

Generalized Linear Modelling

Evol Ecol (2011) 25:1179–1195
DOI 10.1007/s10682-011-9469-7

ORIGINAL PAPER

Maternal and paternal contributions to egg size and egg number variation in the blackfin pearl killifish *Austrolebias nigripinnis*

Mahmoud Moshgani • Tom J. M. Van Dooren



(La Paloma Uruguay)



Generalized Linear Modelling

Evol Ecol (2011) 25:1179–1195

1183

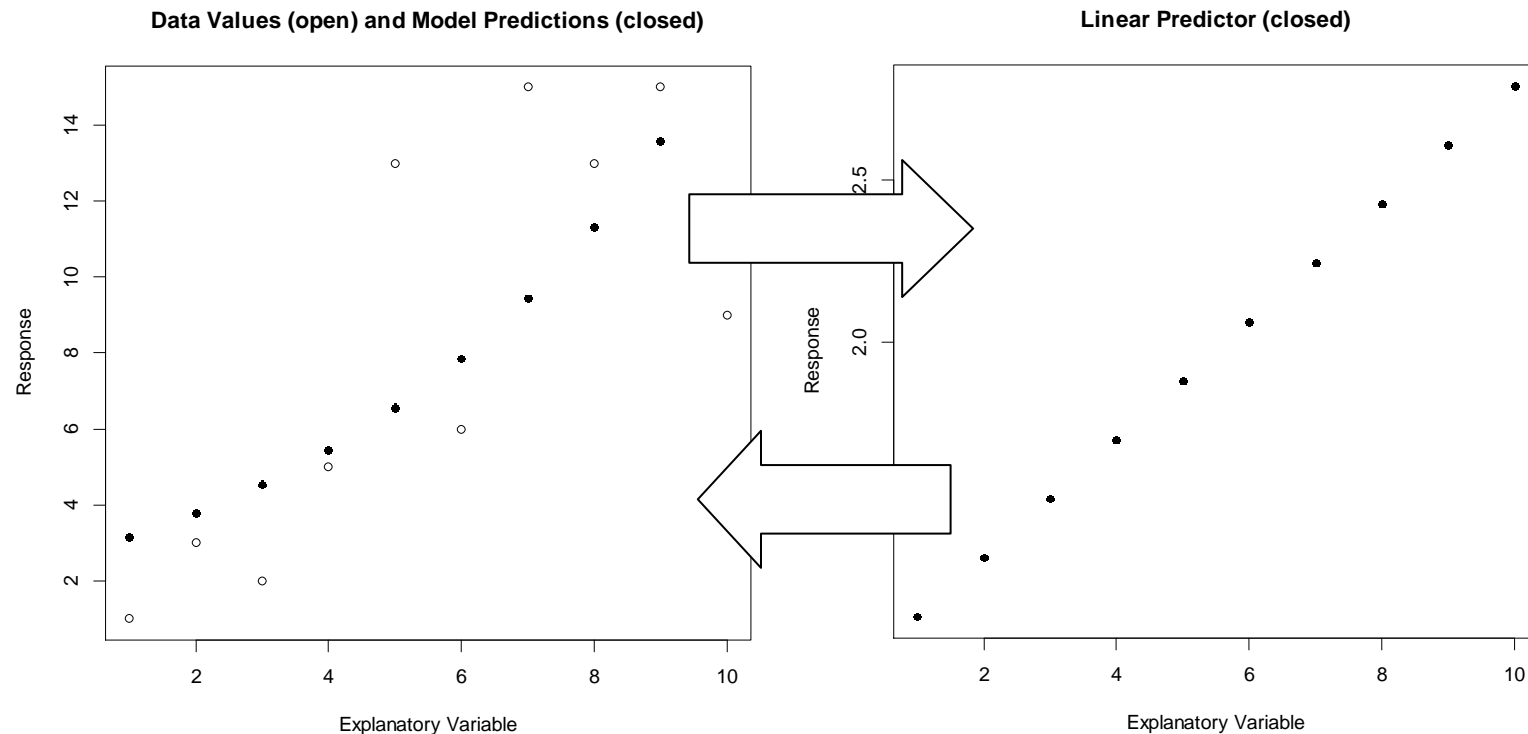
Table 1 Maximal model fitted to each dependent variable

Grouping factor	Description (number of parameters)
Fixed effect	Age (1) Weekday (3, Monday–Wednesday–Friday) Age \times Weekday Interaction (3)
Random effect	
Females	Age (1)
Females	Weekday (3)
Random effect	
Males	Age (1)
Males	Weekday (3)
Random effect	
Female \times male combination	Intercept (1)
Random effect	
Tank	Intercept (1)
Error term	(1, only for Reproductive effort and egg size)

The model contains fixed effects, random effects, and for the reproductive effort and egg size traits also an error variance

Random effects can be estimated if there are repeated observations per level of a grouping factor (fish tank, females, males, or female-male combinations). The weekday variable is categorical with three levels. For the fixed effects, there is a separate parameter estimated per weekday level, for the random effects, a separate variance component is estimated per weekday

```
X<-c(1:10); y<-rpois(10,.7+1.3*x); # partial script!  
predict(glm(y~x,poisson),type="response"); predict(glm(y~x,poisson)
```



→ "linearize" model predictions – link function

← transform a linear model to match data properties – inverse link

Viewed from the side of the explanatory variables:

The linear predictor

The structure of the model relates each observed value y_i of the response to a predicted value η_i

$$\eta_i = \sum_{j=1}^p x_{ij} \beta_j \quad \text{or in matrix notation} \quad \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

x_{ij} are the values of the j -th explanatory variables for observation i .

$(x_{ij})_{ij}$ is called the model matrix

η_i is linear in the coefficients β_i which will be jointly estimated by ML

In the middle: the link function h

Relates the linear predictor to the mean (expectation) of the response

$$E(y_i) = \mu_i \qquad h(\mu_i) = \eta_i \qquad \mu_i = h^{-1}(\eta_i)$$

The linear predictor can in principle produce values in $]-\infty, +\infty[$, the range of the response is often constrained.

Inserting a non-linear link function h can solve this, so that a linear model can be fitted to such data.

Some links are easier to handle mathematically, they are called *canonical* link functions.

Link function h

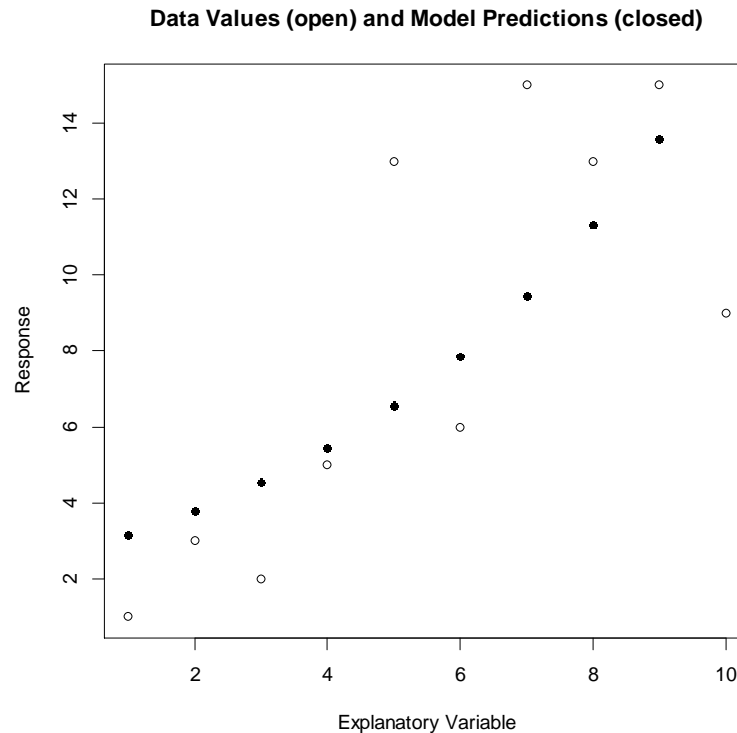
Relates the linear predictor to the mean (expectation) of the response

$$E(y_i) = \mu_i \qquad h(\mu_i) = \eta_i \qquad \mu_i = h^{-1}(\eta_i)$$

Examples of canonical links for two particular distributions, which also solve the "range" issue:

Counts cannot be negative, so h^{-1} has to have a positive range
Using the log function for h gives this property

Proportions vary between zero and one. Using the logit function accommodates this: $h(\mu_i) = \ln \frac{\mu_i}{1-\mu_i}$



The response values in the data are a set of observed values of random variables

Each random variable follows a probability distribution

We can calculate the likelihood of a model given the data

An approach that works for more than a single family of distributions for the response

model response variables in **exponential dispersion family**

observations are independent given explanatory variables
the contribution to the log-likelihood of observation y_i is

$$\ln[f(y_i | \theta_i, \phi)] = \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(\phi, y_i)$$

normal, Poisson, binomial, gamma

In this family, a model (θ_i, ϕ) specifies what the prediction is for each y_i

θ_i is called the canonical parameter, it varies between observations
 ϕ is the dispersion parameter

Table 15.9 Functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ for Constructing the Exponential Families

<i>Family</i>	$a(\phi)$	$b(\theta)$	$c(y, \phi)$
Gaussian	ϕ	$\theta^2/2$	$-\frac{1}{2} \left[y^2/\phi + \log_e(2\pi\phi) \right]$
Binomial	$1/n$	$\log_e(1+e^\theta)$	$\log_e \binom{n}{ny}$
Poisson	1	e^θ	$-\log_e y!$
Gamma	ϕ	$-\log_e(-\theta)$	$\phi^{-2} \log_e(y/\phi) - \log_e y - \log_e \Gamma(\phi^{-1})$
Inverse-Gaussian	ϕ	$-\sqrt{-2\theta}$	$-\frac{1}{2} \left[\log_e(\pi\phi y^3) + 1/(\phi y) \right]$

NOTE: In this table, n is the number of binomial observations, and $\Gamma(\cdot)$ is the gamma function.

a gives the dispersion parameter in this table

$$\ln[f(y_i | \theta_i, \phi)] = \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(\phi, y_i)$$

Properties of distributions in this family are

$$E(y_i) = \mu_i = b'(\theta_i)$$

$$Var(y_i) = \phi b''(\theta_i)$$

Since $\mu_i = b'(\theta_i)$ there is an implied relationship between the canonical parameter and the parameters in the linear predictor. If we choose a link such that $h(\mu_i) = \theta_i$ then this h is called the canonical link

Likelihood of a model given the data



Log-likelihood of a model given the data

Find parameter values of a model that maximize the probability of observing the dataset

When observations in a dataset are independent we can sum their log-likelihoods

Maximum Likelihood estimates $\hat{\beta}$

In general The coefficients in parameter vector $\hat{\beta}$ are maximizing the likelihood of the data given the model. This is a multi-parameter problem.

Often there are no closed form expressions for the elements of $\hat{\beta}$

The ML values are found using numerical approaches

we have a function that we want to maximize

initial estimate \rightarrow updating formula \rightarrow convergence to a maximum

[Expectation-Maximization methods, ML Martin-Magniette]

General scheme: maximize the likelihood in all parameters using a general-purpose numerical optimizer. Get the estimates, the Hessian at the estimated values.

glm() uses a routine for the exponential dispersion family,

Iteratively Reweighted Least Squares

<https://www.stat.wisc.edu/courses/st771-newton/papers/green.pdf>

IRLS/IWLS -Faster - some properties of statistics in the output are different from the general-purpose optimizer.

Inspect what happens:

library(gnrm)

gnlr # general purpose

glm.fit # IRLS

A very simple example
with simulated data

```
# a single independent variable x  
x<-c(1:20) # ~ continuous
```

```
# response variable y  
# regression model: response = intercept + slope  $\times$  x + error
```

```
y<-7+0.3*x+rnorm(20,mean=0,sd=0.1)
```

2. Fit

```
modeln1<-glm(y~x) # our assumptions in 1 are the default settings  
?glm
```

3. Output

`summary(modeln1)`

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.99438	0.04972	140.68	<2e-16 ***
x	0.30206	0.00415	72.78	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.01145456)

Null deviance: 60.88229 on 19 degrees of freedom

Residual deviance: 0.20618 on 18 degrees of freedom

AIC: -28.737

Properties of the IWLS algorithm and the results:

It converges rapidly and reliably, with well chosen initial estimates

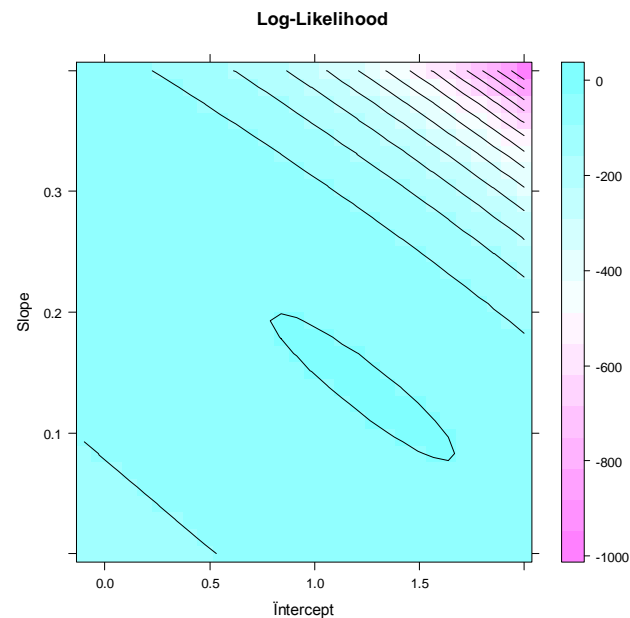
The estimation of β does not depend on the dispersion parameter. That is estimated separately as (or estimated differently using the deviance).

$$\hat{\phi} = \frac{1}{n - p} \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

where V is the variance function: b'' seen as a function of μ .

Log-Likelihood surface

```
x<-c(1:10); y<-rpois(10,.7+1.3*x)
pois.LL<-function(y,m,s){
  LL <- sum(log(dpois(y, m)))
  LL
}
int<-rep(seq(-0.1,2,length.out=30),each=30); slope<-
rep(seq(0,0.4,length.out=30),times=30)
z<-numeric();for(i in 1:length(int))z[i]<-pois.LL(y,exp(int[i]+slope[i]*x))
levelplot(z~int*slope,contour=T,xlab="Intercept",ylab="Slope",main="Log-
Likelihood")
```



Standard deviations of parameter estimates (standard error)

At the ML solution, the diagonal of the inverse of the observed Fisher information matrix contains the variances of parameter estimates

```
# standard deviations of parameter estimates  
sqrt(diag(summary(model)$cov.scaled))
```

Observed Fisher information matrix:

The negative of the second derivative (the Hessian matrix) of the log-likelihood as a function of the parameters.

A measure of the curvatures of the log-likelihood

If the dispersion parameter is estimated the estimates of β divided by their estimated standard deviation are (nearly) student-t distributed

If the dispersion parameter is fixed the estimates of β divided by their known standard deviation are (nearly) normally distributed

```
yp<-rpois(length(y),exp(y))  
modelp1<-glm(yp~x,poisson)
```

```
summary(modelp1)  
# comparison of coefficients
```

```
modeln1$coef  
modelp1$coef
```

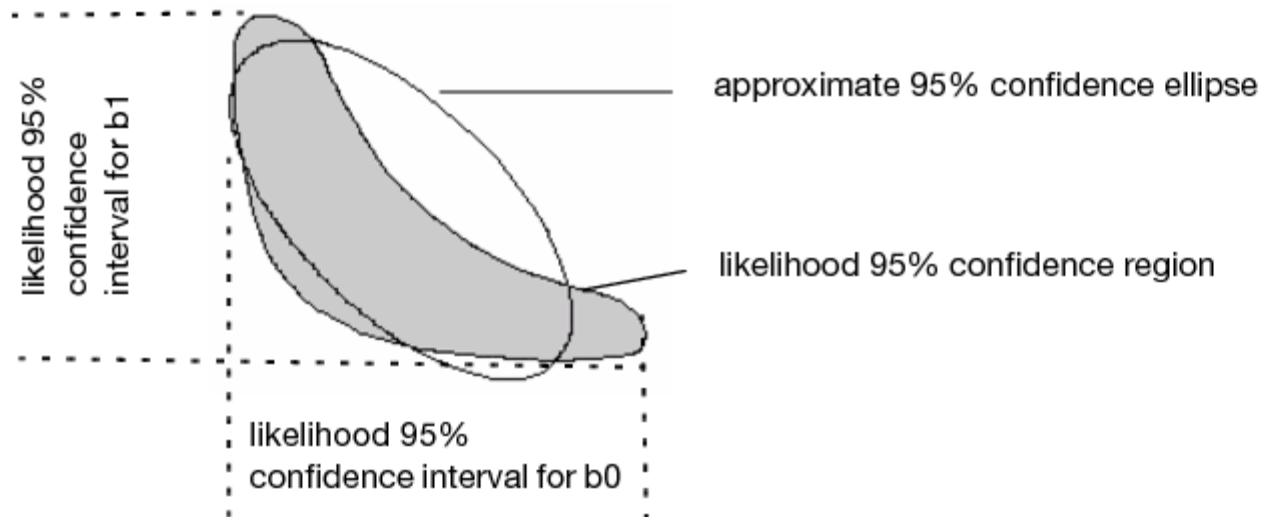
These standardizes ML estimates of β

can be used to construct **symmetric** confidence intervals ♣ assuming that the sampling probability distribution of the parameter has a near-gaussian shape at the ML estimates (Wald-type CI ♣)

Otherwise use **profile likelihood** intervals ♣, most notably when parameter estimates are on the boundary of their allowed range.

Check the pdf of the **Inference** lecture for details or the pdf's at the end of this lecture

Shapes of confidence regions can differ between Wald and profile likelihood CI.



Generalized Linear Modelling

General approach per model

1. **Choose**

linear predictor

family of probability distributions (error structure)

link function

2. **Fit**

fit model to data using Maximum Likelihood by IWLS

3. **Output**

parameter estimates

standard errors of parameter estimates

residuals

deviance

4. **Check** model assumptions

3. Output

`summary(modeln1)`

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.99438	0.04972	140.68	<2e-16 ***
x	0.30206	0.00415	72.78	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.01145456)

Null deviance: 60.88229 on 19 degrees of freedom

Residual deviance: 0.20618 on 18 degrees of freedom

AIC: -28.737

An intercept of the regression is automatically fitted ($y = a + bx$)

Each t -statistics is for H_0 : the parameter estimate is equal to zero! ♣

Hypothesis testing

Null Hypothesis - Alternative Hypothesis

General procedure: assuming the Null is true, what is the probability that the value of the statistic occurs? → Tail probability of a test (p -value)

Type I error: The Null hypothesis is true, but we reject it in favour of the alternative. The frequency at which this happens is controlled to be **not more** than α . (if we can't control exactly, we will be **conservative**)

Type II error: The alternative hypothesis is true, but we don't reject the null hypothesis. This occurs with probability β . Power: $1 - \beta$

Deviance:

assuming $\phi = 1$ to calculate likelihoods

$$D(\mathbf{y}, \boldsymbol{\mu}) = 2 \ln \left(\frac{L(\mathbf{y}|\mathbf{y})}{L(\boldsymbol{\mu}|\mathbf{y})} \right)$$

with $\boldsymbol{\mu}$ is a model for the mean of data vector \mathbf{y} , a model for which $\phi = 1$

(residual) deviance: $-2 \times$ (difference between the log likelihood of a given model and that of a saturated model)

(the log-likelihood of the saturated model can be zero)

null deviance: $-2 \times$ (difference between the log likelihood of the simplest possible model and that of a saturated model)

Scaled Deviance: $\phi \neq 1$

$$D^*(\mathbf{y}, \boldsymbol{\mu}) = D(\mathbf{y}, \boldsymbol{\mu}) / \phi$$

if $\boldsymbol{\mu}$ is a model for the mean of data vector \mathbf{y} and ϕ a parameter which is known.

In practice ϕ will be estimated as $\hat{\phi}$ and plugged in the equation for scaled deviance

Scaled deviance is the likelihood ratio for distributions where $\phi \neq 1$

Kullback Leibler Information

A measure of the distance between the true data generating distribution g and the distribution implied by a specific approximate model f .

$$I(f, g) = \int dy g(y) \ln \frac{g(y)}{f(y | \theta)}$$

*Advance 2 – Estimation of Kullback–Leibler information
(AIC)*

Akaike (1973, 1974) found a formal relationship between K–L information (a dominant paradigm in information and coding theory) and maximum likelihood (the dominant paradigm in statistics) (see deLeeuw 1992). This finding makes it possible to combine estimation (e.g. maximum likelihood or least squares) and model selection under a single theoretical framework – optimisation. Akaike's breakthrough was the finding of an estimator of the expected, relative K–L information, based on the maximised log-likelihood function. Akaike's derivation (which is for large samples) relied on K–L information as averaged entropy and this lead to 'Akaike's information criterion' (AIC),

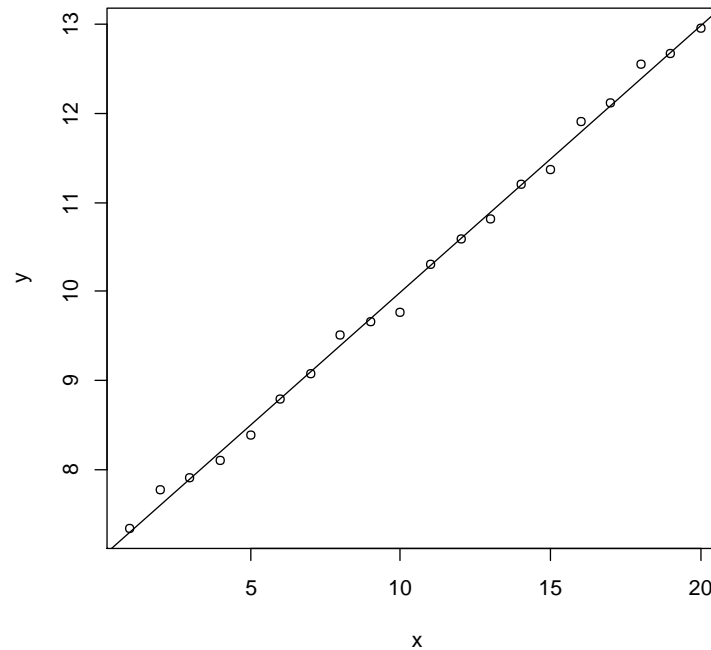
$$AIC = -2\log_e(L(\hat{\theta} \mid data)) + 2K,$$

where $\log_e(L(\hat{\theta} \mid data))$ is the value of the maximised log-likelihood over the unknown parameters (θ), given the data and the model, and K is the number of estimable parameters

Plot the data and the fitted model

`plot(y~x)` # response and explanatory variable

`abline(modeln1)` # regression line



4. Check

Different residuals can be calculated from a glm:

Response residual Difference between fit and data on response scale

Deviance residual Measure of deviance contributed by each observation

Working residual From last iteration of IWLS fit

Standardized residual all residuals are rescaled to an identical variance, using leverages

Pearson residual rescaled by variance function $[\text{variance}(\text{mean})]$

Hat matrix: maps the vector of observed values to the vector of fitted values. $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$

The diagonal elements of the hat matrix are called **leverages**, a measure of influence of an observation on the overall outcome

$1/\text{leverage}$ is called **effective replication** ~ number of observations providing information on each focal one.

Cook's distance uses the leverage to calculate the change to the solution (model estimates) that result from omitting each observation. A distance is thus calculated for each observation

4. Check

`plot(modeln1)`

4a Plot of residuals against fitted values

Is there a systematic pattern in the residuals as a function of the predictions?

4b Scale-Location plot of $\sqrt{\text{residuals}}$ against fitted values

Is the variance of the error term independent of the predicted value?

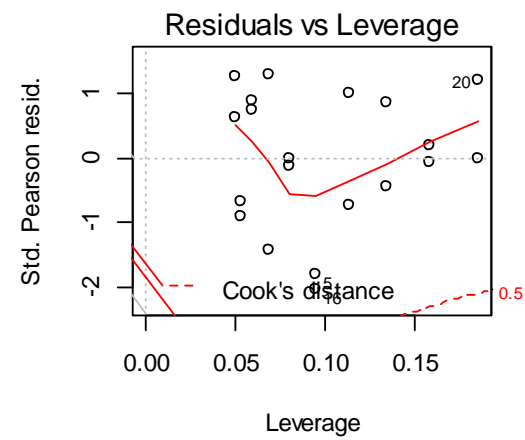
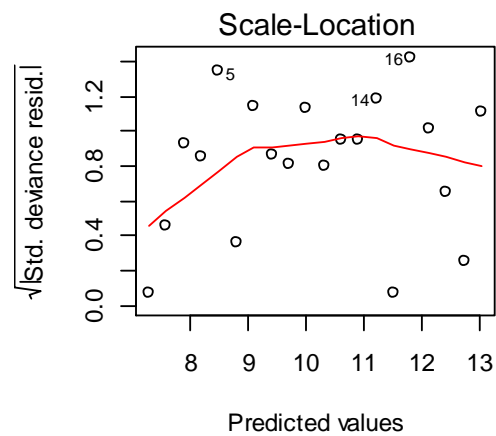
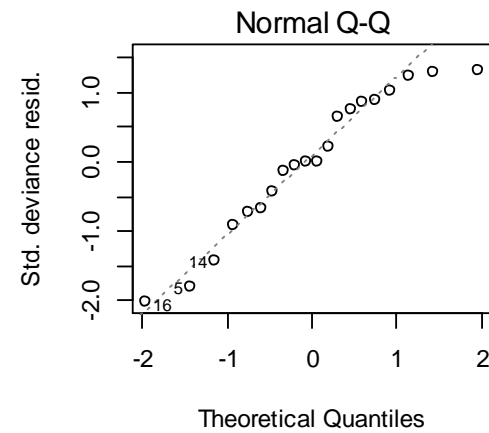
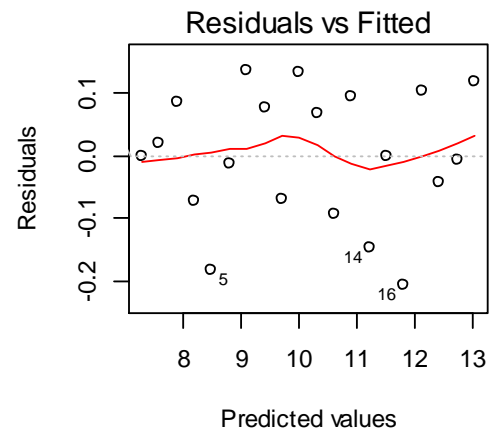
4c Normal Q-Q plot

Are the standardized Pearson residuals approximately normally distributed?

4d Plot of Cook's distances versus row labels or a funnel graph.

Detecting observations with a very large effect on model fit

4. Check



5. Predict

`fitted(modeln1) # fitted values`

1	2	3	4	5	...
7.296440	7.598503	7.900566	8.202630	8.504693	...
17	18	19	20		
12.129453	12.431517	12.733580	13.035643		

`predict(modeln1) # fitted values`

Predict what y is expected to be for new observations of x

`predict(modeln1,newdata=data.frame(x=c(30:40))) # extrapolation`

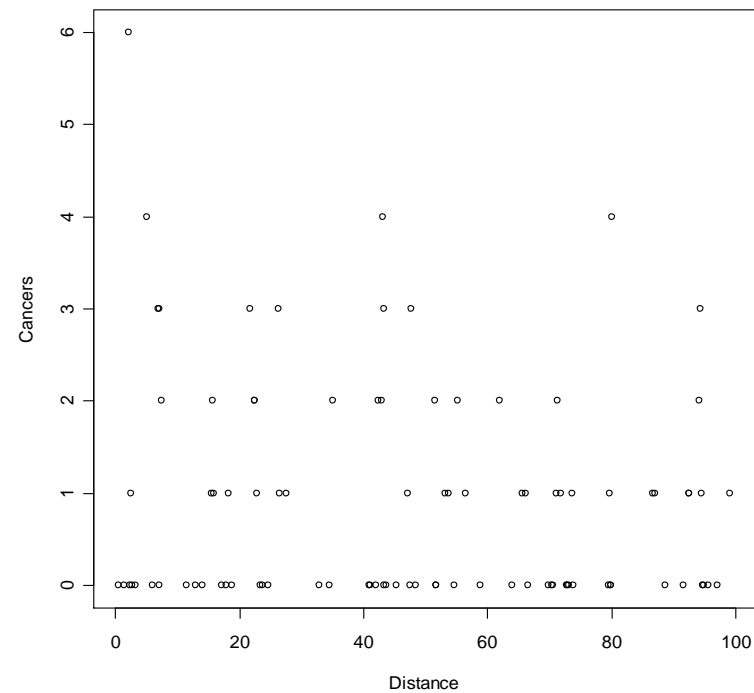
1	2	3	4	...
16.05628	16.35834	16.66040	16.96247	...
9	10	11		
18.47278	18.77485	19.07691		

`simulate(modeln1,3)` # simulate 3 datasets according the normal distribution model

	<code>sim_1</code>	<code>sim_2</code>	<code>sim_3</code>
1	7.252535	7.191326	7.323199
2	7.609283	7.603341	7.563057
3	7.846449	7.799961	7.971919
4	8.252368	8.256203	8.165220
5	8.316619	8.507197	8.388832
6	8.713656	8.692125	8.659978
7	9.082239	9.069606	9.048867
8	9.434953	9.369076	9.386736
9	9.746741	9.694746	9.777400
10	10.134374	10.010319	9.973842
11	10.365654	10.252986	10.223155
12	10.620147	10.526076	10.547619
13	10.907874	10.940264	10.887334
14	11.302542	11.156580	11.199833
15	11.608532	11.548315	11.569124
16	11.866806	11.683517	11.803395
17	12.056659	12.196530	12.144277
18	12.280300	12.466481	12.508674
19	12.777042	12.664291	12.479373
20	13.200004	12.826396	12.949221

Count data

- never negative
- the variance increases with the mean
- the observations are integers



General approach

1. Choose

linear predictor $\eta_i = \sum_{j=1}^p x_{ij} \beta_j$

family of probability distributions: **Poisson**

canonical link function g : **log** function

The predicted mean of the Poisson response variable for observation i is μ_i

$$\eta_i = h(\mu_i) = \log(\mu_i) \Leftrightarrow \mu_i = e^{\eta_i}$$

$\rightarrow \mu_i$ is always non-negative

2. **Fit** model to data using Maximum Likelihood / IWLS

```
modelpcancer<-glm(Cancers~Distance,poisson)
```

3. **Output**

```
summary(modelpcancer)
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.186865	0.188728	0.990	0.3221
Distance	-0.006138	0.003667	-1.674	0.0941 .

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 149.48 on 93 degrees of freedom

Residual deviance: 146.64 on 92 degrees of freedom

4. Check model assumptions

For a Poisson distribution, the mean is equal to the variance.

Sometimes, the variance can be inflated relative to the mean, called **overdispersion**. It requires fitting a dispersion parameter and corrections to statistics in hypothesis testing.

```
modelqpcancer<-glm(Cancers~Distance,quasipoisson)
```

```
summary(model2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.186865	0.235341	0.794	0.429
Distance	-0.006138	0.004573	-1.342	0.183

(Dispersion parameter for quasipoisson family taken to be **1.554966**)

Null deviance: 149.48 on 93 degrees of freedom

Residual deviance: 146.64 on 92 degrees of freedom

AIC: NA

Causes of Overdispersion

Observations are samples from groups which vary randomly in parameter values of their probability generating mechanism. This adds extra heterogeneity to the response.

Causes of Underdispersion

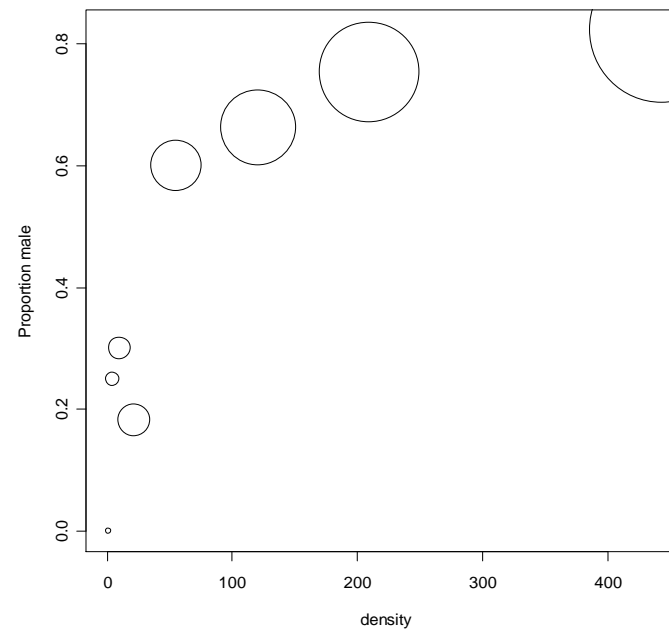
There is some additional regularity, for example, observations are always summed over identical groups (which are not among the independent variables).

- Why not always correct for it? The amount of overdispersion has to be estimated, an estimate > 1 does not mean there really is any.

[There are probability distributions where over- and underdispersion are allowed and are represented by a parameter, e.g. beta-binomial, double binomial, Conway-Maxwell-Poisson]

Proportion data

- proportions are between zero and one.
- Sample sizes on which proportions are calculated matter
- the variance is non-constant
- the observations are integers



General approach

1. Choose

linear predictor $\eta_i = \sum_{j=1}^p x_{ij} \beta_j$

family of probability distributions: **Binomial**

canonical link function g : **logit** function

The predicted mean of the Binomial response variable for observation i is μ_i

$$\eta_i = h(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) \Leftrightarrow \mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

$\rightarrow \mu_i$ is always between zero and one.

When fitting, mind that numbers of positive and negative responses need to be given as responses.

```
totals<-rpois(10,4)
positives<-rbinom(length(totals),totals,0.5)
ybin<-cbind(positives,totals-positives) #!!!
```

2. **Fit:** fit model to data using Maximum Likelihood / IWLS

```
modelb1<-glm(ybin~c(1:10),quasibinomial)
```

3. Output

```
summary(modelb1)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0807368	0.2888702	0.279	0.7893
density	0.0035101	0.0009532	3.683	0.0103 *

(Dispersion parameter for quasibinomial family taken to be 3.471613)

Null deviance: 71.159 on 7 degrees of freedom

Residual deviance: 22.091 on 6 degrees of freedom

Generalized Linear Modelling GLM - short summary

- model response variables in **exponential dispersion family**
- a single response variable
- investigate importance/relevance of explanatory variables
[These can be **continuous** or **discrete** - we can include **interactions** between them] ♣
- Checking assumptions: Investigate model fit graphically (or with a test)
- [start from a maximal model - model selection - hypothesis testing likelihood & frequentist approaches] ♣

other family/link possibilities with glm

binomial(link = "logit" or "probit", 'cauchit', "log" and "cloglog")

gaussian(link = "identity" or "log" or "inverse")

Gamma(link = "inverse", or "identity" or "log")

inverse.gaussian(link = "1/mu^2", "identity" or "log" or "inverse")

poisson(link = "log", or "identity" or "sqrt")

quasi(link = "identity", variance = "constant")

quasibinomial(link = "logit" or nearly any of the others)

quasipoisson(link = "log" or nearly any of the others)

methyKit – differential methylation in genomic data

diffMeth.R

```
diffMeth.R diffMethbetatestmeanwithdifferentva...

Treat <- c( rep(1,length(numC1.ind)),rep(0,length(numC2.ind)) ) # get the treatment vector

Cs=as.matrix(set[,c(numC1.ind,numC2.ind)])
Ts=as.matrix(set[,c(numT1.ind,numT2.ind)])

sums=Cs+Ts # get number of Cs and Ts
Ps =Cs/sums # get probability of Cs

Tmod=model.matrix(~Treat)
glm.bare<-function(X,Y) # X probs > Ps, Y total number > sums
{
  #cat(Tmod)
  obj=glm.fit(Tmod,X,weights=Y,family=binomial(link=logit))
  deviance <- obj$null.deviance - obj$deviance
  dispersion=1 #(if binomial or poisson)
  aliased <- is.na(coef(obj))
  p <- obj$rank
  if (p > 0) { # if clause and the rest to get the t-value or wald statistic
    p1 <- 1L:p
    Qr <- obj$qr
    coef.p <- obj$coefficients[Qr$pivot[p1]]
    covmat.unscaled <- chol2inv(Qr$qr[p1, p1, drop = FALSE])
    dimnames(covmat.unscaled) <- list(names(coef.p), names(coef.p))
    covmat <- dispersion * covmat.unscaled
    var.cf <- diag(covmat)
    s.err <- sqrt(var.cf)
    tvalue <- (coef.p/s.err)[2]
  }else if (obj$df.residual<=0)
  { tvalue=NaN }

  wald <- tvalue
  betal <- obj$coefficients[2]
  p.value <- 1-pchisq(deviance,df=1)

  return( c(wald,betal,p.value) )
}
```


General Many other models for counts, proportions, quantitative measures and durations can be fitted without using IWLS (the distributions are not exponential dispersion family)
They are available in packages "gnlm" (Jim Lindsey) or "bbmle" (Ben Bolker)

Extended glm techniques check these:

`bayesglm()` # bayesian approach

`library(brglm)` # binomial bias-reduced

`library(aod)` # overdispersed data

`vglm()` # zero-truncated distributions

`library(pscl)` # zero-inflated models (mixtures)

References

<http://www.stat.wisc.edu/courses/st849-bates/lectures/GLMH.pdf>

<http://www.stat.wisc.edu/courses/st849-bates/lectures/GLMDeviance.pdf>

<https://www.statistics.ma.tum.de/fileadmin/w00bdb/www/czado/lec2.pdf>

Crawley M (2005) Statistics: An introduction using R

Faraway J (2006) Extending the Linear Model with R. CRC Press

Pawitan, Y (2001) In all Likelihood. Oxford University Press