can rephrase to:

$$\mu = \mathbf{X}\beta + \epsilon$$

Linear models

$$\mu = \mathbf{X}\beta + \epsilon$$

Expected Response

Explanatory variable(s)

Parameters
(e.g. slope, intercept)

Error
(variance in Y that is not explained by X)

Includes:
Ordinary regression
Multiple regression
ANOVA/MANOVA
ANCOVA

Linear models

Assumptions

- *linearity* of the relationship between explanatory variable(s) and response.
- *independence* of the errors.
- *constant variance* (homoscedasticity) vs. the response variable
- *normality* of the error distribution.

Generalised linear models

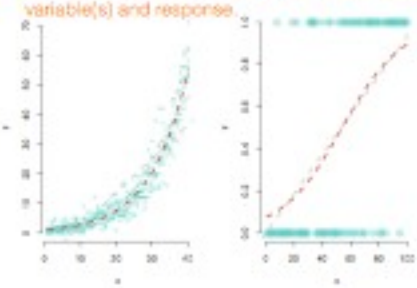
Assumptions

- *linearity* of the relationship between explanatory variable(s) and response.
- *independence* of the errors.
- *constant variance* (homoscedasticity) vs. the response variable
- *normality* of the error distribution.

Generalised Linear Models relax these assumptions.

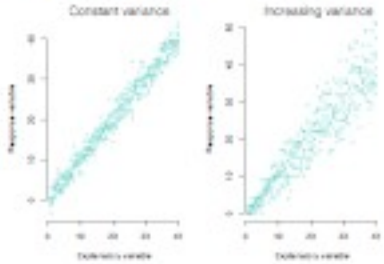
Linearity

- *linearity* of the relationship between explanatory variable(s) and response.



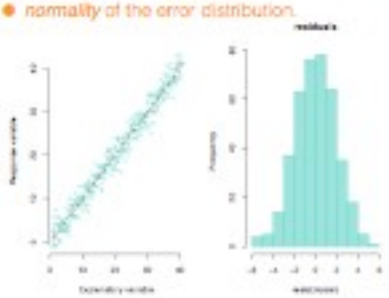
Constant variance

- *constant variance* (homoscedasticity) vs. the response variable



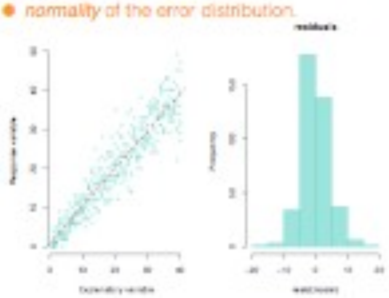
Normality

- *normality* of the error distribution.



Normality

- *normality* of the error distribution.



Normality

- *normality* of the error distribution.

Generalised linear models

GLMs have 3 components

- *linear predictor*
- *link function*

} allows to go beyond linear relationship.

- *variance function* - allows non-normal error distributions.

Generalised linear models

$$\eta = \mathbf{X}\beta + \epsilon$$

"Linear predictor of the response" / Predicted response

Explanatory variable(s)

Parameters

Error

Linear predictor & link function

$$\eta = \mathbf{X}\beta + \epsilon$$

How do we get back to the predicted response?

$$g(\mu) = \eta = \mathbf{X}\beta + \epsilon$$

If $f = g^{-1}$ then $f(\eta) = \mu$

Linear predictor & link function

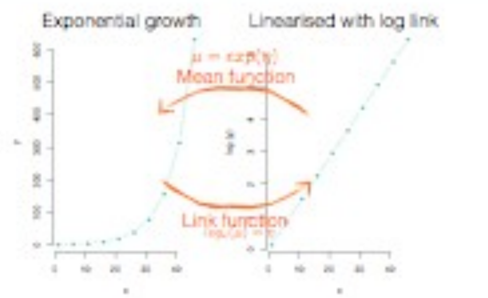
So we have 2 functions, f and g:

link function	mean function
$g(\mu) = \eta$	$f(\eta) = \mu$

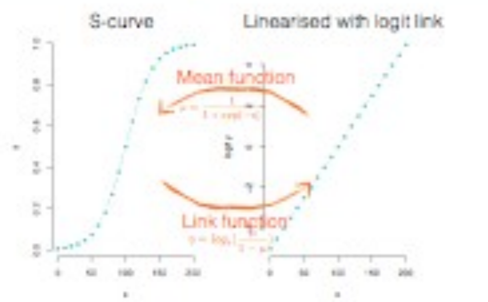
The value of the linear predictor is obtained by transforming the predicted mean using the link function.

The predicted response value is obtained by applying the mean function to the linear predictor.

Log link



Logit link



Variance function

Also called: "Family" or "Error structure"

Describes how variance of response depends on mean - variance is a function of the mean.

- *Normal* - constant variance 
- *Poisson* - variance proportional to mean 
- *Binomial* - variance proportional to $\mu(1-\mu)$ 
- *Gamma* - variances increases faster than linearly 

Choices, choices, choices...

Must choose (1) link (2) error structure:

- Best choice depends on the data.
- Ask yourself:
    - is my response discrete/continuous?
    - what are the bounds?
      (e.g. can I go less than 0, more than 1)
- To make things easier, there are standard links associated with each error structure.

Choices, choices, choices...

Discrete response (e.g. counts):

- poisson $[0,\infty]$, log link.
- binomial $[0,N]$, logit link.

Continuous response (e.g. measurements):

- normal $[-\infty,\infty]$, identity link.
- gamma $[0,\infty]$, inverse link.

Standard link and error

Error distribution	*Bounds*	*Link name*	*Link function*	*Mean function*
Normal	$[-\infty,\infty]$	identity	μ	$\mathbf{X}\beta$
Exponential	$[0,\infty]$	inverse	$-\mu^{-1}$	$-(\mathbf{X}\beta)^{-1}$
Gamma	$[0,\infty]$	inverse	$-\mu^{-1}$	$-(\mathbf{X}\beta)^{-1}$
Poisson	$[0,\infty]$	log	$log_e(\mu)$	$exp(\mathbf{X}\beta)$
Binomial	$[0,N]$	logit	$log_e(\frac{\mu}{1-\mu})$	$\frac{1}{1+exp(-\mathbf{X}\beta)}$
Bernoulli	$[0,1]$	logit	$log_e(\frac{\mu}{1-\mu})$	$\frac{1}{1+exp(-\mathbf{X}\beta)}$

Implementation

Ordinary linear model:

```
model1 = lm(y ~ x, data = mydata)            (simple regression)
model1 = lm(y ~ x1 + x2, data = mydata)      (multiple regression)
model1 = lm(y ~ f1, data = mydata)           (1 - way ANOVA)
model1 = lm(y ~ f1 + f2, data = mydata)      (2 - way ANOVA)
model1 = lm(y ~ x + f1, data = mydata)       (ANCOVA)
```

Generalised linear model:

```
model1 = glm(y~x, data=mydata, family = poisson(link = "log"))
```

Handout

The handout runs through:

1. Examining data to decide on the error family and link.
2. Constructing GLM models for Poisson and Binomial data.
3. Interpreting outputs.
4. Plotting the model.