Day4: GLMs

Generalising the Linear Model

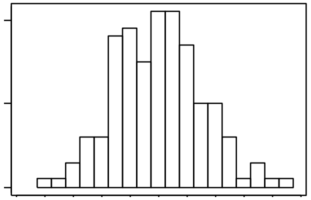
Part I

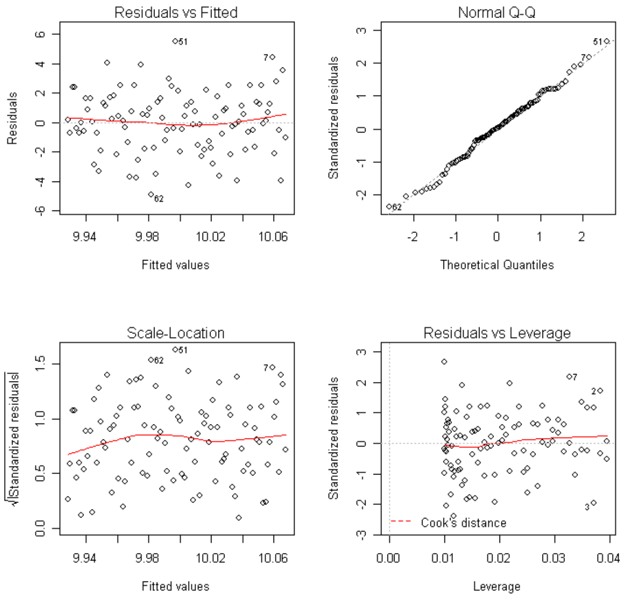
Linear Models

Classical linear models make two important assumptions:

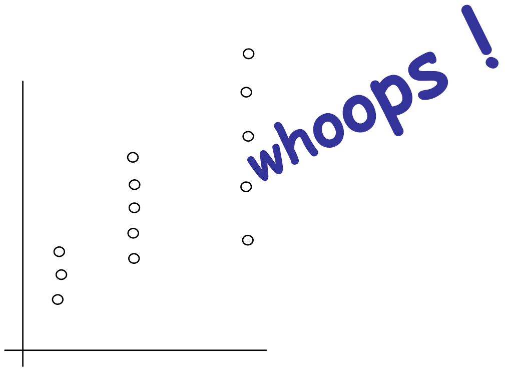
- *__Normal error distribution__* (i.e. Gaussian, or bell-shaped)
- *__Constant variance__*

  (independent of mean) = *homo-sced-a-sti-city*



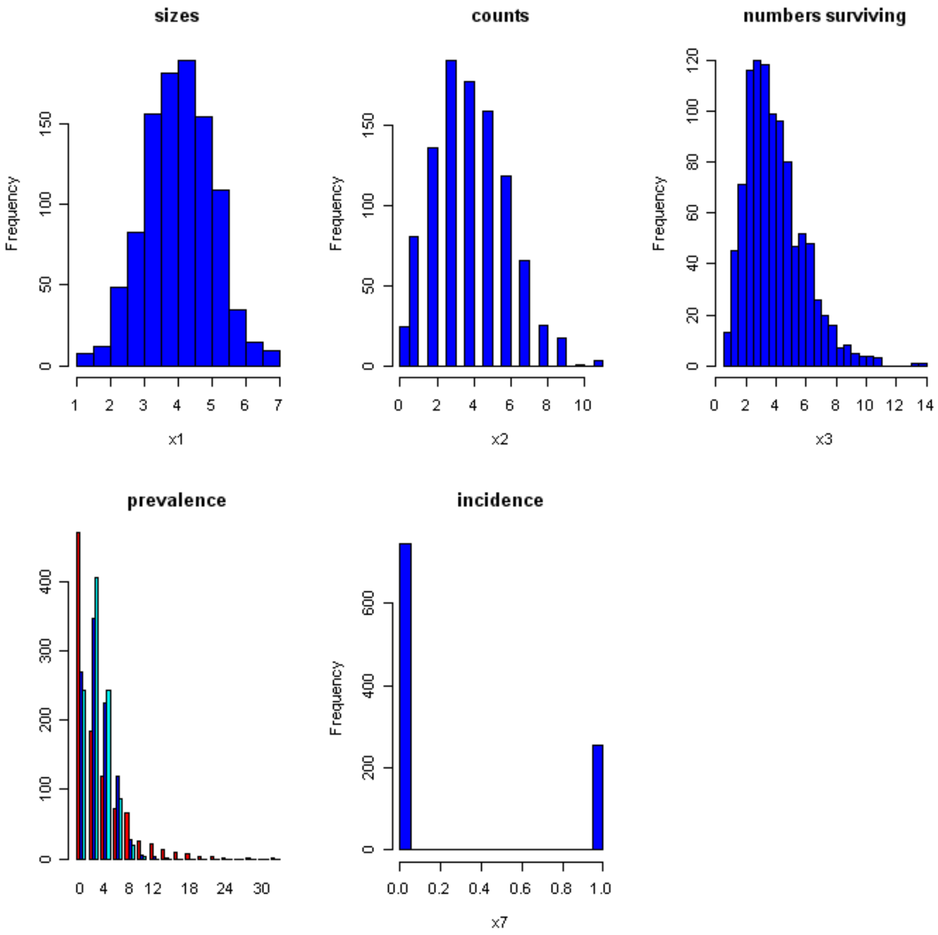


as well as assumptions about independence of data and that X is measured without error – we'll relax these when we come to multilevel models...



Many types of data do not conform to these assumptions!

- Counts
  - Moths in a trap
  - Parasites
- Presence/Absence
  - Survival/Proportions/Binary
  - Incidence
- Timing of Events
  - Death, flowering, maturity



What do we do?

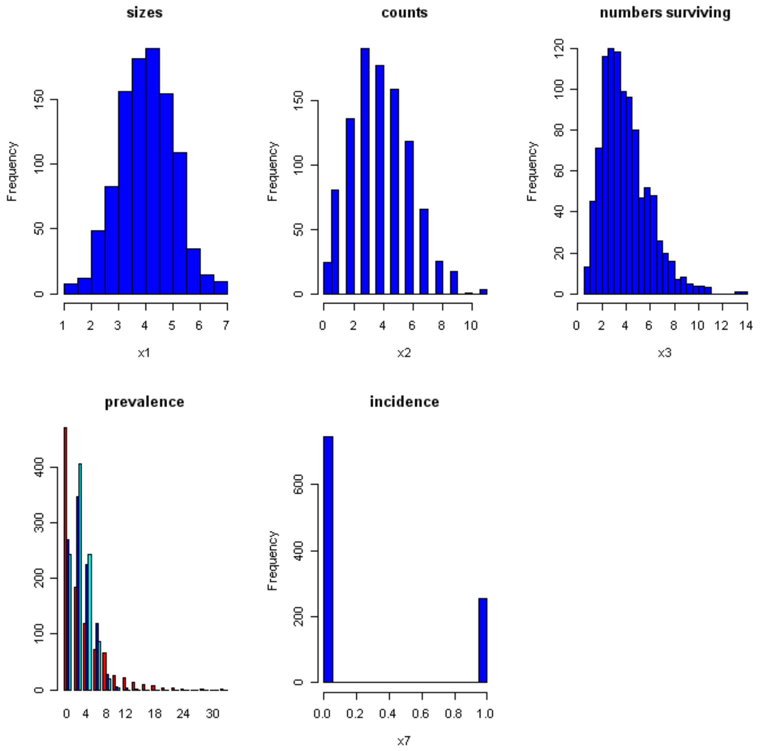
What do we do? We generalise the LM

- *Generalised linear modelling*
  - *for data producing non-Normally distributed errors*.
- Uses the appropriate distribution of error (the residuals) to fit the model and get probabilities
- Different distributions have different variance properties (e.g. gamma has var increasing with the mean$^{a>1}$)

Gaussian

Poisson

gamma/ exponential

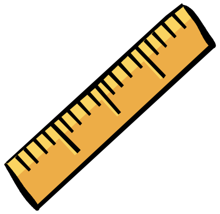


neg. bin.

binomial

Distributions

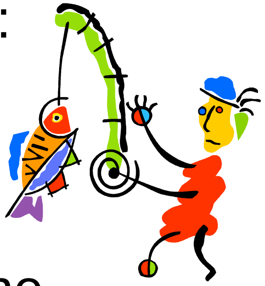
- A **normal** (Gaussian) distribution arises when many random values are *added* together (Central Limit Theorem)
- A **lognormal** distribution arises when many random values are *multiplied* together



A Bernoulli trial is a random event with a (1/0) outcome:

- A **binomial** distribution models the number of 1's from a fixed number of Bernoulli trials (if lots, → **normal**).
- A **negative binomial** distribution models the number of trials before a specified number of 1s is achieved (**geometric** distribution is special case for the *first* 1)

A Poisson process generates independent events in time or space:



- A **Poisson** distribution models the number of events occurring in a given interval (if common, → **normal**).
- A **gamma** distribution models the time taken to get a specified number of events (**exponential** distribution is special case for the *first* event).

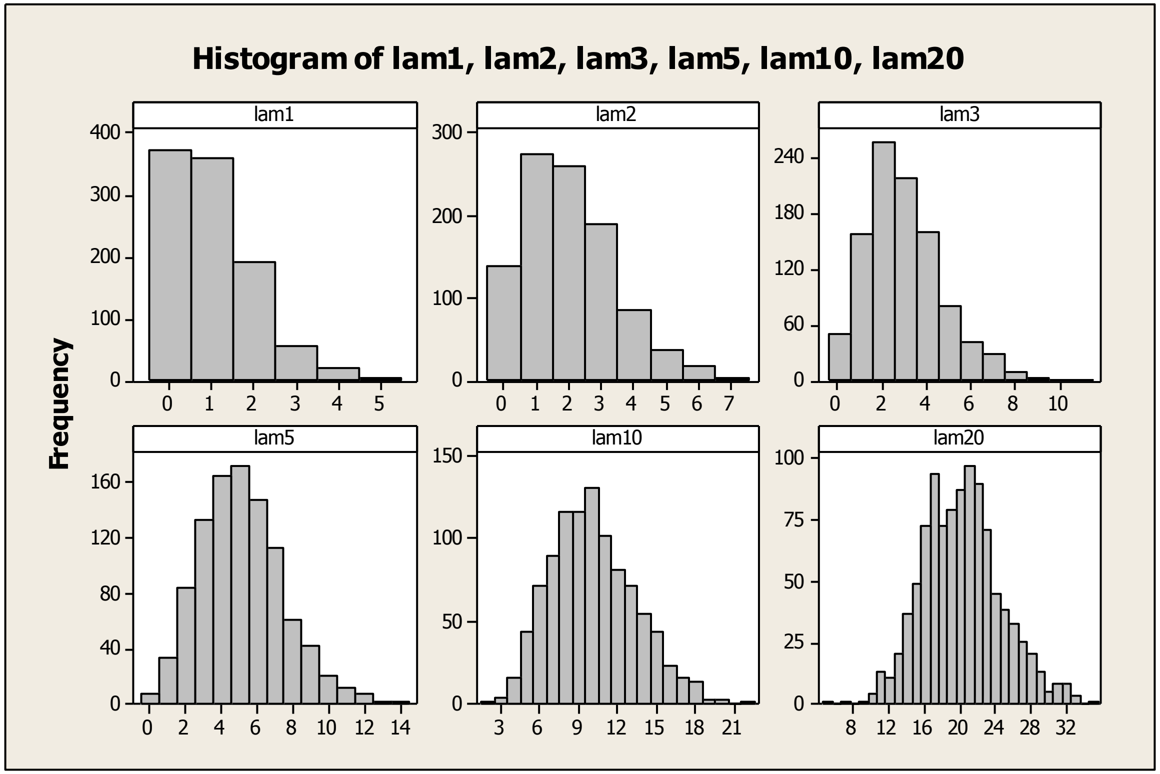
Example: Poisson

Normal has two parameters: μ (mean), σ (SD)

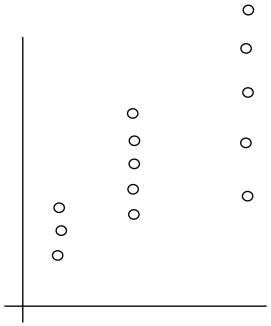
Poisson has only one: λ = mean = var

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$



Components of a GLM



$$data \xleftrightarrow{link\_function} linear\_predictor + error$$

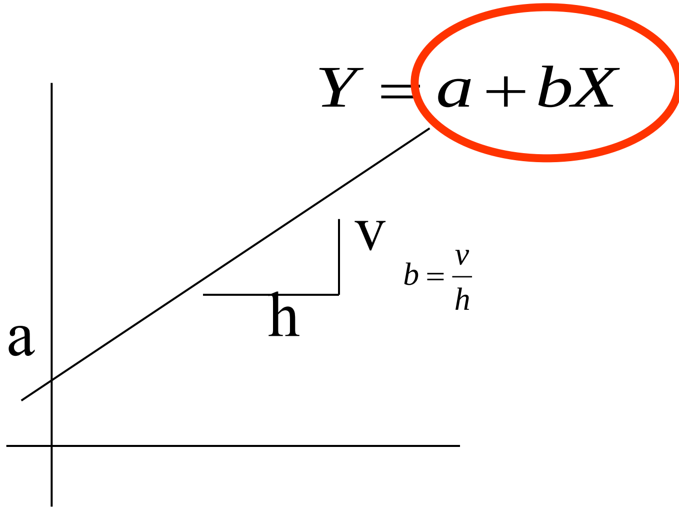
A generalised linear model has three important components:

- **1. the linear predictor (the equation)**
- **2. the link function**
- **3. the error distribution**

N.B. In general linear models (LMs) the link function is assumed to be the “identity” link and error is Gaussian.

1. Linear Predictor

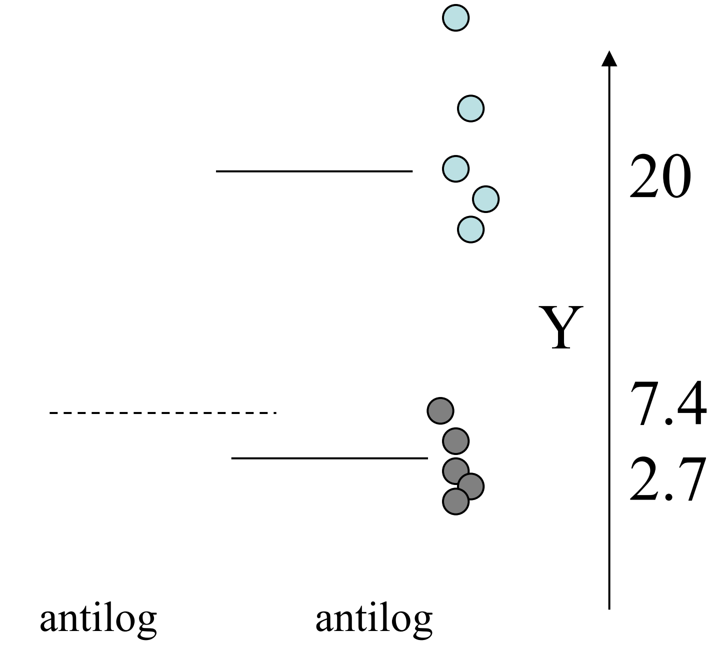
- The linear predictor is the equation predicting the value of ***Y*** on a *transformed* scale determined by the **link function**.



$$data \xleftrightarrow{link\_function} linear\_predictor + error$$

$$Y \xleftrightarrow{link\_function} a + bX_1 + cX_2 + \varepsilon$$

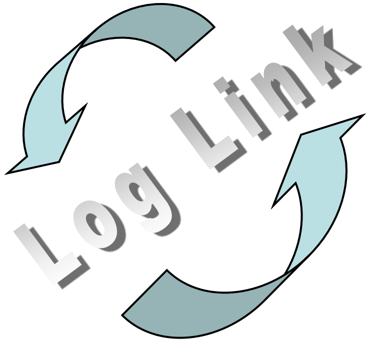
e.g. bird counts in site 1 and site 2

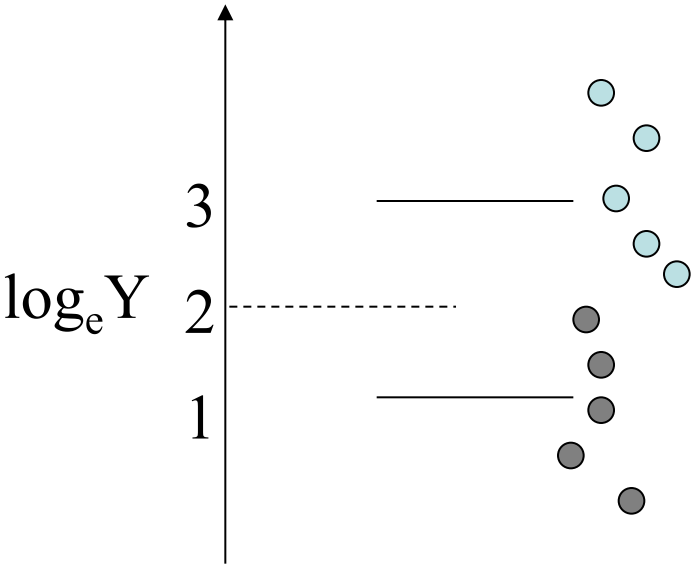


antilog antilog

Intercept ± coef ± error

(not Gaussian…)





Intercept ± coef ± error

2. Link Functions:

linking the expected value of Y, μ, *to a linear predictor*, η

So: **Mean + Treatment effect + Error** $= \eta$ and:

Link function	Formula	Use
Identity	$\eta = \mu$	*Normal errors* (classic LM)
Log	$\eta = \log \mu$	Count data: *Poisson* or *neg.bin. errors* (log-linear models)
Logit	$\eta = \log (\mu / n - \mu)$	Proportion or binary data: *binomial errors*
Reciprocal	$\eta = 1 / \mu$	Continuous data: *gamma errors*
Probit	$\eta = \Phi^{-1} (\mu / n)$	Proportion or binary data (bioassays)
Complementary log-log	$\eta = \log[-\log(1 - \mu / n)]$	Proportion data (dilution assays)
Square root	$\eta = \sqrt{\mu}$	Count data
Exponent	$\eta = \mu^{a}$	Power functions

2. Link Functions:

choosing one

- The most important criterion is to choose a link function that ensures that fitted values stay within the possible range. E.g.:
  - If analysing count data, we’d know they have to be positive (you can’t have a negative count!).
  - This could be achieved by using a log link, so the predicted values will be antilogs of the linear predictor, therefore > 0.
- Statisticians may try a range of different link functions and compare their fits.
- But in practice, it is usually sufficient to use the **default option** for a given error distribution

  (see previous slide).

3. Error Distributions

Type of data	Example	Error distribution
Metric data	body weight, height, tail length	Normal* (Gaussian)
Count data	moths in traps, daisies per quadrat, number of prey per predator, flock size etc (something that occurs with constant probability in time or space)	Poisson *(arrivals per time; counts per unit area)*
Over-dispersed count data	worms per rabbit, mites per swallow, generally: parasites and many zero's	negative binomial *(trials between successes)*
Binary	nestlings surviving, sex ratios, animals doing different behaviours, gene frequencies	binomial *(yes/no trials)*
Survival times	mortality rates, time to death (where Pr(death) is age-independent)	exponential (*waiting time till success*)
Metric data	if variance increases with the mean e.g. variance in body size increases with population density, functional responses (any relationships described by inverse polynomials)	gamma

*Normal comes from additive normal processes, log-normal from multiplicative

Fitting GLMs

Remember: classical stats asks how likely the data would be for a given model. So we always fit a model to have the maximum likelihood of producing our observed data.

- Classical LMs have a simple algorithm for this: find parameters to minimise the error variance i.e. sum of squares (hence technique is “least squares”).
- But this only works for normal errors. So:
- GLMs use other ways to *maximise the likelihood*...

Fitting GLMs

Maximum likelihood

$$L(\theta \mid x) \propto P(x \mid \theta)$$

The likelihood (L) for a set of parameters in a model (θ), given the data you have (x), just means the probability of sampling those data given those parameters.*

Maximum likelihood occurs with the set of parameters which maximise the probability of the actual data being observed.

*which we can easily calculate, using the appropriate underlying probability model (PDF)

Fitting GLMs

In GLMs, parameter estimates have to be obtained by an *iterative* method (e.g. Newton-Raphson).

1. an initial guestimate of the parameter values is taken.
2. This is then changed slightly (up or down) and the likelihood of the new model is compared with that of the previous one.
3. The better fit is retained and then compared with a new estimate.
4. This procedure is repeated iteratively until the fit of the model is within specified limits (the “tolerance”).

Diagnostics

Residuals in Poisson and Binomial models: What do we expect?

- For a well-fitting model, the residual deviance should be similar to the residual d.f.
- i.e. {residual deviance} / {residual d.f.} should be ≈ 1
- called the "scale parameter" (measure of dispersion)

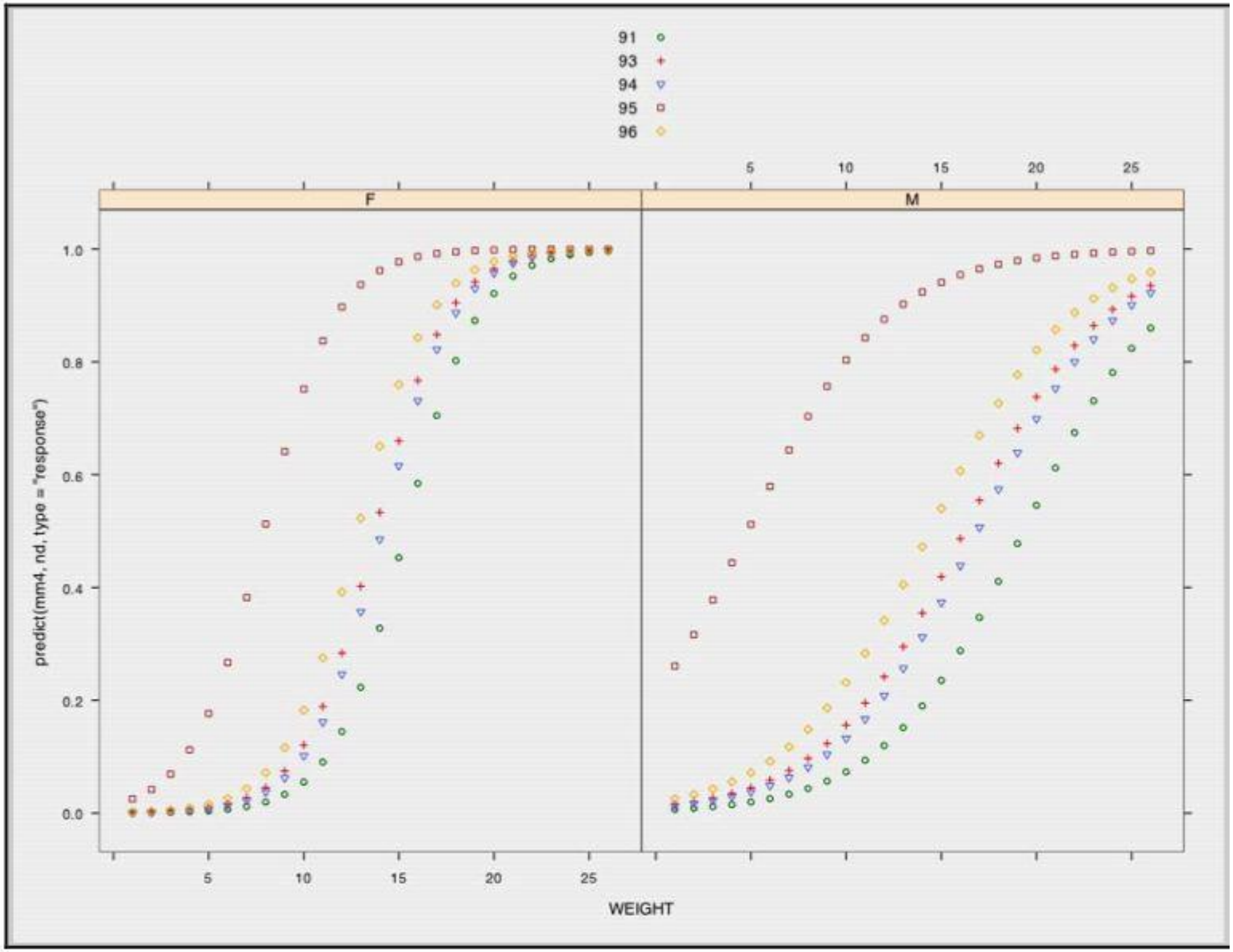
Diagnostics

```
> summary(counts.glm)
```

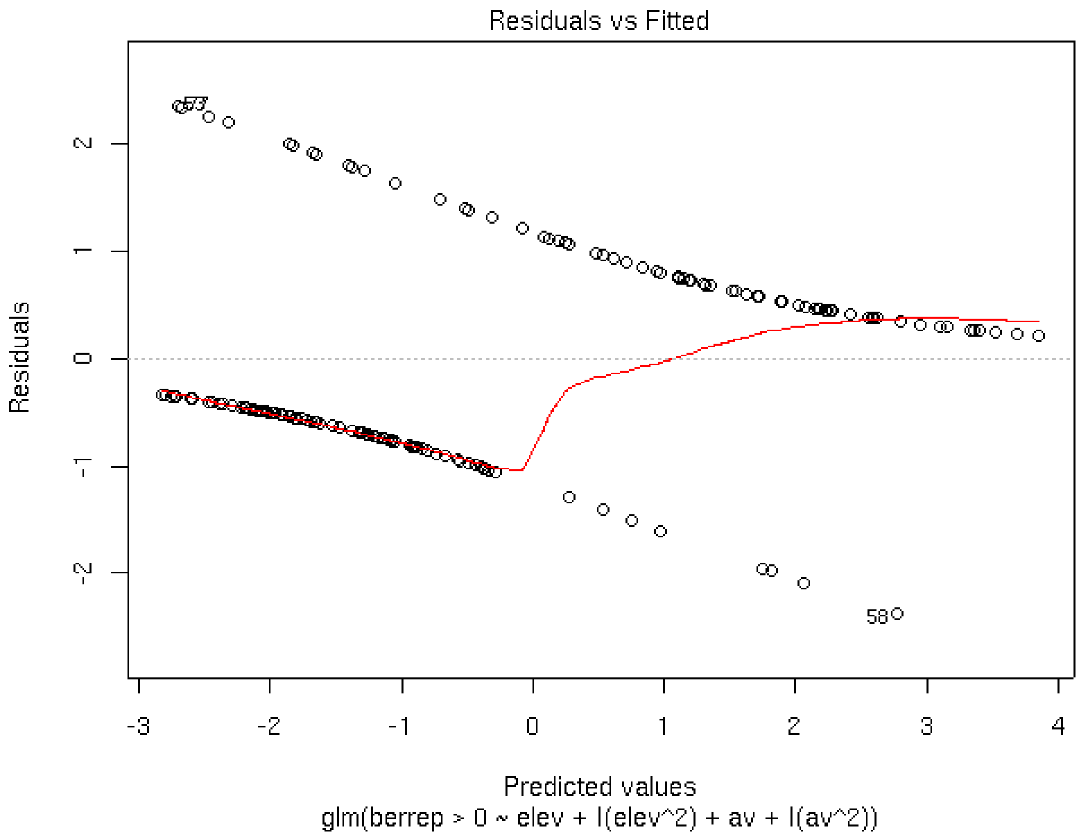
```
glm(formula = counts ~ outcome + treatment, family = poisson())
Coefficients:
              Estimate Std. Error   z value Pr(>|z|)
(Intercept)  3.045e+00  1.709e-01    17.815   <2e-16 ***
outcome2    -4.543e-01  2.022e-01    -2.247   0.0246 *
outcome3    -2.930e-01  1.927e-01    -1.520   0.1285
treatment2  -2.263e-16  2.000e-01 -1.13e-15   1.0000
treatment3  -1.251e-16  2.000e-01 -6.26e-16   1.0000
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 10.5814  on 8  degrees of freedom
Residual deviance:  5.1291  on 4  degrees of freedom  ☺
AIC: 56.761
```

Logistic regression

Logistic regression



Logistic regression



Method Consistency with LMs

- Plot
- Fit a model
- Diagnostics
- Scale / Dispersion
- Stick with or change model

Summary

- GLMs are a generalisation of the LM to cope with non-normal errors
- No reason NOT to do GLMs
  - May seem complex now, but this method is much better than “traditional” stats and is MORE likely to get the right answer than by the traditional route of TRANSFORMING the data or (heaven forbid!) doing a non-parametric test!

Practical

- Day GeneralisedLMs

(material for today and ….. plenty)