Introduction to Statistical Learning

Notes by Tim Brown and Owen Jones

Module 8: Exponential families and GLM's

Contents

1 Exponential families **2**
1.1 Big Picture- F6.1 . . . 2
1.2 Definition - F6.1 . . . 2
1.3 Examples - F6.1 . . . 2
1.4 Exponential Families: mean and variance - F6.1 . . . 4

2 GLMs **4**
2.1 Definition - F6.1 . . . 4
2.2 Canonical link - F6.1 . . . 5
2.3 GLM estimation - F6.1 . . . 5
2.4 Newton Raphson Method - not in F . . . 6
2.5 Newton's method for finding a maximum - not in F . . . 6
2.6 Newton's method with Fisher scoring- F6.1 . . . 7

3 Weighted LS **8**
3.1 Newton's method with Fisher scoring applied to GLM- F6.2 . . . 8
3.2 Example: Inseticide efficacy- F6.2 . . . 9
3.3 Example: Variance of MLE - F6.2 . . . 10

4 Inference **11**
4.1 Example: Deviance - F6.3 . . . 11
4.2 Examples- F6.3 . . . 11
4.3 Hypothesis Testing in Nested Models- F6.3 . . . 12
4.4 Example Insects- F6.3 . . . 12
4.5 AIC for GLM's- not in F Ch 6 . . . 14
4.6 Example Galapogos islands- not in F . . . 14

5 Diagnostics **17**
5.1 Residuals- F6.4 . . . 17
5.2 Example Galapogos islands- F6.4 . . . 18
5.3 Leverage- F6.4 . . . 18
5.4 Studentized and Jack-knifed Residuals- F6.4 . . . 22
5.5 Cook's distance- F6.4 . . . 22
5.6 Residual Plots- F6.4 . . . 23
5.7 Example: Galapagos - F6.4 . . . 23

1 Exponential families

1.1 Big Picture- F6.1

Where we have been.

In modules 3 to 6, the theory and practice of linear models were developed. This enabled estimation, confidence intervals and hypothesis testing in a wide variety of situations. Big data analysis, particularly machine learning, uses these tools routinely with a heavy emphasis on estimation. Questions of model selection, reliability, repeatability and robustness, however, require the distribution theory for confidence intervals and hypothesis tests. All that linear algebra and theory of multivariate normal, χ^2, independence and $F-$ distributions constitute the building blocks that enable the data scientist to respond in new, as well as standard, applications.

Where are we going?

In module 7, linear models were extended to generalised linear models involving regression with Binomial rather than normal distributions. A *link function* formed the bond between the linear function of explanatory variables - the *linear predictor* - and the mean of the Binomial distribution. This allows extension of the techniques of linear models to Binomial regression which can be used for large data sets, provided there is some underlying binary response variable. Questions of model selection, reliability, repeatability and robustness still require the distribution theory for confidence intervals and hypothesis tests. In this chapter, the extension to generalised linear models will cover a wide variety of distributions called *exponential families*.

1.2 Definition - F6.1

Exponential families

Y comes from an exponential family if it has a probability density or probability mass function (pdf or pmf) of the form

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

where

θ is the *canonical parameter* (captures location)

ϕ is the *dispersion parameter* (captures scale).

Observe that it is the log of the density function that must have a specific form involving θ and ϕ.

1.3 Examples - F6.1

Example: normal

$Y \sim N(\mu, \sigma^2)$

$$\begin{aligned} f(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}} \\ &= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \tfrac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right] \\ &= \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \end{aligned}$$

where $\theta = \mu$, $\phi = \sigma^2$, and

$$\begin{aligned} b(\theta) &= \theta^2/2 \\ a(\phi) &= \phi \\ c(y,\phi) &= -\tfrac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right) \end{aligned}$$

Example: Poisson

$Y \sim \text{Poisson}(\lambda)$

$$\begin{aligned} f(y) &= e^{-\lambda}\lambda^y/y! \text{ for } y = 0, 1, 2, \ldots \\ &= \exp\left[y\log\lambda - \lambda - \log y!\right] \\ &= \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \end{aligned}$$

where $\theta = \log\lambda$, $\phi = 1$, and

$$\begin{aligned} b(\theta) &= e^\theta \\ a(\phi) &= \phi \\ c(y,\phi) &= -\log y! \end{aligned}$$

Example: Binomial

$Y \sim \text{Binomial}(m, p)$ for known m (not a parameter)

$$\begin{aligned} f(y) &= \binom{m}{y} p^y (1-p)^{m-y} \text{ for } y = 0, 1, \ldots, m \\ &= \exp\left[y\log\frac{p}{1-p} + m\log(1-p) + \log\binom{m}{y}\right] \\ &= \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \end{aligned}$$

where $\theta = \log\frac{p}{1-p}$, $\phi = 1$, and

$$\begin{aligned} b(\theta) &= m\log(1+e^\theta) \\ a(\phi) &= \phi \\ c(y,\phi) &= \log\binom{m}{y} \end{aligned}$$

Example: Scaled Binomial

$X \sim \text{Binomial}(m, p)$ and $Y = X/m$

$$\begin{aligned} f(y) &= \binom{m}{my} p^{my}(1-p)^{m(1-y)} \text{ for } y = 0, 1/m, \ldots, 1 \\ &= \exp\left[\frac{y \log \frac{p}{1-p} + \log(1-p)}{1/m} + \log\binom{m}{my}\right] \\ &= \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \end{aligned}$$

where $\theta = \log \frac{p}{1-p}$, $\phi = 1/m$, and

$$\begin{aligned} b(\theta) &= \log(1 + e^{\theta}) \\ a(\phi) &= \phi \\ c(y, \phi) &= \log\binom{m}{my} \end{aligned}$$

Other examples of exponential families are the gamma and the inverse Gaussian - see Lab and Workshop 9.

1.4 Exponential Families: mean and variance - F6.1

Exponential family: mean and variance

Lemma If Y is from an exponential family then

$$\begin{aligned} \mathbb{E}Y &= b'(\theta) \\ \operatorname{Var} Y &= b''(\theta)a(\phi) \end{aligned}$$

The demonstration is part of Assignment 3.

Exponential family: variance function

Let $\mu = \mathbb{E}Y$ and write

$$\operatorname{Var} Y = v(\mu)a(\phi)$$

so $v = b'' \circ (b')^{-1}$. v is called the *variance function*

Examples:

normal $v(\mu) = 1$

Poisson $v(\mu) = \mu$

binomial $v(\mu) = \mu(1 - \mu/m)$

scaled binomial $v(p) = p(1-p)$

2 GLMs

2.1 Definition - F6.1

Generalised Linear Model

Definition: Y is said to come from a *Generalised Linear Model (GLM)* if the pdf/pmf is from an exponential family, and

$$\mu := \mathbb{E}Y = g^{-1}(\mathbf{x}^T\boldsymbol{\beta})$$

where

g is a monotonic differentiable function called the *link function*.

$\mathbf{x}$ is a vector of independent (explanatory or predictor) variables, and

$\boldsymbol{\beta}$ is a vector of parameters

Remark: We model *location* using $\eta = \mathbf{x}^T\boldsymbol{\beta}$, and let the *scale* be determined by the family of distributions. That is, we do not model the scale explicitly in terms of explanatory variables.

2.2 Canonical link - F6.1

Canonical link

Recall Y is from an exponential family if

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

If $g(\mu) = g(\mathbb{E}Y) = \theta$ then g is called the *cannonical* link. Since $\mu = b'(\theta)$, it follows that the canonical link must be $(b')^{-1}$.

Examples: canonical links

normal $\theta = \mu$, $g(\mu) = \mu$

Poisson $\theta = \log\lambda = \log\mu$, $g(\mu) = \log\mu$

Binomial $\theta = \log\frac{p}{1-p} = \log\frac{\mu}{m-\mu}$, $g(\mu) = \log\frac{\mu}{m-\mu}$

scaled Binomial $\theta = \log\frac{p}{1-p} = \log\frac{\mu}{1-\mu}$, $g(\mu) = \log\frac{\mu}{1-\mu}$

2.3 GLM estimation - F6.1

GLM estimation

We fit GLMs using maximum likelihood.

Suppose we have independent observations y_i from an exponential family, with canonical parameter θ_i and dispersion parameter ϕ, for $i = 1,\ldots,n$.

Furthermore suppose that y_i has mean

$$\mu_i = b'(\theta_i) = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$$

If g is the canonical link then $\theta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, so the multiplier of y is the log pdf/pmf is the linear predictor $\mathbf{x}_i^T\boldsymbol{\beta}$.

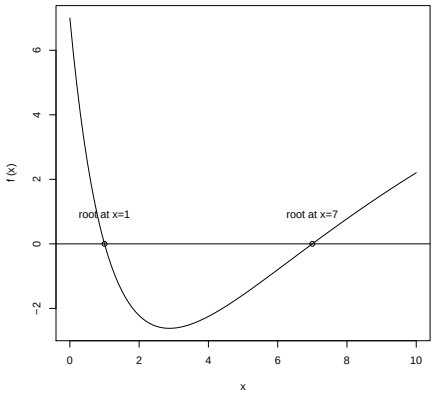
The log-likelihood is then

$$l(\boldsymbol{\beta},\phi;\mathbf{y}) = \sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right)$$

2.4 Newton Raphson Method - not in F

To find the maximum likelihood estimate, the log-likelihood needs to be differentiated and the result equated to zero. This requires a procedure which works well for our generalised linear models. Fortunately one is available using Newton's method. Newton's method is discussed in the next slides.

Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a continuous function. A *root* of f is a solution to the equation $f(x) = 0$



The Newton-Raphson method

Suppose the function f is differentiable with continuous derivative f' and a root a. Let $x_0 \in \mathbb{R}$ and think of x_0 as our current 'guess' at a. The straight line through the point $(x_0, f(x_0))$ with slope $f'(x_0)$ is the local straight line approximation to the function $f(x)$. The equation of this straight line is given by

$$y = f(x_0) + f'(x_0)(x - x_0).$$

This straight line crosses the x-axis at a point x_1, which should be a better approximation than x_0 to a. To find x_1 transform the previous equation:

$$f'(x_0) = \frac{f(x_0) - 0}{x_0 - x_1} \quad \text{and so} \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

2.5 Newton's method for finding a maximum - not in F

Newton's method

This is the Newton-Raphson method applied to finding a maximum or minimum. So the root of the function is replaced by finding the root of the derivative. The current approximation to the root of the derivative, x_n, generates the next approximation via:

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}.$$

Often, in implementation, the process finishes if successive approximations are sufficiently close. Convergence is not guaranteed and the second derivative needs to be non-zero.

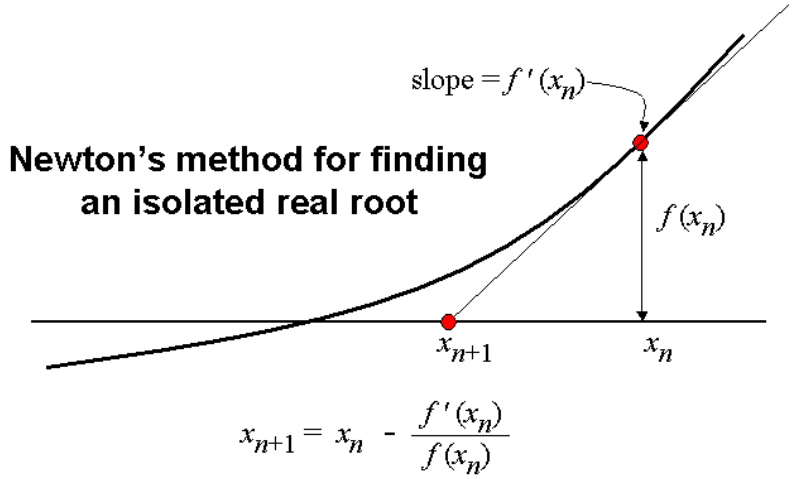


Figure 1: Illustration of one step in Newton's method

Newton's method - higher dimensions

The one dimensional Newton's method extends to functions which have more than one variable - like our log-likelihood. Suppose we have a vector of variables $\mathbf{x} = (x_1, x_2, \ldots, x_k)^T$, and some scalar function of them:

$$z = f(\mathbf{x}).$$

Recall, from Module 2 Section 8, that the derivative of z with respect to $\mathbf{y}$ is defined as:

$$\frac{\partial z}{\partial \mathbf{x}} = \begin{bmatrix} \partial z/\partial x_1 \\ \partial z/\partial x_2 \\ \vdots \\ \partial z/\partial x_k \end{bmatrix}.$$

Newton's method - higher dimensions

Further, let H be the $k \times k$ *Hessian* matrix defined by:

$$H_{ij}(\mathbf{x}) = \frac{\partial^2 z}{\partial x_i \partial x_j}$$

In the multivariate case, the current approximation to the root of the derivative, x_n, generates the next approximation via:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - H^{-1}(\mathbf{x}_n)\frac{\partial z}{\partial \mathbf{x}}(\mathbf{x}_n).$$

2.6 Newton's method with Fisher scoring- F6.1

Fisher scoring

Suppose we wish to maximise a log likelihood $l(\boldsymbol{\theta})$ using Newton's method. Our update step is

$$\boldsymbol{\theta}(n+1) = \boldsymbol{\theta}(n) - H(\boldsymbol{\theta}(n))^{-1}\frac{\partial l}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}(n))$$

where $H(\boldsymbol{\theta}) = (H_{ij}(\boldsymbol{\theta}))$ for

$$H_{ij}(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}$$

That is, $-H(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})$, the observed information.

If we replace $\mathcal{J}$ by $\mathcal{I}$, the Fisher information, then the algorithm is called *Fisher scoring*.

The Fisher information is often easier/quicker to calculate, and is guaranteed to be positive definite (unlike the observed information).

3 Weighted LS

3.1 Newton's method with Fisher scoring applied to GLM-F6.2

Weighted Least Squares

For a GLM, a remarkable result is that the Fisher scoring version of Newton's method produces an iteration step which is the solution of a weighted least squares problem.

This problem involves the data random variables

$$\begin{aligned} Z_i &:= g(\mu_i) + (Y_i - \mu_i) g'(\mu_i) \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \text{ say} \end{aligned}$$

where

$$\operatorname{Var} \epsilon_i = (g'(\mu_i))^2 \operatorname{Var} Y_i.$$

The random variable Z_i is the straight line approximation to $g(Y_i)$ starting at $(\mu_i, g(\mu_i))$.

Weighted LS for GLM

From Module 4, Section 10.2, that if $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\operatorname{Var}(\mathbf{Y}) = \Sigma$ with X and Σ full rank, then the BLUE estimator of $\boldsymbol{\beta}$ is

$$(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \mathbf{Y}.$$

This suggests that $\boldsymbol{\beta}$ is estimated using

$$\hat{\boldsymbol{\beta}} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \mathbf{Z}.$$

Problem: Clearly z_i depends on $\boldsymbol{\beta}$, but $\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$, so Σ_{ii} also depends on $\boldsymbol{\beta}$.

Solution: Iterate! This produces the Fisher scoring version of Newton's algorthm for finding $\boldsymbol{\beta}$.

Note that the $a(\phi)$ factor in the expression for $\hat{\boldsymbol{\beta}}$ cancels out and that the covariance matrix Σ is diagonal.

Newton's method with Fisher scoring or Iterated Weighted Least Squares (IWLS)

1. Start with $\hat{\mu}(0) = \mathbf{y}$.
2. Given $\hat{\mu}(n)$ calculate
   $z_i(n) = g(\hat{\mu}_i(n)) + (y_i - \hat{\mu}_i(n))g'(\hat{\mu}_i(n))$ and
   $W_{ii}(n) = 1/[g'(\hat{\mu}_i(n))^2 v(\hat{\mu}_i(n))]$, for each i.
3. Put $\hat{\boldsymbol{\beta}}(n+1) = (X^T W(n) X)^{-1} X^T W(n)\mathbf{z}(n)$ and
   $\hat{\mu}_i(n+1) = g^{-1}(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}(n+1))$ for each i.
4. If $\hat{\boldsymbol{\beta}}(n+1)$ is sufficiently close to $\hat{\boldsymbol{\beta}}(n)$ then stop, otherwise return to (2).

Mostly, for GLM's the algorithm converges and if it does then the resultant estimator is the maximum likelihood estimator.

3.2 Example: Inseticide efficacy- F6.2

Example: insecticide efficacy

An experiment measuring death rates for insects, with 30 insects at each of five treatment levels.

```
library(faraway)
data(bliss)
bliss
```

```
##   dead alive conc
## 1    2    28    0
## 2    8    22    1
## 3   15    15    2
## 4   23     7    3
## 5   27     3    4
```

We model this with a binomial regression model, and fit using IWLS...

Example: start

```
# IWLS
y <- bliss$dead
m <- bliss$dead + bliss$alive

mu <- y
eta <- logit(mu/m)
z <- eta + (y - mu)*m/mu/(m - mu)
w <- mu*(m - mu)/m
lmod <- lm(z ~ conc, weights=w, bliss)
coef(lmod)
```

```
## (Intercept)        conc
##   -2.302462    1.153587
```

Example: iteration

```
for (i in 1:5) {
eta <- lmod$fit
mu <- m*ilogit(eta)
z <- eta + (y - mu)*m/mu/(m - mu)
w <- mu*(m - mu)/m
lmod <- lm(z ~ conc, weights=w, bliss)
cat(i, coef(lmod), "\n")}

## 1 -2.323672 1.161847
## 2 -2.32379 1.161895
## 3 -2.32379 1.161895
## 4 -2.32379 1.161895
## 5 -2.32379 1.161895
```

3.3 Example: Variance of MLE - F6.2

Variance of $\hat{\boldsymbol{\beta}}$

Suppose that the IWLS algorithm converges to the estimate $\hat{\boldsymbol{\beta}}$, then

$$\hat{\boldsymbol{\beta}} = (X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}\mathbf{z}$$

where for each $i = 1, \cdots, n$

$$\begin{aligned}\hat{\mu}_i &= g^{-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})\\ z_i &= \mathbf{x}_i^T\hat{\boldsymbol{\beta}} + (y_i - \hat{\mu}_i)g'(\hat{\mu}_i)\\ \hat{\Sigma}_{ii} &= (g'(\hat{\mu}_i))^2 v(\hat{\mu}_i)a(\phi)\end{aligned}$$

Since $\operatorname{Var}\mathbf{z} = \hat{\Sigma}$,

$$\begin{aligned}\operatorname{Var}\hat{\boldsymbol{\beta}} &= [(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}]\hat{\Sigma}[(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}]^T\\ &= (X^T\hat{\Sigma}^{-1}X)^{-1}\end{aligned}$$

Note that the $a(\phi)$ term in $\hat{\Sigma}$ does not cancel here, as it did in the IWLS algorithm, so we need to estimate it.

Now $(Y_i - \mu_i)/\sqrt{v(\mu_i)}$ has mean 0 and variance $a(\phi)$, so it should come as no surprise that

$$X^2 := \sum_i \frac{(Y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} \approx a(\phi)\chi^2_{n-p}$$

where p is the number of parameters used to estimate $\boldsymbol{\mu}$.

Thus $X^2/(n-p)$ will be an estimator for $a(\phi)$.

X^2 is called Pearson's χ^2 statistic, and it can be shown that $X^2/(n-p)$ is a consistent estimator for $a(\phi)$.

4 Inference

4.1 Example: Deviance - F6.3

Deviance

From here on, by reinterpreting the parameter if necessary, assume $a(\phi) = \phi$.

Definition: the *scaled deviance*, $\frac{D^A}{\phi}$, for model A is

$$\frac{D^A}{\phi} = -2 \log \frac{\mathcal{L}(\hat{\boldsymbol{\beta}}^A)}{\mathcal{L}(\text{full})}$$

where

- $\hat{\boldsymbol{\beta}}^A$ is the MLE of $\boldsymbol{\beta}^A$, the true parameter value for model A, and
- $\mathcal{L}(\text{full})$ is the maximum likelihood for the "full" model with one parameter for each observation.

The *deviance* is just D^A.

4.2 Examples- F6.3

Example: Normal

The full normal model has MLE, y_i, in estimating one mean, μ_i, for each of the n observations.

The deviance can be written as $D = \sum_i d_i$ where

$$d_i = (y_i - \hat{\mu}_i)^2$$

where $\hat{\mu}_i$ is the fitted mean using the model, so that D is the residual sum of squares for the fitted model.

Example: Poisson

The full Poisson model has MLE y_i for $\mu_i = \lambda_i$.

The deviance (and the scaled deviance since $\phi = 1$) can be written as $D = \sum_i d_i$ where

$$d_i = -2 \left(y_i \log \frac{\hat{\mu}_i}{y_i} - (\hat{\mu}_i - y_i) \right)$$

where $\hat{\mu}_i$ is the fitted mean using the model.

Example: Binomial

The full Binomial model has MLE y_i for $\mu_i = m_i p_i$.

Equivalently y_i/m_i is the full model MLE of p_i.

The deviance (and the scaled deviance since $\phi = 1$) can be written as $D = \sum_i d_i$ where

$$d_i = -2 \left(y_i \log \frac{\hat{\mu}_i}{y_i} + (m_i - y_i) \log \frac{m_i - \hat{\mu}_i}{m_i - y_i} \right)$$

where $\hat{\mu}_i$ is the fitted mean using the model.

4.3 Hypothesis Testing in Nested Models- F6.3

If the model is adequate then the scaled deviance will often (but not always) be $\approx \chi^2_{n-p}$, where the full model has n parameters (equal to the number of observations) and the fitted model has p parameters.

For nested models, if the smaller model is correct then the difference between two scaled deviances is the log likelihood ratio and will be $\approx \chi^2_s$ for large n, where s is the difference in the number of parameters.

The scaled deviance can be used to test model adequacy, but the χ^2approximation for the full model scaled deviance is not as reliable, for large n, as the approximation for the deviation difference.

If the difference between two scaled deviances is large enough, then the null hypothesis of the adequacy of the smaller model can be rejected (this is called a *log likelihood ratio test*).

For the Binomial and Poisson models $\phi = 1$ and the scaled deviance is just the deviance.

For these models the scaled deviance will be approximately χ^2 when the individual responses are somewhat normal. As a rule of thumb we need the Poisson mean or the Binomial mean of both failures and successes to be at least 5.

For the Normal, Gamma or Inverse Gaussian models, $X^2/(n-p)$, or the residual deviance divided by the degrees of freedom, estimates the scale parameter ϕ.

So, for models with ϕ a parameter, the residual deviance can't be used to test model adequacy.

For a linear model A nested within linear model B with , under the null hypothesis that model A is correct ,

$$\frac{(D^A - D^B)/s}{X^2/(n-p)} \sim F_{s,n-p}$$

where there are n observations, A has $p-s$ parameters and B has p parameters.

In R X^2 (Pearson's chi-squared) is calculated using the fitted model A.

For other GLM's this distributional result only holds approximately, but it can still be used for comparing models. In particular it can be used to compare Gamma models (alternatively the AIC can be used).

4.4 Example Insects- F6.3

Blisss(1935) dose-response

Recall that 5 sets of 30 insects were exposed to 5 different concentrations of insecticide and after exposure the number alive was recorded.

```
str(bliss)
```

```
## 'data.frame': 5 obs. of  3 variables:
##  $ dead : num  2 8 15 23 27
##  $ alive: num  28 22 15 7 3
##  $ conc : int  0 1 2 3 4
```

GLM command

The GLM command can fit the logistic model as follows:

```
modl <- glm(cbind(dead,alive) ~ conc,family=binomial, data=bliss)
summary(modl)
```

```
##
## Call:
## glm(formula = cbind(dead, alive) ~ conc, family = binomial, data = bliss)
##
## Deviance Residuals:
##       1        2        3        4        5
## -0.4510   0.3597   0.0000   0.0643  -0.2045
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3238     0.4179  -5.561 2.69e-08 ***
## conc          1.1619     0.1814   6.405 1.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 64.76327  on 4  degrees of freedom
## Residual deviance:  0.37875  on 3  degrees of freedom
## AIC: 20.854
##
## Number of Fisher Scoring iterations: 4
```

Model adequacy

The residual deviance p-value can be found from:

```
pchisq(0.37585,df=3,lower=FALSE)
```

```
## [1] 0.9451849
```

There is a very large chance that a chi-square rv with 3 df would be more than the residual deviance so there is no evidence of model inadequacy.

Adequacy of null model

The null deviance is the deviance if only one probability is fitted for all 5 concentrations.

The p-value for the null model is:

```
pchisq(64.76327,df=4,lower=FALSE)
```

```
## [1] 2.886305e-13
```

There is a tiny chance that a chi-square rv with 4 df would be more than the observed null deviance so the model with one probability for all concentrations is not adequate.

Significance of concentration

The null deviance is the deviance if only one probability is fitted for all 5 concentrations.

The difference between this and the residual deviance is the contribution of the insecticide concentration in explaining the response.

This difference is $64.76327 - 0.37585 = 64.38742$ and this has 1 degree of freedom with p-value

```
pchisq(64.38742,df=1,lower=FALSE)

## [1] 1.022087e-15
```

There is a tiny chance that a chi-square rv with 1 df would be more than the observed difference so concentration is significant in the presence of an intercept term.

Use of GLM anova

An additional model could be fitted with a quadratic term as follows. This is done only for illustrative purposes since the low residual deviance makes any refinement of the model dubious.

```
mod12 <- glm(cbind(dead, alive) ~ conc+I(conc^2),
family=binomial,bliss)
anova(modl,mod12,test="Chi")

## Analysis of Deviance Table
##
## Model 1: cbind(dead, alive) ~ conc
## Model 2: cbind(dead, alive) ~ conc + I(conc^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         3    0.37875
## 2         2    0.19549  1  0.18325   0.6686
```

The large p-value confirms that the quadratic term is not necessary.

4.5 AIC for GLM's- not in F Ch 6

Akaike Information Criterion (AIC)

For model selection, the AIC often produces better models than comparing (scaled) deviances, and choosing a significance level is not necessary. The AIC is defined as

$$\text{AIC} = 2p - 2\log\mathcal{L}(\hat{\boldsymbol{\beta}}^{)}$$

where p is the number of parameters in the model. The model with smallest AIC is preferred.

If model B has s more parameters than model A (not necessarily nested within B), then

$$\begin{aligned}\text{AIC}^B - \text{AIC}^A &= 2s - 2\log\mathcal{L}(\hat{\boldsymbol{\beta}}^B) + 2\log\mathcal{L}(\hat{\boldsymbol{\beta}}^A) \\ &= 2s + \frac{D^B}{\phi} - \frac{D^A}{\phi}.\end{aligned}$$

Like the log likelihood ratio test, the AIC needs an estimate of ϕ if this not fixed by the model.

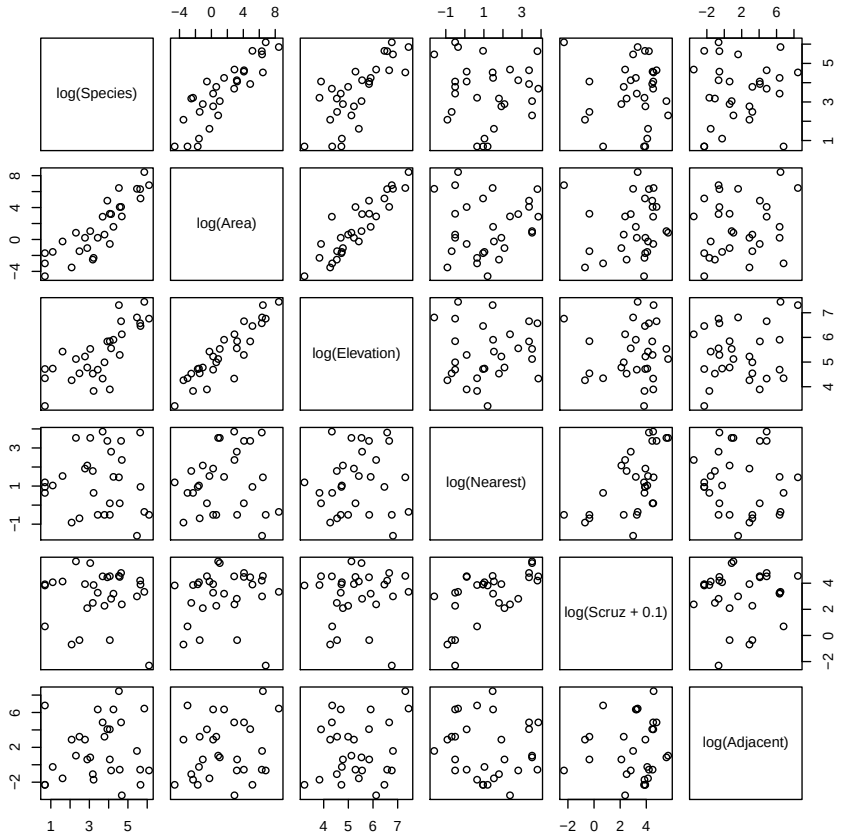
4.6 Example Galapogos islands- not in F

Data description

Data was collected on the number of species on each of 30 islands in the Galapogos Islands, famous for their role in Darwin's theory of evolution.

There were 5 geographic variables that can be used to explain the number of species on each island:

Figure 2: Pairwise plots of logs of Galapogos variables



- Area - the area of the island (km^2)
- Elevation - the highest elevation of the island (m)
- Nearest - the distance from the nearest island (km)
- Scruz - the distance from Santa Cruz island (km)
- Adjacent - the area of the adjacent island (square km)

A lot of variability indicates taking logs might help. Pairwise plots are shown in Figure 2.

Poisson GLM

```
modp <- glm(Species ~ log(Area) + log(Elevation) +
log(Nearest) + log(Scruz+0.1) + log(Adjacent),
family=poisson, gala)
summary(modp)


##
## Call:
## glm(formula = Species ~ log(Area) + log(Elevation) + log(Nearest) +
```

```
##     log(Scruz + 0.1) + log(Adjacent), family = poisson, data = gala)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.4479  -2.6717  -0.4547   2.5613   8.2970
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.287941   0.284661  11.550  < 2e-16 ***
## log(Area)         0.348445   0.018029  19.327  < 2e-16 ***
## log(Elevation)    0.036421   0.056983   0.639  0.52272
## log(Nearest)     -0.040644   0.013781  -2.949  0.00318 **
## log(Scruz + 0.1) -0.030045   0.010492  -2.864  0.00419 **
## log(Adjacent)    -0.089014   0.006948 -12.812  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3510.73  on 29  degrees of freedom
## Residual deviance:  359.12  on 24  degrees of freedom
## AIC: 531.96
##
## Number of Fisher Scoring iterations: 5
```

Stepwise using AIC

The command to see what is the best model starting with this full model is:

```
step(modp, scope = ~ .)
```

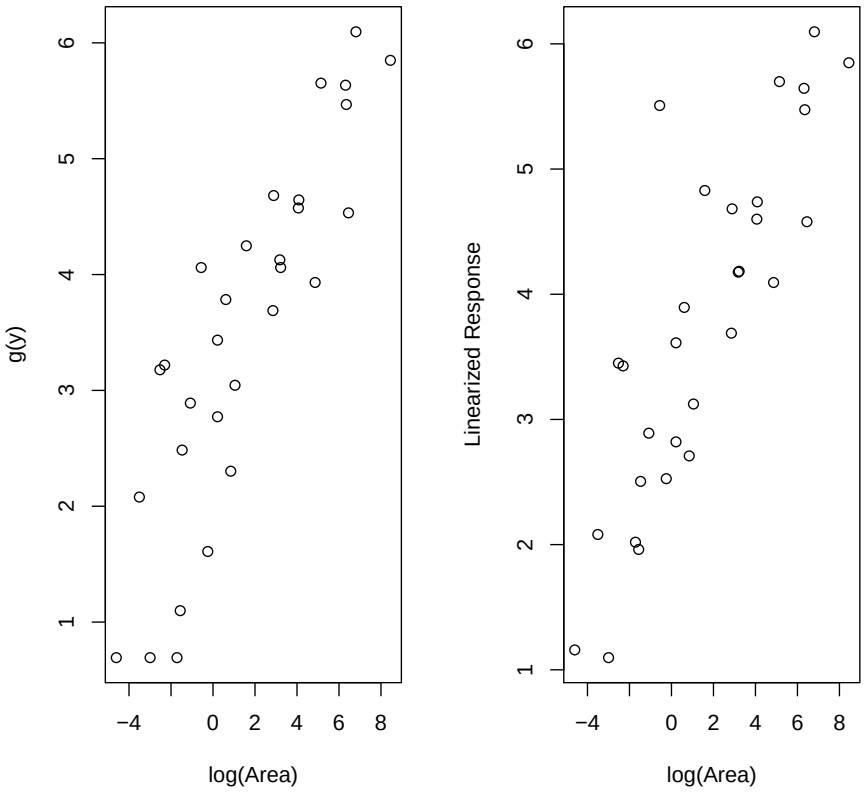
```
## Start:  AIC=531.96
## Species ~ log(Area) + log(Elevation) + log(Nearest) + log(Scruz +
##     0.1) + log(Adjacent)
##
##                   Df Deviance    AIC
## - log(Elevation)    1   359.54 530.37
## <none>                  359.12 531.96
## - log(Scruz + 0.1)  1   367.27 538.10
## - log(Nearest)      1   367.79 538.62
## - log(Adjacent)     1   525.13 695.96
## - log(Area)         1   714.98 885.81
##
## Step:  AIC=530.37
## Species ~ log(Area) + log(Nearest) + log(Scruz + 0.1) + log(Adjacent)
##
##                   Df Deviance    AIC
## <none>                 359.5  530.4
## + log(Elevation)    1    359.1  532.0
## - log(Scruz + 0.1)  1    367.7  536.6
## - log(Nearest)      1    368.5  537.3
## - log(Adjacent)     1    528.6  697.4
## - log(Area)         1   3266.1 3434.9
##
## Call:  glm(formula = Species ~ log(Area) + log(Nearest) + log(Scruz +
##     0.1) + log(Adjacent), family = poisson, data = gala)
##
## Coefficients:
##      (Intercept)         log(Area)      log(Nearest)  log(Scruz + 0.1)
##          3.46648           0.35871          -0.04112          -0.03010
##    log(Adjacent)
##         -0.08822
##
## Degrees of Freedom: 29 Total (i.e. Null);  25 Residual
## Null Deviance:     3511
## Residual Deviance: 359.5  AIC: 530.4
```

Plots comparing linearized response Z with log Species

Figure 3 shows the model final linearized response $\mathbf{Z} = (\texttt{Species} - \mu)/\mu$ as well as log `Species` plotted against log `Area`

```
par(mfrow=c(1,2))
# g(y) vs log(Area)
plot(log(Species) ~ log(Area), gala, ylab="g(y)")
# linearised response vs log(Area)
mu <- predict(modp, type="response")
```

Figure 3: Plots comparing linearized response **z** with log Species




```
z <- predict(modp) + (gala$Species - mu)/mu
plot(z ~ log(Area), gala, ylab="Linearized Response")
```

5 Diagnostics

5.1 Residuals- F6.4

Diagnostics: residuals

Response residuals: $y_i - \hat{\mu}_i$

Unless the variance function $v(\mu)$ is constant, as in the Gaussian case, plots with the response residuals are not very useful in assessing whether the model is reasonable.

Pearson residuals:

$$r_P(i) = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}$$

Pearson residuals are (approximately) homoskedastic, and $\sum_i r_P(i)^2 = X^2$.

Deviance residuals:

$$r_D(i) = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

where the deviance is $D = \sum_i d_i = \sum_i r_D(i)^2$.

For GLMs deviance residuals are often the most useful, noting that the need for additional parameters in a model is assessed through reduction in deviance.

As for linear models, patterns in the residuals indicate structure in the data that has not been captured by the model.

We can plot the residuals against predictor variables, the responses, or the fitted means. Often a plot against the linear predictors, $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, works well.

With count data residual plots exhibit banding due to the discrete nature of the responses, and this can make it hard to see other patterns.

In this case we can use a smoothed fit of the residuals to help spot trends/patterns.

5.2 Example Galapogos islands- F6.4

Residual plot commands

Three different plots of residuals are in Figures 4, 5 and 6.

```
par(mfrow=c(1,1))
#  Deviance residuals versus fitted values
plot(residuals(modp) ~ predict(modp,type="response"),
xlab=expression(hat(mu)), ylab="Deviance residuals",
main="Cramped in x")
# Deviances residuals versus linear predictor
plot(residuals(modp) ~ predict(modp,type="link"),
xlab=expression(hat(eta)), ylab="Deviance residuals",
main="Easier to look for patterns - none observed")
#  Residuals versus linear predictor
plot(residuals(modp,type="response") ~ predict(modp, type="link"),
xlab=expression(hat(eta)), ylab="Response residuals",
main= "Heteroskedastic")
```

5.3 Leverage- F6.4

Leverage

Just as in linear models, the leverage measures the potential influence of a point on the fitted model. The definition of leverage comes from the theory of linear models, using the hat matrix from the IWLS fitting.

In the IWLS scheme $\Sigma_{ii} = (g'(\mu_i))^2 v(\mu_i)\phi$ (assuming $a(\phi) = \phi$) and

$$\mathbf{Z}' = \Sigma^{-1/2}\mathbf{Z} = \Sigma^{-1/2}X\boldsymbol{\beta} + \boldsymbol{\varepsilon}'$$

where $\text{Var}\,\boldsymbol{\varepsilon}' = I$. The hat matrix for $\hat{\mathbf{Z}}'$ is the matrix H' such that $H'\mathbf{Z}' = \hat{\mathbf{Z}}'$. From the theory of linear models

$$H' = \Sigma^{-1/2}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1/2}$$

Thus $\hat{\mathbf{Z}} = X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{Z}$, and the hat matrix for $\hat{\mathbf{Z}}$ is

$$H = X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}.$$

Figure 4: Deviance residuals versus the fitted values for number of species

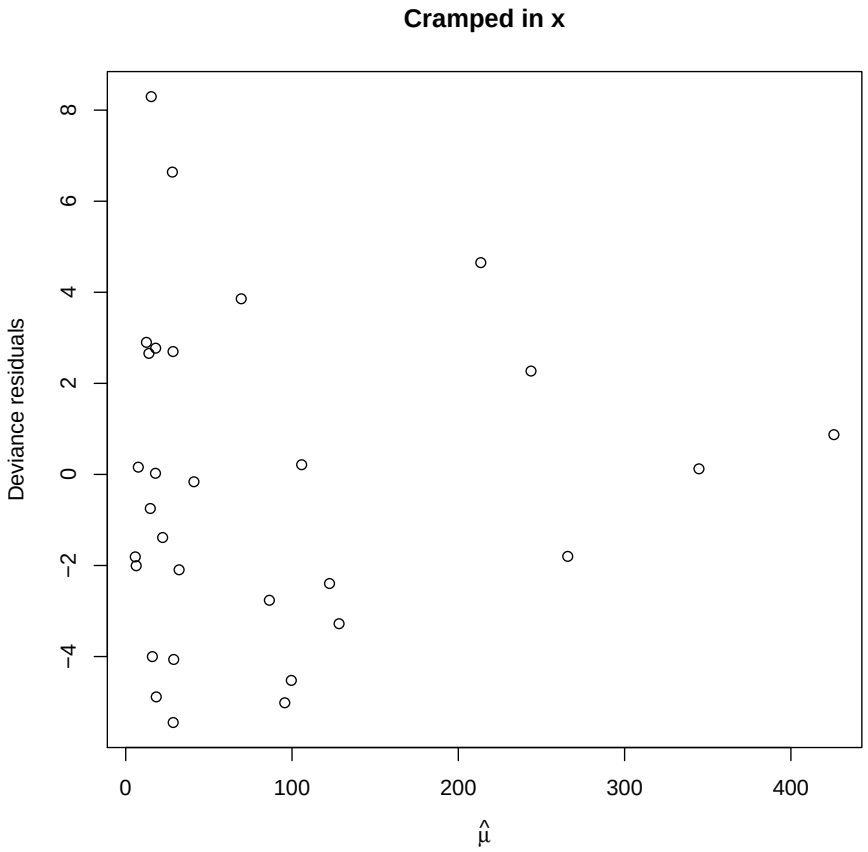


Figure 5: Deviance residuals versus the fitted values on linear predictor scale

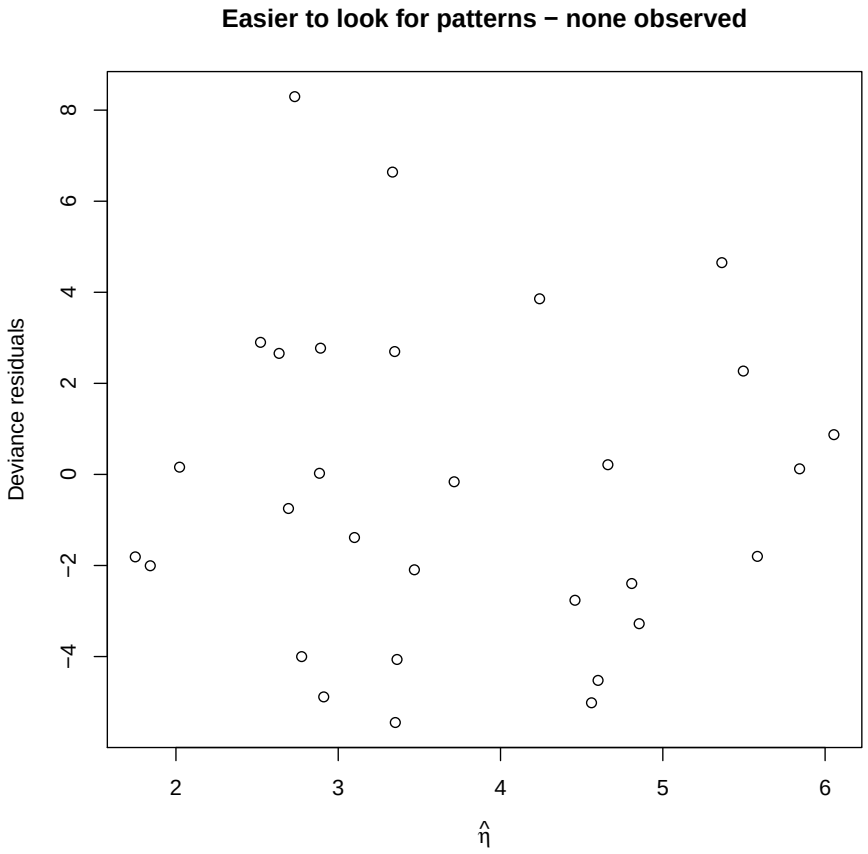
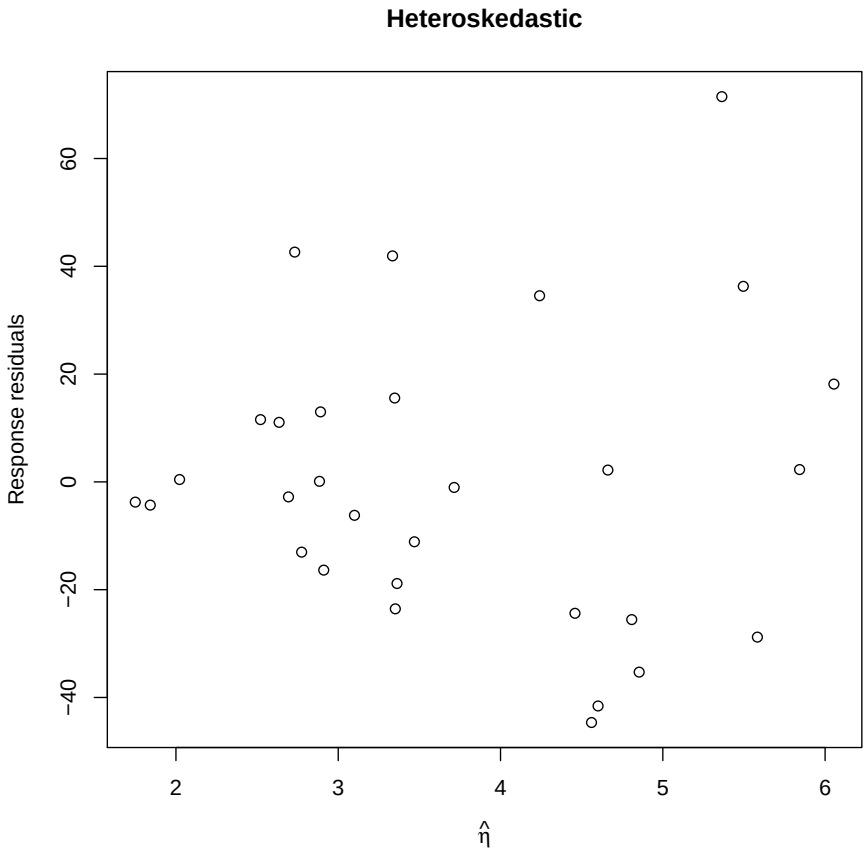


Figure 6: Residuals versus the fitted values on linear predictor scale



The leverage for the i-th observation is H_{ii}, the i-th diagonal element of H.

Note that H does not depend on ϕ (it cancels out).

The variance of $\hat{\mathbf{Z}}$ is

$$\begin{aligned} & H\text{Var}\,\mathbf{Z}H^T \\ = \; & [X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}]\Sigma[\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T] \\ = \; & X(X^T\Sigma^{-1}X)^{-1}X^T \quad = \quad H\Sigma \end{aligned}$$

This leads to $\text{Var}\,(\mathbf{Z}-\hat{\mathbf{Z}}) = (I-H)\Sigma$.

5.4 Studentized and Jack-knifed Residuals- F6.4

Studentised residuals

From the above we have $\text{Var}\,\frac{Z_i-\hat{Z}_i}{\sqrt{(1-H_{ii})g'(\mu_i)^2v(\mu_i)\phi}} = 1$. Also

$$\begin{aligned} Z_i - \hat{Z}_i &= g(\mu_i) + (Y_i-\mu_i)g'(\mu_i) - [g(\mu_i) + (\hat{Y}_i-\mu_i)g'(\mu_i)] \\ &= (Y_i-\hat{Y}_i)g'(\mu_i) \end{aligned}$$

so, noting that $\hat{Y}_i = \hat{\mu}_i$,

$$\text{Var}\,\frac{Y_i-\hat{\mu}_i}{\sqrt{(1-H_{ii})v(\hat{\mu}_i)\hat{\phi}}} \approx 1$$

The LHS is just $r_P(i)/\sqrt{(i-H_{ii})\hat{\phi}} =: r_{SP}(i)$, which we call the i-th **studentised Pearson residual**.

Jack-knife residuals

By analogy we define the i-th **studentised deviance residual** to be

$$r_{SD}(i) = \frac{r_D(i)}{\sqrt{(1-H_{ii})\hat{\phi}}}$$

A large leverage does not necessarily mean a point *has* influenced the fit.

A direct measure of the influence of a point is the **jack-knife residual**, which is the change in $\hat{\mu}_i$ when you remove y_i from the set of observations, then scaled to standardise the variance.

The jack-knife residual can be approximated by

$$\text{sign}(y_i-\hat{\mu}_i)\sqrt{(1-H_{ii})r_{SD}^2(i) + H_{ii}r_{SP}^2(i)}.$$

5.5 Cook's distance- F6.4

Cook's distance

Another measure of the influence of the i-th observation is **Cook's distance**:

$$\frac{(\hat{\boldsymbol{\beta}}^{(i)}-\hat{\boldsymbol{\beta}})^T X^T W X(\hat{\boldsymbol{\beta}}^{(i)}-\hat{\boldsymbol{\beta}})}{p\hat{\phi}}$$

where $\hat{\boldsymbol{\beta}}^{(i)}$ is the estimate of $\boldsymbol{\beta}$ obtained when y_i is omitted. (Note that $W/\hat{\phi} = \hat{\Sigma}^{-1}$)

Recall that for linear models the Cook's distance can be expressed in terms of the leverage and the studentised (Pearson) residual, and is large when both of these are large. Values greater than 1 are usually of interest and/or values substantially larger than the majority.

The jack-knife residual and Cook's distance are both useful for detecting potential outliers.

5.6 Residual Plots- F6.4

Plotting residuals

When looking at residuals it is helpful to consider them ordered by absolute size.

If we plot the ordered absolute values against the percentage points of a half-normal dsitribution then it is easier to see if the largest values are in keeping with the others.

That is, plot the i-th ordered absolute residual against $\Phi^{-1}((n+i)/(2n+1))$, for $i = 1, \ldots, n$. If all is well we expect to see a smooth plot, while a jump or kink in the tail indicates a potential problem.

Note that for a glm the residuals will not in general be normal, so don't expect a straight line.

Checking linearity

A non-linear link g (anything except the identity) makes it harder to check the assumption that $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$.

The easiest thing to do is to plot $g(y_i)$ against $\{x_{ij}\}_{i=1}^n$ for each j and look for linear relationships. A more sophisticated approach is to plot $z_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$ against $\{x_{ij}\}_{i=1}^n$, where $\hat{\mu}_i$ substitutes for μ_i. From the IWLS scheme —provided the predictor variables are linearly independent— these plots should be linear.

If there are non-linearities present then we can consider transforming $\{x_{ij}\}_{i=1}^n$ or adding extra variables.

Transforming the responses y_i is often not a good idea for a glm, as this can break assumptions made about the distribution of Y_i.

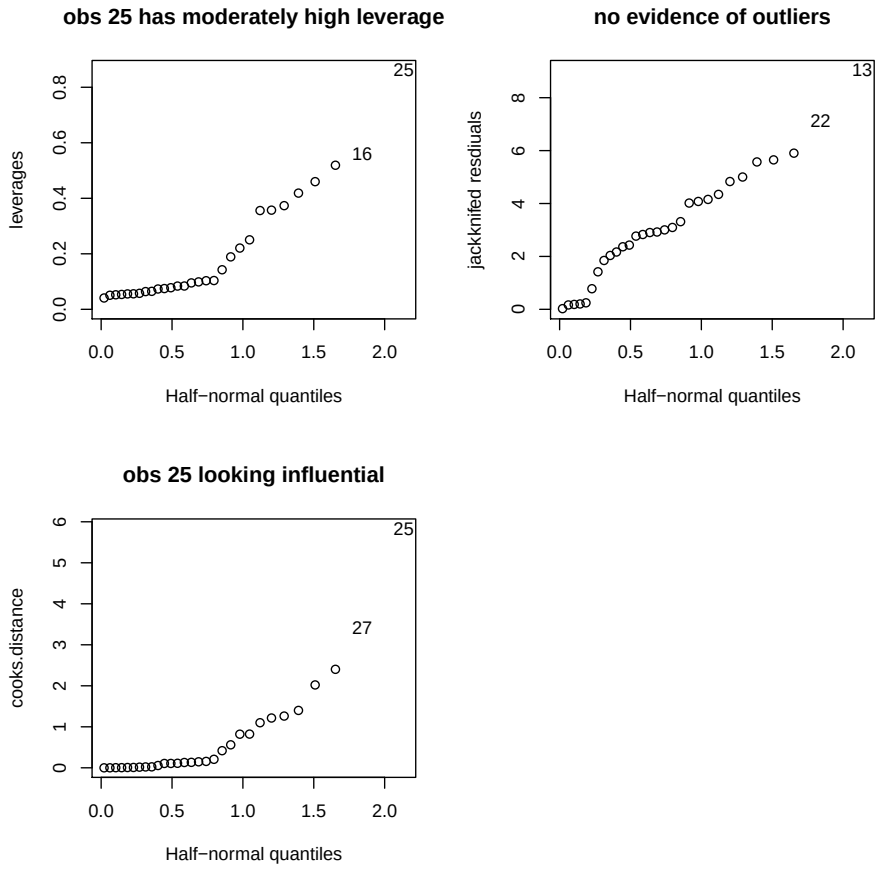
5.7 Example: Galapagos - F6.4

Leverage, Jack-knife residuals, Cook's distance

Plots illustrating these are in Figure 7 .

```
par(mfrow=c(2,2))
# leverages
halfnorm(influence(modp)$hat,
ylab="leverages",
main= "obs 25 has moderately high leverage")
# jackknife residuals
halfnorm(rstudent(modp),
ylab="jackknifed resdiuals",
main= "no evidence of outliers")
```

Figure 7: Leverage, Jack-knife residuals, Cook's distance




```
# Cook's distance - obs 25 looking influential
halfnorm(cooks.distance(modp),
ylab="cooks.distance",
main = "obs 25 looking influential")
```

What to do about 25?

```
# observation 25 has Scruz=0,
# which gives -infty when taking log
# our artificial adjustment by 0.1 could be the problem

# effect of removing obs 25 on model
modp2 <- glm(Species ~ log(Area) + log(Elevation) +
log(Nearest) + log(Scruz+0.1) + log(Adjacent),
family=poisson, gala, subset=-25)
summary(modp2)
step(modp2)
```

```
##
## Call:
## glm(formula = Species ~ log(Area) + log(Elevation) + log(Nearest) +
##     log(Scruz + 0.1) + log(Adjacent), family = poisson, data = gala,
##     subset = -25)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.7237  -2.7539  -0.3181   2.6401   7.9333
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.05070    0.30033  10.158  < 2e-16 ***
## log(Area)         0.33453    0.01883  17.770  < 2e-16 ***
## log(Elevation)    0.05960    0.05743   1.038 0.299325
## log(Nearest)     -0.05255    0.01469  -3.578 0.000347 ***
## log(Scruz + 0.1)  0.01592    0.02218   0.718 0.472998
## log(Adjacent)    -0.08852    0.00696 -12.717  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2707.88  on 28  degrees of freedom
## Residual deviance:  353.42  on 23  degrees of freedom
## AIC: 518.32
##
## Number of Fisher Scoring iterations: 5
## Start:  AIC=518.32
## Species ~ log(Area) + log(Elevation) + log(Nearest) + log(Scruz +
##     0.1) + log(Adjacent)
##
##                    Df Deviance    AIC
## - log(Scruz + 0.1)  1   353.94 516.84
## - log(Elevation)    1   354.51 517.41
## <none>                  353.42 518.32
## - log(Nearest)      1   366.21 529.11
## - log(Adjacent)     1   516.83 679.73
## - log(Area)         1   663.37 826.27
##
## Step:  AIC=516.84
## Species ~ log(Area) + log(Elevation) + log(Nearest) + log(Adjacent)
##
##                  Df Deviance    AIC
## - log(Elevation)  1   354.83 515.72
## <none>                353.94 516.84
## - log(Nearest)    1   368.20 529.09
## - log(Adjacent)   1   519.96 680.86
## - log(Area)       1   679.00 839.90
##
## Step:  AIC=515.72
## Species ~ log(Area) + log(Nearest) + log(Adjacent)
##
##                Df Deviance     AIC
## <none>              354.83  515.72
## - log(Nearest)  1   369.86  528.76
## - log(Adjacent) 1   521.71  680.60
## - log(Area)     1  2679.93 2838.82
##
## Call:  glm(formula = Species ~ log(Area) + log(Nearest) + log(Adjacent),
##     family = poisson, data = gala, subset = -25)
##
## Coefficients:
##   (Intercept)      log(Area)   log(Nearest)  log(Adjacent)
##       3.38465        0.35292       -0.04788       -0.08662
##
## Degrees of Freedom: 28 Total (i.e. Null);  25 Residual
## Null Deviance:     2708
## Residual Deviance: 354.8  AIC: 515.7
```

Final model

```
# without obs 25 log(Scruz+0.1) and log(Elevation) not significant
# refit using all data but omitting these variables
modp3 <- glm(Species ~ log(Area) + log(Nearest) + log(Adjacent),
family=poisson, gala)
summary(modp3)
```

```
##
## Call:
## glm(formula = Species ~ log(Area) + log(Nearest) + log(Adjacent),
##     family = poisson, data = gala)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5064  -2.9908  -0.3175   2.6705   7.9670
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.397868   0.048637  69.862  < 2e-16 ***
```

```
## log(Area)      0.362669   0.008200  44.229  < 2e-16 ***
## log(Nearest)  -0.061141   0.011695  -5.228 1.71e-07 ***
## log(Adjacent) -0.096593   0.006168 -15.660  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3510.73  on 29  degrees of freedom
## Residual deviance:  367.73  on 26  degrees of freedom
## AIC: 536.56
##
## Number of Fisher Scoring iterations: 5
```