Advanced Data Analysis

Additional Aspects of Generalised Linear Models

Generalised Linear Models

- Very large topic – could easily fill up a course.
- Still relatively young – started in early 1970s.
- Lots of methodology options.
- Quite a bit of jargon.
- Diagnostics more delicate.

Exponential Family

$$\ln[y;\eta] = \frac{y\eta - b(\eta)}{\phi} + c(y,\phi)$$

$\eta = g(\mu)$, where $\mu = E(y)$.

η known as the canonical parameter.

g known as the (canonical) link function.

ϕ known as the dispersion parameter.

Exponential Family Examples

name	canonical link	$b(\eta)$	$c(y,\phi)$	ϕ
Bernoulli	$\eta = \text{logit}(\mu)$	$\log(1+e^{\eta})$	0	1
Poisson	$\eta = \ln(\mu)$	e^{η}	$-\ln(y!)$	1
$N(\mu,\sigma^2)$	$\eta = \mu$	$\eta^2/2$	$(y^2/\sigma^2 - \ln(2\pi\sigma^2))/2$	σ^2

Connections Involving the b Function

$$\mu = E(y) = b'(\eta)$$

$$\text{Var}(y) = \phi b''(\eta)$$

General GLM Fitting with Canonical Link

The log-likelihood is

$$\ell(\beta) = \{y^T X\beta - 1^T b(X\beta)\}/\phi + \text{const.}$$

So need maximimise $S(\beta) = y^T X\beta - 1^T b(X\beta)$.

Newton-Raphson reduces to iteratively reweighted least squares.

$$\widehat{\beta} = \text{estimate of } \beta$$

$$\widehat{\text{se}}(\widehat{\beta}) = \sqrt{\text{diagonal entries of } [X^T \text{diag}\{b''(X\widehat{\beta})/\phi\} X]^{-1}}$$

$\widehat{\eta}$ and $\widehat{\mu}$ Vectors

Convenient notation is

$$\widehat{\eta} = X\widehat{\beta} = \text{vector of fitted values on the link scale}$$

and

$$\widehat{\mu} = g^{-1}(X\widehat{\beta}) = \text{vector of fitted values on the response scale}$$

Deviance

Write the log-likelihood $\ell(\beta)$ as a function of $\mu = g^{-1}(X\beta)$.

e.g. in Bernoulli case

$$\begin{aligned} \ell(\beta) &= y^T X\beta - 1^T \ln(1 + e^{X\beta}) \\ &= y^T \ln(\mu) + (1-y)^T \ln(1-\mu) = \ell_\mu(\mu), \quad \text{(say)} \end{aligned}$$

The deviance is:

$$D = 2\phi\{\ell_\mu(y) - \ell_\mu(\widehat{\mu})\}$$

Examples of Deviance Expressions

Normal: $D = (y - \widehat{\mu})^T (y - \widehat{\mu})$

Bernoulli: $D = 2[y^T \ln(y/\widehat{\mu}) + (1-y)^T \ln\{((1-y)/(1-\widehat{\mu})\}]$

Poisson: $D = 2\{y^T \ln(y/\widehat{\mu}) - 1^T (y - \widehat{\mu})\}$

Interpretation of Deviance

The

smaller the deviance

the

better the model fits the data.

Overfitting Caveat

Can achieve

zero deviance

just by fitting with very high degree polynomials.

Need to trade-off against parsimony.

AIC

This is short for Akaike's Information Criterion

AIC $=$ Deviance$+2\times$no. of parameters$+$const(y).

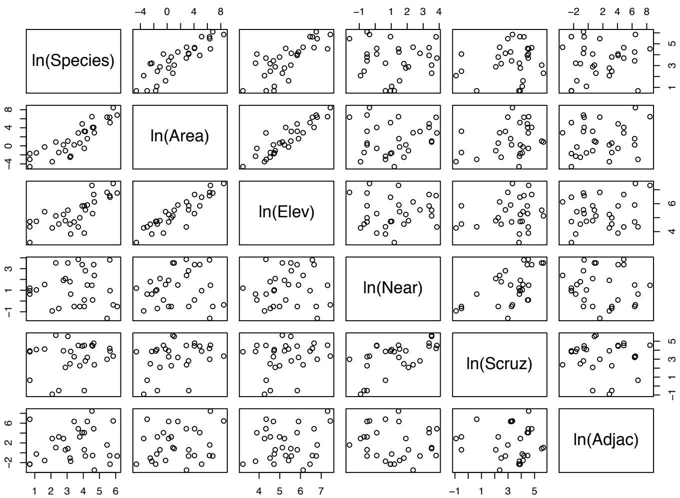
(Generalises Mallow's C_p from linear regression.)

AIC is often used for model selection in GLM.

Galapagos Turtle Example

Species	the number of species of tortoise found on the island
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)

Zoological interesting in modelling Species in terms of predictors.



Initial Model

Since response variable `Species` is a count we try the model:

$$\texttt{Species}_i \sim \text{Poisson}(\exp(\beta_0 + \beta_1 \texttt{Area}_i + \beta_2 \texttt{Elevation}_i + \beta_3 \texttt{Scruz}_i + \beta_4 \texttt{Adjacent}_i))$$


```
glm(formula = Species ~ ., family = poisson, data = galapagos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2752  -4.4966  -0.9443   1.9168  10.1849

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 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:  716.85  on 24  degrees of freedom
AIC: 889.68
```


Model Checking Essential!

For most scientific purposes we

CANNOT STOP HERE

We now have to conduct diagnostic checks on the model.

Diagnostic Checks

These are required to check:

- Linearity assumption.
- Leverage and influence.
- Distributional assumptions

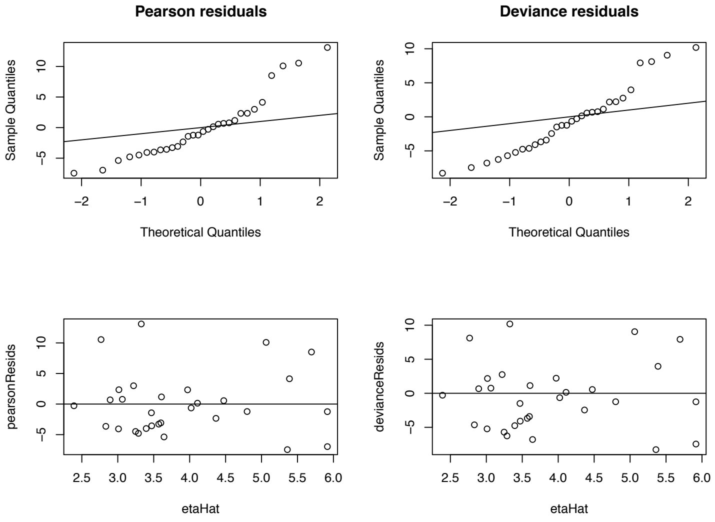
  (e.g. Poisson versus overdispersed).

GLM Residuals

ith Pearson residual: $r_i^P = (y_i - \widehat{\mu}_i)/\sqrt{\phi b''(\widehat{\eta}_i)}$

ith Deviance residual: $r_i^D = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i}$
where $D = \sum_{i=1}^{n} d_i$

If model is correct then these should each behave approximately like a
$N(0,1)$ sample.



Comments on Previous Slide

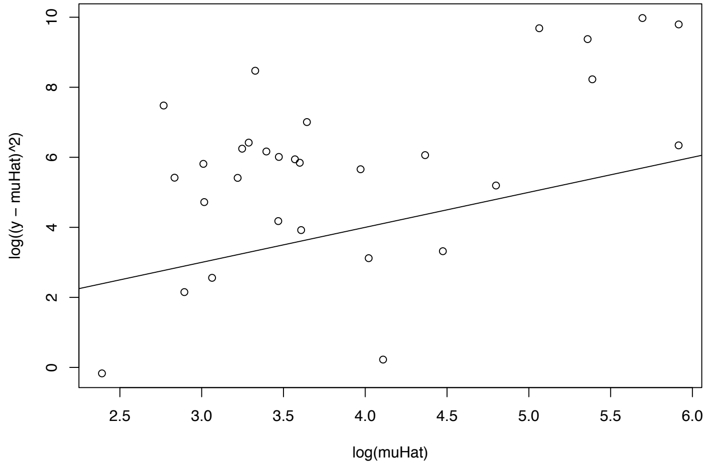
Upper Panels

These are normal Q-Q plots with a 1:1 line added.

Both Pearson and deviance residuals are much more dispersed (large spread) than $N(0,1)$ data indicating strong violation of Poisson assumption.

Lower Panels

These are the residuals against $\widehat{\eta}_i$ values (with zero line added). These show no strong patterns.



Comments on Previous Slide

This is a plot of

$$\ln\{(y_i - \widehat{\mu}_i)^2\} \quad \text{versus} \quad \ln(\widehat{\mu}_i)$$

with a 1:1 line added. The fact that the most points are well above the line indicates an overdispersion problem.

Overdispersion

Underpinning overdispersion is the reality:

Just because my data is a count variable,

doesn't mean that it is Poisson!

Count Data Example I

Following data are number of Prussian soldiers kicked to death by their horse for each unit and year in the period 1875–1894.

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	3	3	3	4																						

One can show that the Poisson model fits these data well:

$$[x; \lambda] = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, \ldots$$

Prussian Horse Kick Deaths (cont.)

Note that

sample mean = 0.622

sample variance = 0.636 $\simeq$ sample mean

which is consistent with well-known result:

$E(X) = \text{Var}(X) = \lambda$ result for $X \sim \text{Poisson}(\lambda)$

Count Data Example II

These data are daily number of respiratory deaths in Milan in 1980:

2	5	0	1	1	6	2	4	1	4	2	4	3	5	5	3	3	3	3	5	1	4	4	3	5
4	6	3	4	3	5	7	1	1	3	5	1	2	6	3	6	5	6	5	2	4	3	3	9	2
7	1	5	7	5	2	5	8	5	7	4	4	6	3	9	12	10	4	8	4	5	6	8	2	3
3	2	3	5	8	5	2	3	1	0	3	3	4	6	4	1	4	4	6	3	5	1	4	7	3
3	6	2	3	6	3	4	3	4	4	1	1	1	1	4	3	6	2	2	1	4	1	3	4	1
3	5	4	3	6	1	2	5	4	4	3	1	1	0	1	1	1	0	4	1	2	0	2	1	1
2	4	2	0	3	3	2	3	5	2	4	6	1	3	2	5	0	6	5	2	2	3	2	3	4
5	1	1	0	2	1	1	1	1	2	0	3	3	4	3	3	0	1	2	1	0	1	2	1	0
0	2	5	1	1	2	2	0	2	3	3	2	1	1	5	9	4	3	1	8	5	6	4	4	4
2	4	2	0	1	2	3	1	2	2	1	3	1	1	0	0	2	1	1	1	2	3	2	2	0
2	3	1	2	1	1	5	2	3	2	0	2	4	2	5	2	2	1	3	0	2	0	2	3	5
2	1	0	1	3	3	2	0	1	2	1	2	1	2	1	3	2	0	2	3	4	1	2	1	3
3	2	5	2	5	2	1	2	1	2	3	5	3	2	0	1	4	8	4	1	2	7	3	3	1
4	5	0	5	1	5	2	3	2	4	1	3	8	2	4	2	2	1	6	5	2	2	2	4	3
5	6	0	4	3	3	5	2	3	1	2	5	5	4	2										

sample mean = 2.9:   sample variance = 4.1 $>$ 2.9.

This indicates incompatibility with the Poisson distribution.
The data are (mildly) overdispersed.

Quasi-likelihood

A remedy is quasi-likelihood where the Poisson log-likeliood

$$\ell(\beta) = y^T X\beta - 1^T e^{X\beta}$$

is replaced by

$$q\ell(\beta) = \frac{y^T X\beta - 1^T e^{X\beta}}{\phi}$$

where ϕ is estimated from the data. Note that $q\ell$ does not correspond to a 'proper' probability distribution (hence the quasi label).


```
Call:
glm(formula = Species ~ ., family = quasipoisson, data = galapagos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2752  -4.4966  -0.9443   1.9168  10.1849

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.1548079  0.2915901  10.819 1.03e-10 ***
Area        -0.0005799  0.0001480  -3.918 0.000649 ***
Elevation    0.0035406  0.0004925   7.189 1.98e-07 ***
Nearest      0.0088256  0.0102622   0.860 0.398292
Scruz       -0.0057094  0.0035251  -1.620 0.118380
Adjacent    -0.0006630  0.0001653  -4.012 0.000511 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```



```
(Dispersion parameter for quasipoisson family taken to be 31.74921)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Comment on Previous Output

The over-dispersion parameter is estimated as 31.74921

The coefficient estimates haven't changed much

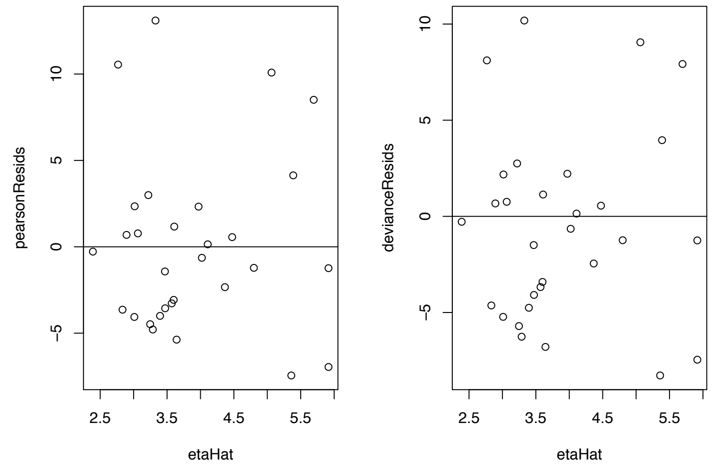
BUT

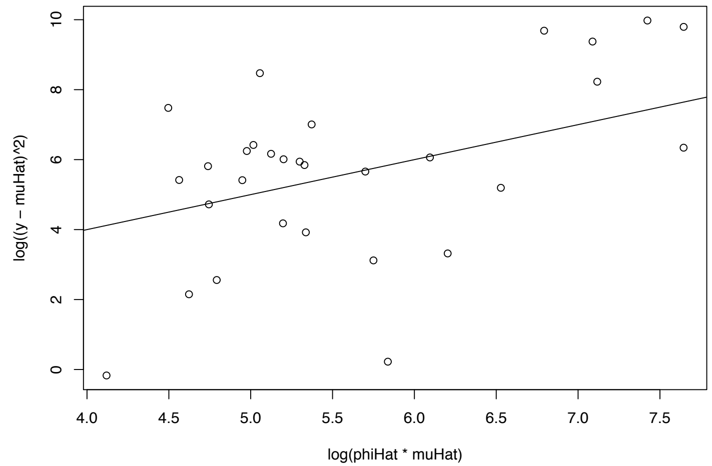
the standard errors have changed enormously!

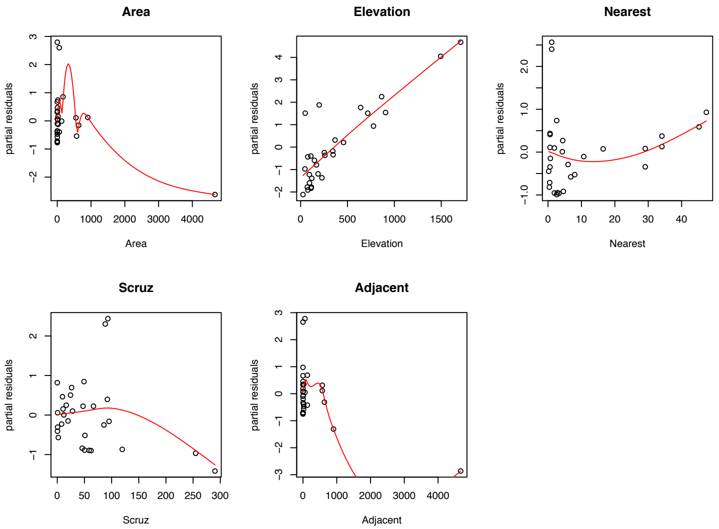
This can have a big effect on statist. significance.

Diagnosis of New Model

The following figures show residual plots for the quasi-likelihood extension:

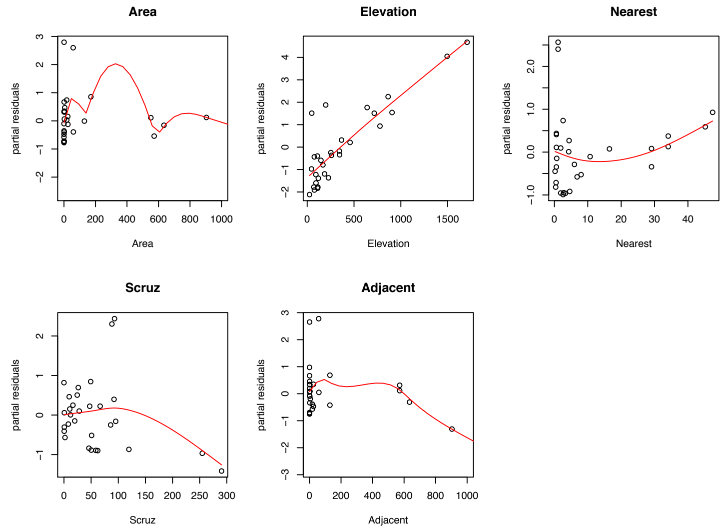






Two of the panels in the previous slide are distorted by a single high value on the x-axis.

We will re-do these plots with a restricted range.



Comment on Previous Graphic

These are partial residual plots for each predictor.

Linearity of the red curve indicates that the linearity assumption is reasonable.

e.g. Elevation looks fine, but Adjacent appears to have non-linear a effect.

(Nearest and Scruz also look non-linear; but are not significant).

Even though the red curve for Area looks weird, the points are reasonably well-behaved (on a straight line).

Later in this course we will look at ways to address the non-linear effects.

Other GLMs

With the aim of fitting data better, several other GLMs have emerged. These include:

- Negative binomial regression models for count data.
- Quasi-likelihood models with variance modelling.

- Models that include offsets.

Model Selection

AIC-based strategies for ‘automatic’ selection of a model are now available.

In `R` the relevant function is `step()`.

(Assignment 6 illustrates `step()`)

Further reading

Faraway, J.J. (2006). *Extending the Linear Model with R*, Chapman & Hall/CRC.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models (Second Edition)*. London: Chapman and Hall.

Practical GLM Beyond Advanced Data Analysis

Recommendation if you have to do lots of GLM analyses in a future role (e.g. job, thesis research):

- Spend $100 on the Faraway book.
- Work through its GLM examples.