Modern Regression Methods

Anthony Davison

http://stat.epfl.ch

Dictionary

□ **Regression**: (statistics) a measure of the relation between the mean value of one variable (e.g., output) and corresponding values of other variables (e.g., time and cost).

□ Regression: the dependence of a **response variable**, y, (treated as random) on **explanatory variables**, x, (treated as fixed):

$$Y \sim f(y;\theta,x),$$

where f is a (usually) known density that depends on parameters θ.

□ Classical normal linear model:

$$Y \sim N(x^{\mathrm{T}}\beta,\sigma^2), \quad \theta=(\beta,\sigma^2)\in\mathbb{R}^p\times\mathbb{R}_+.$$

Linear model

□ Data consist of observations $(x_1,y_1),\ldots,(x_n,y_n)$

□ The y_j are assumed to be realisations of observations Y_j, such that we can write

$$Y_j = x_j^{\mathrm{T}}\beta+\sigma\varepsilon_j, \quad j=1,\ldots,n,$$

or, with X having jth row x_j^{T},

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1}+\varepsilon_{n\times 1},$$

where the 'errors' satisfy

$$\mathrm{E}(\varepsilon)=0_{n\times 1}, \quad \mathrm{var}(\varepsilon)=I_n,$$

□ Two distributional assumptions in general use:

- **second-order assumptions**, $\mathrm{E}(\varepsilon)=0$, $\mathrm{var}(\varepsilon)=I_n$, as above;
- **normal assumptions**, $\varepsilon\sim N_n(0,I_n)$, giving a (mostly) exact distribution theory

□ Under above assumptions have elegant theory of minimum variance unbiased estimation, tests, analysis of variance, confidence intervals, model-checking (residuals), ..., all linked to the geometry of projections in $\mathbb{R}^n$

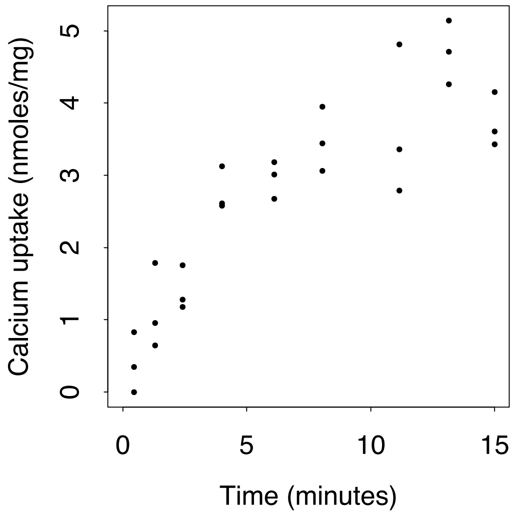
□ Can extend to weighted least squares, with $\mathrm{var}(\varepsilon)=V$, to robust fitting, penalised estimators, etc.

Calcium data

Table 1: Calcium uptake (nmoles/mg) of cells suspended in a solution of radioactive calcium, as a function of time suspended (minutes).

Time (minutes)	Calcium uptake (nmoles/mg)		
0.45	0.34170	–0.00438	0.82531
1.30	1.77967	0.95384	0.64080
2.40	1.75136	1.27497	1.17332
4.00	3.12273	2.60958	2.57429
6.10	3.17881	3.00782	2.67061
8.05	3.05959	3.94321	3.43726
11.15	4.80735	3.35583	2.78309
13.15	5.13825	4.70274	4.25702
15.00	3.60407	4.15029	3.42484

Calcium data



Calcium data

□ Possible differential equation describing the uptake is

$$\frac{dy}{dx} = (\beta_0 - y)/\beta_1,$$

with $y = 0$ when $x = 0$, and solution $\beta_0\{1 - \exp(-x/\beta_1)\}$.

□ Hence model

$$y = \beta_0\{1 - \exp(-x/\beta_1)\} + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$ represents measurement error.

□ Can write this as

$$\begin{aligned} y_j &\sim N\{\mu(\beta; x_j), \sigma^2\}, \\ \mu(\beta; x) &= \beta_0\{1 - \exp(-x/\beta_1)\}, \quad j = 1, \ldots, n: \end{aligned}$$

a nonlinear model with normal response distribution.

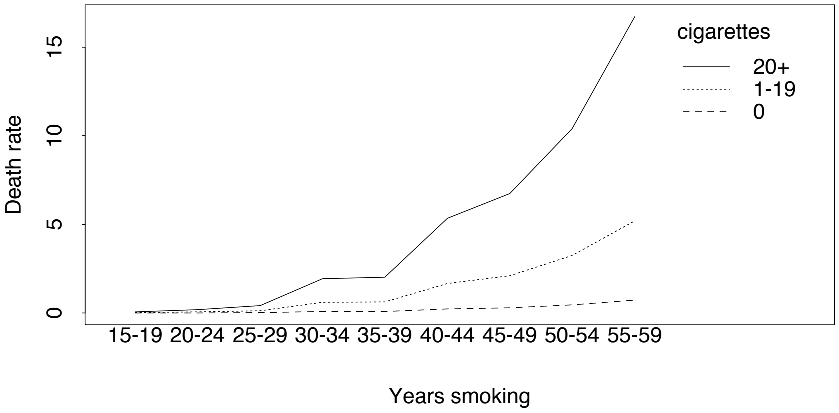
Smoking data

Table 2: Lung cancer deaths in British male physicians (Doll and Hill, 1952). The table gives man-years at risk T/number of cases y of lung cancer, cross-classified by years of smoking t, taken to be age minus 20 years, and number of cigarettes smoked per day, d.

Years of smoking t	Daily cigarette consumption d						
	Nonsmokers	1–9	10–14	15–19	20–24	25–34	35+
15–19	10366/1	3121	3577	4317	5683	3042	670
20–24	8162	2937	3286/1	4214	6385/1	4050/1	1166
25–29	5969	2288	2546/1	3185	5483/1	4290/4	1482
30–34	4496	2015	2219/2	2560/4	4687/6	4268/9	1580/4
35–39	3512	1648/1	1826	1893	3646/5	3529/9	1336/6
40–44	2201	1310/2	1386/1	1334/2	2411/12	2424/11	924/10
45–49	1421	927	988/2	849/2	1567/9	1409/10	556/7
50–54	1121	710/3	684/4	470/2	857/7	663/5	255/4
55–59	826/2	606	449/3	280/5	416/7	284/3	104/1

Smoking data

Lung cancer deaths in British male physicians. The figure shows the rate of deaths per 1000 man-years at risk, for each of three levels of daily cigarette consumption.



Smoking data

□ Suppose number of deaths y has Poisson distribution, mean $T\lambda(d,t)$, where T is man-years at risk, d is number of cigarettes smoked daily and t is time smoking (years).

□ Take

$$\lambda(d,t) = \beta_0 t^{\beta_1}\left(1+\beta_2 d^{\beta_3}\right) :$$

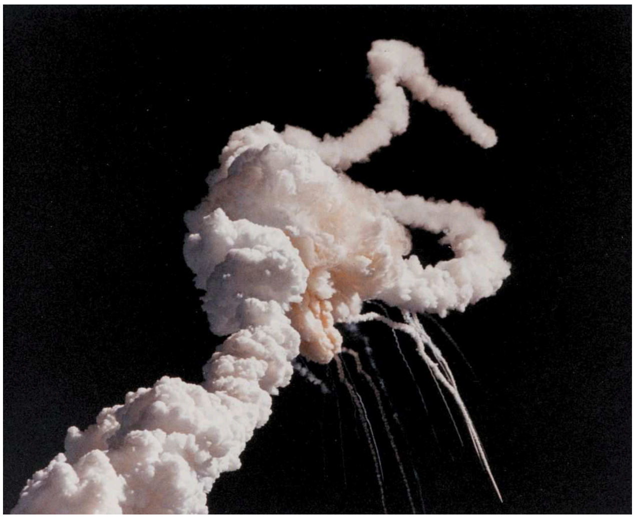
- background rate of lung cancer is $\beta_0 t^{\beta_1}$ for non-smoker,
- additional risk due to smoking d cigarettes/day is $\beta_2 d^{\beta_3}$.

□ With $x_j = (T_j, d_j, t_j)$, can write this as

$$\begin{aligned} y_j &\sim \operatorname{Poiss}\{\mu(\beta; x_j)\}, \\ \mu(\beta; x) &= T\beta_0 t^{\beta_1}\left(1+\beta_2 d^{\beta_3}\right), \quad j = 1, \ldots, n : \end{aligned}$$

a nonlinear model with Poisson-distributed response.

Challenger data



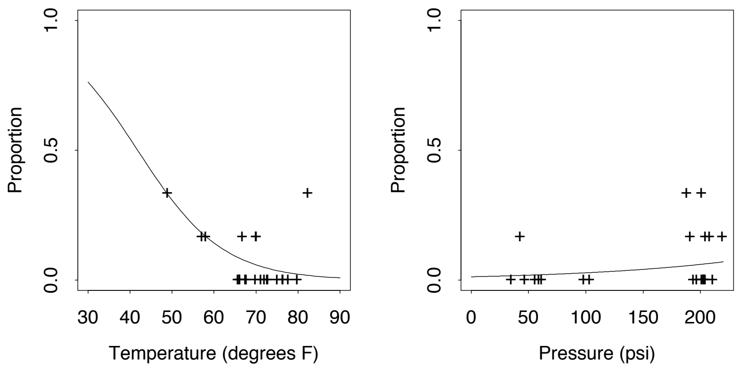
Challenger data

Table 3: O-ring thermal distress data. r is the number of field-joint O-rings showing thermal distress out of 6, for a launch at the given temperature (°F) and pressure (pounds per square inch)

Flight	Date	Number of O-rings with thermal distress, r	Temperature (°F) x_1	Pressure (psi) x_2
1	21/4/81	0	66	50
2	12/11/81	1	70	50
$\vdots$				
51-F	29/7/85	0	81	200
51-I	27/8/85	0	76	200
51-J	3/10/85	0	79	200
61-A	30/10/85	2	75	200
61-B	26/11/86	0	76	200
61-C	21/1/86	1	58	200
61-I	28/1/86	—	31	200

Challenger data

Figure 1: O-ring thermal distress data. The left panel shows the proportion of incidents as a function of joint temperature, and the right panel shows the corresponding plot against pressure. The x-values have been jittered to avoid overplotting multiple points. The solid lines show the fitted proportions of failures under a logistic regression model.



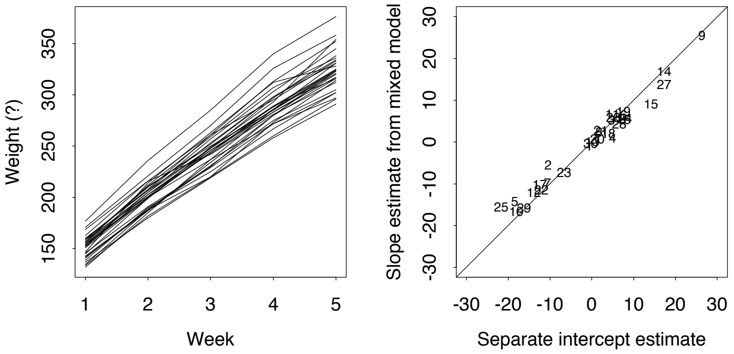
Rat growth data

Table 4: Weights (units unknown) of 30 young rats over a five-week period

	Week						Week				
	1	2	3	4	5		1	2	3	4	5
1	151	199	246	283	320	16	160	207	248	288	324
2	145	199	249	293	354	17	142	187	234	280	316
3	147	214	263	312	328	18	156	203	243	283	317
4	155	200	237	272	297	19	157	212	259	307	336
5	135	188	230	280	323	20	152	203	246	286	321
6	159	210	252	298	331	21	154	205	253	298	334
7	141	189	231	275	305	22	139	190	225	267	302
8	159	201	248	297	338	23	146	191	229	272	302
9	177	236	285	340	376	24	157	211	250	285	323
10	134	182	220	260	296	25	132	185	237	286	331
11	160	208	261	313	352	26	160	207	257	303	345
12	143	188	220	273	314	27	169	216	261	295	333
13	154	200	244	289	325	28	157	205	248	289	316
14	171	221	270	326	358	29	137	180	219	258	291
15	163	216	242	281	312	30	153	200	244	286	324

Rat growth data

Figure 2: Rat growth data. Left: weekly weights of 30 young rats. Right: shrinkage of individual slope estimates towards overall slope estimate; the solid line has unit slope, and the estimates from the mixed model lie slightly closer to zero than the individual estimates.



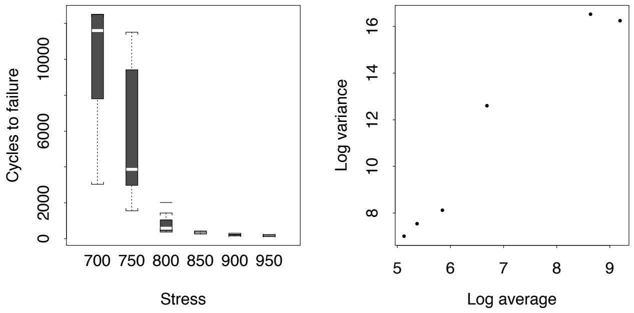
Spring failure data

Table 5: Failure times (in units of 10^3 cycles) of springs at cycles of repeated loading under the given stress . + indicates that an observation is right-censored. The average and estimated standard deviation for each level of stress are $\overline{y}$ and s.

	Stress (N/mm^2)					
	950	900	850	800	750	700
	225	216	324	627	3402	12510+
	171	162	321	1051	9417	12505+
	198	153	432	1434	1802	3027
	189	216	252	2020	4326	12505+
	189	225	279	525	11520+	6253
	135	216	414	402	7152	8011
	162	306	396	463	2969	7795
	135	225	379	431	3012	11604+
	117	243	351	365	1550	11604+
	162	189	333	715	11211	12470+
$\overline{y}$	168	215	348	803	5636	9828
s	33	43	58	544	3864	3355

Spring failure data

Figure 3: Failure times (in units of 10^3 cycles) of springs at cycles of repeated loading under the given stress. The left panel shows failure time boxplots for the different stresses. The right panel shows a rough linear relation between log average and log variance at the different stresses.



Motivation

□ Normal linear model $y = X\beta + \varepsilon$

- applicable for continuous response $y \in \mathbb{R}$
- assumes linear dependence of mean response $\mathrm{E}(y)$ on covariates X
- assumes $y \sim$ normal, constant variance σ^2, uncorrelated

□ Most data not like this

□ Need extensions for

- nonlinear dependence on covariates
- arbitrary response distribution (binomial? Poisson? exponential? ...)
- dependent responses
- variance non-constant (and related to mean?)
- censoring, truncation, ...
- layers of variation
- basis function expansions
- ...

Simple fixes

□ Just fit a linear model anyway

- Might work as an approximation, but usually extrapolates really badly.

□ Fit a linear model to transformed responses

- E.g., take variance-stabilising transformation for y, such as $2\sqrt{y}$ when y is Poisson
- Can be helpful, but usually the obvious transformation can't give linearity.

□ Instead we attempt to fit the model using likelihood estimation.

Likelihood

Definition 1 *Let y be a data set, assumed to be the realisation of a random variable $Y \sim f(y;\theta)$, where the unknown parameter θ lies in the parameter space $\Omega_\theta \subset \mathbb{R}^p$. Then the* **likelihood** *(for θ based on y) and the corresponding* **log likelihood** *are*

$$L(\theta) = L(\theta; y) = f_Y(y;\theta), \quad \ell(\theta) = \log L(\theta), \quad \theta \in \Omega_\theta.$$

The **maximum likelihood estimate** *(MLE) $\widehat{\theta}$ satisfies $\ell(\widehat{\theta}) \geq \ell(\theta)$, for all $\theta \in \Omega_\theta$.*
Often $\widehat{\theta}$ is unique and in many cases it satisfies the **score (or likelihood) equation**

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0,$$

which is interpreted as a vector equation of dimension $p \times 1$ if θ is a $p \times 1$ vector.
The **observed information** *and* **expected (Fisher) information** *are defined as*

$$J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}, \quad I(\theta) = \mathrm{E}\left\{J(\theta)\right\};$$

these are $p \times p$ matrices if θ has dimension p.

Maximum likelihood estimator

□ In large samples from a **regular model** in which the true parameter is $\theta^0_{p\times 1}$, the maximum likelihood estimator $\widehat{\theta}$ has an approximate normal distribution,

$$\widehat{\theta} \mathrel{\dot\sim} \mathcal{N}_p\left\{\theta^0, J(\widehat{\theta})^{-1}\right\},$$

so we can compute an approximate $(1-2\alpha)$ confidence interval for the rth parameter θ^0_r as

$$\widehat{\theta}_r \pm z_\alpha v_{rr}^{1/2},$$

where v_{rr} is the rth diagonal element of the matrix $J(\widehat{\theta})^{-1}$.

□ This is easily implemented:

- we code the negative log likelihood $-\ell(\theta)$ (and check the code carefully!);
- we minimise $-\ell(\theta)$ numerically, ensuring that the minimisation routine returns $\widehat{\theta}$ and the Hessian matrix $J(\widehat{\theta}) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^{\mathrm{T}}|_{\theta=\widehat{\theta}}$
- we compute $J(\widehat{\theta})^{-1}$, and use the square roots of its diagonal elements, $v_{11}^{1/2}, \ldots, v_{dd}^{1/2}$, as standard errors for the corresponding elements of $\widehat{\theta}$.

Aside: Regular model

We say that a statistical model $f(y;\theta)$ is **regular (for likelihood inference)** if

1. the true value θ^0 of θ is interior to the parameter space $\Omega_\theta \subset \mathbb{R}^p$;
2. the densities defined by any two different values of θ are distinct;
3. there is an open set $\mathcal{I} \subset \Omega_\theta$ containing θ^0 within which the first three derivatives of the log likelihood with respect to elements of θ exist almost surely, and

$$|\partial^3 \log f(Y_j;\theta)/\partial\theta_r\partial\theta_s\partial\theta_t| \le g(Y_j)$$

uniformly for $\theta \in \mathcal{I}$, where $0 < \mathrm{E}_0\{g(Y_j)\} = K < \infty$; and
4. for $\theta \in \mathcal{I}$ we can interchange differentation with respect to θ and integration, that is,

$$\frac{\partial}{\partial\theta}\int f(y;\theta)\,dy = \int \frac{\partial f(y;\theta)}{\partial\theta}\,dy, \quad \frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\int f(y;\theta)\,dy = \int \frac{\partial^2 f(y;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\,dy.$$

The results are also true under weaker conditions, for non-identically distributed and dependent data.

Aside: Comments on regular models

Condition

1. is needed so that $\widehat{\theta}$ can lie 'on both sides' of θ^0 and hence can have a limiting normal distribution, once standardized—**fails**, for example, if θ has a discrete component (e.g. changepoint $\gamma \in \{1,\ldots,n\}$);
2. is needed to be able to identify the model on the basis of the data;
3. ensures the validity of Taylor series expansions of $\ell(\theta)$—not usually a problem;
4. ensures that the score statistic has a limiting normal distribution—can **fail** in some models — sometimes good news, leading to faster convergence than $n^{-1/2}$.

All the above assumes the postulated model is correct! — there is a literature on what happens when we fit the wrong model, or if the parameter dimension increases with n, or ... usually there are no generic results for such cases.

Likelihood ratio statistic

□ Model $f_B(y)$ is **nested** within model $f_A(y)$ if A reduces to B on restricting some parameters:

- for example, the model $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ is nested within the model $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, because the first is obtained from the second by setting $\mu = 0$;
- the maximised log likelihoods satisfy $\widehat{\ell}_A \geq \widehat{\ell}_B$, because the more comprehensive model A contains the simpler model B.

□ The **likelihood ratio statistic** for comparing them is

$$W = 2(\widehat{\ell}_A - \widehat{\ell}_B).$$

□ If the model is regular, the simpler model is true, and A has q more parameters than B, then

$$W \dot{\sim} \chi^2_q.$$

□ This implicitly assumes that ML inference for model A is OK, so that the approximation $\widehat{\theta}_A \dot{\sim} \mathcal{N}\{\theta_A, J_A(\widehat{\theta}_A)^{-1}\}$ is adequate.

Profile log likelihood

□ Consider a regular log likelihood $\ell(\psi, \lambda)$, where the **parameter of interest** ψ is variation independent of the **nuisance parameter** λ, i.e., $(\psi, \lambda) \in \Omega_\psi \times \Omega_\lambda$, and the overall MLE is $(\widehat{\psi}, \widehat{\lambda})$.

□ For a confidence set for ψ, without reference to λ, we use the **profile log likelihood**

$$\ell_{\mathrm{p}}(\psi) = \max_{\lambda \in \Omega_\lambda} \ell(\psi, \lambda) = \ell(\psi, \widehat{\lambda}_\psi),$$

say, and, based on the limiting distribution of the likelihood ratio statistic, take as $(1 - 2\alpha)$ confidence region the set

$$\left\{ \psi \in \Omega_\psi : 2\{\ell(\widehat{\psi}, \widehat{\lambda}) - \ell(\psi, \widehat{\lambda}_\psi)\} \leq \chi^2_{\dim\psi}(1 - 2\alpha) \right\}.$$

□ When ψ is scalar, this yields

$$\left\{ \psi \in \Omega_\psi : \ell(\psi, \widehat{\lambda}_\psi)\} \geq \ell(\widehat{\psi}, \widehat{\lambda}) - \tfrac{1}{2}\chi^2_1(1 - 2\alpha) \right\},$$

and $\frac{1}{2}\chi^2_1(0.95) = 1.92$.

□ Such intervals are generally better than the standard interval $\widehat{\psi} \pm z_\alpha \mathrm{SE}$, particularly when the distribution of $\widehat{\psi}$ is asymmetric, but require more computation, since they involve many maximisations of ℓ.

Iterative weighted least squares slide 27

Model setup

□ Independent random variables $Y_1, \dots, Y_n$, with observed values $y_1, \dots, y_n$, and covariates $x_1, \dots, x_n$.

□ Suppose that probability density of Y_j is $f(y_j; \eta_j, \phi)$, where $\eta_j = \eta(\beta, x_j)$, and ϕ is common to all models.

□ Log likelihood is

$$\ell(\beta, \phi) = \sum_{j=1}^{n} \ell_j(\beta, \phi) = \sum_{j=1}^{n} \log f\{y_j; \eta(\beta, x_j), \phi\}.$$

□ More generally, just let $\ell_j(\beta, \phi)$ denote the log likelihood contribution from the $j\mathrm{th}$ observation.

□ Suppose ϕ known (for now), suppress it, and estimate β.

Example 2 (Normal linear model) *Express the normal linear model in the terms above.*

Note to Example 2

A simple calculation gives

$$\eta_j = x_j^{\mathrm{T}}\beta, \quad \phi = \sigma^2, \quad \ell_j \equiv -\tfrac{1}{2}\{(y_j - \eta_j)^2/\phi + \log \phi\}.$$

Iterative weighted least squares (IWLS)

□ General approach for estimation in regression models, based on Newton–Raphson iteration

□ Assume that ϕ is fixed, and write

$$\ell(\beta) = \sum_{j=1}^{n} \ell_j\{\eta_j(\beta)\}.$$

□ MLEs $\widehat{\beta}$ usually satisfy

$$\frac{\partial \ell(\widehat{\beta})}{\partial \beta_r} = 0, \qquad r = 1, \dots, p,$$

or equivalently

$$\frac{\partial \ell(\widehat{\beta})}{\partial \beta} = \frac{\partial \eta^{\mathrm{T}}}{\partial \beta} \frac{\partial \ell}{\partial \eta} = \frac{\partial \eta^{\mathrm{T}}}{\partial \beta} u(\widehat{\beta}) = 0, \tag{1}$$

where $u(\beta)$ is $n \times 1$ vector with $j\mathrm{th}$ element $\partial \ell / \partial \eta_j$.

Derivation of IWLS algorithm

□ To find the maximum likelihood estimate $\widehat{\beta}$ starting from a trial value β, we make a Taylor series expansion in (1), to obtain

$$\frac{\partial \eta^{\mathrm{T}}(\beta)}{\partial \beta} u(\beta) + \left\{ \sum_{j=1}^{n} \frac{\partial \eta_j(\beta)}{\partial \beta} \frac{\partial^2 \ell_j(\beta)}{\partial \eta_j^2} \frac{\partial \eta_j(\beta)}{\partial \beta^{\mathrm{T}}} + \sum_{j=1}^{n} \frac{\partial^2 \eta_j(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}} u_j(\beta) \right\} (\widehat{\beta} - \beta) \doteq 0. \qquad (2)$$

If we denote the $p \times p$ matrix in braces on the left by the $p \times p$ matrix $-J(\beta)$, assumed invertible, we can rearrange (2) to obtain

$$\widehat{\beta} \doteq \beta + J(\beta)^{-1} \frac{\partial \eta^{\mathrm{T}}(\beta)}{\partial \beta} u(\beta). \qquad (3)$$

This suggests that maximum likelihood estimates may be obtained by starting from a particular β, using (3) to obtain $\widehat{\beta}$, then setting β equal to $\widehat{\beta}$, and iterating (3) until convergence. This is the Newton–Raphson algorithm applied to our particular setting. In practice it can be more convenient to replace $J(\beta)$ by its expected value

$$I(\beta) = \sum_{j=1}^{n} \frac{\partial \eta_j(\beta)}{\partial \beta} \mathrm{E}\left(-\frac{\partial^2 \ell_j}{\partial \eta_j^2} \right) \frac{\partial \eta_j(\beta)}{\partial \beta^{\mathrm{T}}};$$

the other term vanishes because $\mathrm{E}\{u_j(\beta)\} = 0$. We write

$$I(\beta) = X(\beta)^{\mathrm{T}} W(\beta) X(\beta), \qquad (4)$$

where $X(\beta)$ is the $n \times p$ matrix $\partial \eta(\beta)/\partial \beta^{\mathrm{T}}$ and $W(\beta)$ is the $n \times n$ diagonal matrix whose jth diagonal element is $\mathrm{E}(-\partial^2 \ell_j / \partial \eta_j^2)$.

□ If we replace $J(\beta)$ by $X(\beta)^{\mathrm{T}} W(\beta) X(\beta)$ and reorganize (3), we obtain

$$\widehat{\beta} = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W (X\beta + W^{-1} u) = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W z, \qquad (5)$$

say, where the dependence of the terms on the right on β has been suppressed. That is, starting from β, the updated estimate $\widehat{\beta}$ is obtained by weighted linear regression of the $n \times 1$ vector **adjusted dependent variable**

$$z = X(\beta)\beta + W(\beta)^{-1} u(\beta)$$

on the columns of $X(\beta)$, using weight matrix $W(\beta)$. The maximum likelihood estimates are obtained by repeating this step until the log likelihood, the estimates, or more often both are essentially unchanged. The variable z plays the role of the response or dependent variable in the weighted least squares step.

□ Often the structure of a model simplifies the estimation of an unknown value of ϕ. It may be estimated by a separate step between iterations of $\widehat{\beta}$, by including it in the step (3), or from the profile log likelihood $\ell_{\mathrm{p}}(\phi)$.

IWLS II

- □ Newton–Raphson update step:
$$\widehat{\beta} = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wz,$$
where
$$\begin{aligned} X_{n\times p} &= \partial\eta/\partial\beta^{\mathrm{T}}, \quad \text{(design matrix)}\\ W_{n\times n} &= \mathrm{diag}\{\mathrm{E}(-\partial^2\ell_j/\partial\eta_j^2)\}, \quad \text{(weights)}\\ z_{n\times 1} &= X\beta + W^{-1}u, \quad \text{(adjusted dependent variable)}\end{aligned}$$
- □ Thus to obtain MLEs $\widehat{\beta}$ we use the **IWLS algorithm**:
- □ take an initial $\widehat{\beta}$. Repeat
  - compute X, W, u, z;
  - compute new $\widehat{\beta}$;

  until changes in $\ell(\widehat{\beta})$ (or, sometimes, $\widehat{\beta}$, or both) are lower than some tolerance.
- □ Sometimes a line search is added, if $\ell(\widehat{\beta}_{\mathrm{new}}) < \ell(\widehat{\beta}_{\mathrm{old}})$: i.e., we half the step length and try again.

Examples

Example 3 (Normal linear model) *Give the components of the IWLS algorithm for the normal linear model.*

Example 4 (Normal nonlinear model) *Give the components of the IWLS algorithm for the normal nonlinear model.*

Example 5 (Gumbel linear model) *Give the components of the IWLS algorithm for fitting the linear model*
$$y_j = \beta_0 + \beta_1(x_j - \overline{x}) + \tau\varepsilon_j, \quad j = 1,\ldots,n,$$
with Gumbel errors having density function
$$f(y_j;\eta_j,\tau) = \tau^{-1}\exp\left\{-\frac{y_j-\eta_j}{\tau} - \exp\left(-\frac{y_j-\eta_j}{\tau}\right)\right\},$$
where $\tau > 0$ and $\eta_j = \beta_0 + \beta_1(x_j - \overline{x})$; this distribution is natural for maxima; note that τ^2 is not the variance.

Note to Example 11

- □ In the normal linear model, we write $\eta_j = x_j^{\mathrm{T}}\beta$. If the y_j are independently normally distributed with means η_j and variances $\phi = \sigma^2$, we have

$$\ell_j(\eta_j, \sigma^2) \equiv -\tfrac{1}{2}\left\{\log\sigma^2 + \frac{1}{\sigma^2}(y_j - \eta_j)^2\right\},$$

so

$$u_j(\eta_j) = \frac{\partial\ell_j}{\partial\eta_j} = \frac{1}{\sigma^2}(y_j - \eta_j), \qquad \frac{\partial^2\ell_j}{\partial\eta_j^2} = -\frac{1}{\sigma^2};$$

the $j\mathrm{th}$ element on the diagonal of W is the constant σ^{-2}. The (j, r) element of the matrix $\partial\eta/\partial\beta^{\mathrm{T}}$ is $\partial\eta_j/\partial\beta_r = x_{jr}$, so $X(\beta)$ is simply the $n \times p$ design matrix X. We see that

$$z = X(\beta)\beta + W^{-1}(\beta)u(\beta) = y,$$

because in this situation $X(\beta)\beta = X\beta$ and $W^{-1}(\beta)u(\beta) = \sigma^2(y - X\beta)/\sigma^2$.

- □ Here iterative weighted least squares converges in a single step.
- □ The maximum likelihood estimate of σ^2 is $\widehat{\sigma}^2 = SS(\widehat{\beta})/n$, where $SS(\beta)$ is the sum of squares $(y - X\beta)^{\mathrm{T}}(y - X\beta)$.

Note to Example 4

- □ Here the mean of the $j\mathrm{th}$ observation is $\eta_j = \eta_j(\beta)$. The log likelihood contribution $\ell_j(\eta_j)$ is the same as in the previous example, so u and W are the same also. However, the $j\mathrm{th}$ row of the matrix $X = \partial\eta/\partial\beta^{\mathrm{T}}$ is $(\partial\eta_j/\partial\beta_0, \ldots, \partial\eta_j/\partial\beta_{p-1})$, and as η_j is nonlinear as a function of β, X depends on β. After some simplification, we see that the new value for $\widehat{\beta}$ given by (5) is

$$\widehat{\beta} \doteq (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(X\beta + y - \eta), \tag{6}$$

where X and η are evaluated at the current β. Here $\eta \neq X\beta$ and (6) must be iterated.

- □ The log likelihood is a function of β only through the sum of squares, $SS(\beta) = \sum_{j=1}^{n}\{y_j - \eta_j(\beta)\}^2$. The profile log likelihood for σ^2 is

$$\ell_{\mathrm{p}}(\sigma^2) = \max_{\beta}\ell(\beta, \sigma^2) \equiv -\tfrac{1}{2}\left\{n\log\sigma^2 + SS(\widehat{\beta})/\sigma^2\right\},$$

so the maximum likelihood estimator of σ^2 is $\widehat{\sigma}^2 = SS(\widehat{\beta})/n$. Although $S^2 = SS(\widehat{\beta})/(n - p)$ is not unbiased when the model is nonlinear, it turns out to have smaller bias than $\widehat{\sigma}^2$, and is preferable in applications.

- □ In some cases the error variance depends on covariates, and we write the variance of the $j\mathrm{th}$ response as $\sigma_j^2 = \sigma^2(x_j, \gamma)$. Such models may be fitted by alternating iterative weighted least squares updates for β treating γ as fixed at a current value with those for γ with β fixed, convergence being attained when neither estimates nor log likelihood change materially.

Note to Example 5

□ As the data are annual maxima, it is more appropriate to suppose that y_j has the Gumbel density

$$f(y_j;\eta_j,\tau)=\tau^{-1}\exp\left\{-\frac{y_j-\eta_j}{\tau}-\exp\left(-\frac{y_j-\eta_j}{\tau}\right)\right\}, \tag{7}$$

where τ is a scale parameter and $\eta_j=\beta_0+\beta_1(x_j-\overline{x})$; here we have replaced the γs with βs for continuity with the general discussion above.

□ In this case

$$\ell_j(\eta_j,\tau)=-\log\tau-\frac{y_j-\eta_j}{\tau}-\exp\left(-\frac{y_j-\eta_j}{\tau}\right), \tag{8}$$

and it is straightforward to establish that

$$\frac{\partial\ell_j(\eta_j,\tau)}{\partial\eta_j}=\tau^{-1}\left\{1-\exp\left(-\frac{y_j-\eta_j}{\tau}\right)\right\},\quad \mathrm{E}\left\{-\frac{\partial^2\ell_j(\eta_j,\tau)}{\partial\eta_j^2}\right\}=\tau^{-2},$$

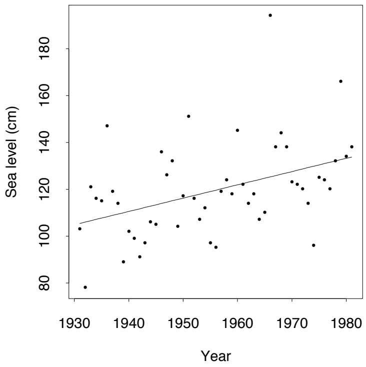
that $\partial\eta/\partial\beta^{\mathrm{T}}=X$ is the $n\times 2$ matrix whose $j\mathrm{th}$ row is $(1,x_j-\overline{x})$, and $W=\tau^{-2}I_n$. Hence (5) becomes $\widehat{\beta}\doteq(X^{\mathrm{T}}X)^{-1}(X\beta+\tau^2u)$, where the $j\mathrm{th}$ element of u is $\tau^{-1}[1-\exp\{-(y_j-\eta_j)/\tau\}]$.

□ Here it is simplest to fix τ, to obtain $\widehat{\beta}$ by iterating (5) for each fixed value of τ, and then to repeat this over a range of values of τ, giving the profile log likelihood $\ell_{\mathrm{p}}(\tau)$ and hence confidence intervals for τ. Confidence intervals for β_0 and β_1 are obtained from the information matrix.

□ With starting value chosen to be the least squares estimates of β, and with $\tau=5$, 19 iterations of (5) were required to give estimates and a maximized log likelihood whose relative change was less than 10^{-6} between successive iterations. We then took $\tau=5.5,\ldots,40$, using $\widehat{\beta}$ from the preceding iteration as starting-value for the next; in most cases just three iterations were needed. The left panel of Figure 4 shows a close-up of $\ell_{\mathrm{p}}(\tau)$; its maximum is at $\widehat{\tau}=14.5$, and the 95% confidence interval for τ is $(11.9,18.1)$. The maximum likelihood estimates of β_0 and β_1 are 111.4 and 0.563, with standard errors 2.14 and 0.137; these compare with standard errors 2.61 and 0.177 for the least squares estimates. There is some gain in precision in using the more appropriate model.

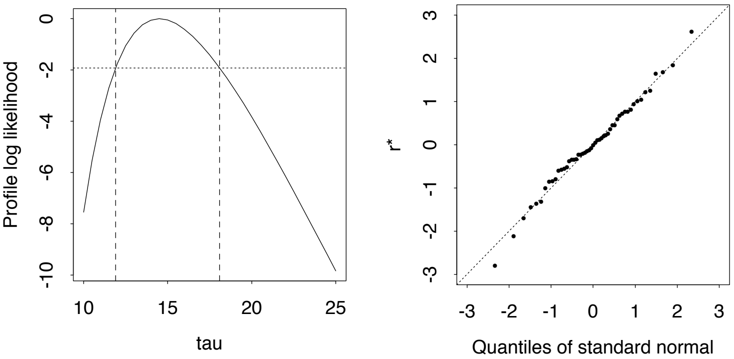
Venice data

Example 6 (Venice sea level data) *The figure below shows annual maximum sea levels in Venice, from 1931–1981. The very large value in 1966 is not an outlier. The fit of a Gumbel model to the data using IWLS gives MLEs (SEs)* $\widehat{\beta}_0 = 111.4\ (2.14)$ *(cm) and* $\widehat{\beta}_1 = 0.563\ (0.137)$ *(cm/year). The standard errors for LSEs are* 2.61, 0.177, *larger than for MLEs with Gumbel model — gain in precision through using appropriate model.*



Venice data

Figure 4: Gumbel analysis of Venice data. Left panel: profile log likelihood $\ell_p(\tau) = \max_\beta \ell(\beta, \tau)$, with 95% confidence interval $(11.9, 18.1)$ (cm) for τ. Right panel: normal probability plot of residuals r_j^*.



Deviance

- □ Let $\widehat{\eta}_j = \eta_j(\widehat{\beta}, x_j)$, where $\widehat{\beta}$ is MLE of β, giving maximised log likelihood $\ell(\widehat{\beta})$ and $\widehat{\eta}^{\mathrm{T}} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_n)$.
- □ Let $\tilde{\eta}_j$ be the value of η_j that maximises $\log f(y_j; \eta_j)$, and let $\tilde{\eta}^{\mathrm{T}} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_n)$. This corresponds to the **saturated model**, with

$$\#\text{parameters in } \eta = \#\text{observations in } y,$$

  which will give the largest likelihood possible.
- □ Define the **scaled deviance**:

$$D = 2\sum_{j=1}^{n} \{\log f(y_j; \tilde{\eta}_j) - \log f(y_j; \widehat{\eta}_j)\} \geq 0.$$

- □ Small D implies $\widehat{\eta} \approx \tilde{\eta}$, so model fits well.
- □ Large D implies poor fit — like $SS(\widehat{\beta})$ in linear model.

Differences of deviances

- □ Consider two models:
  - Model A: $\beta^{\mathrm{T}} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ vary freely — MLEs $\widehat{\eta}^A = \eta(\widehat{\beta}^A)$;
  - Model B: $(\beta_1, \ldots, \beta_q) \in \mathbb{R}^q$ vary freely, but $\beta_{q+1}, \ldots, \beta_p$ are fixed — hence q free parameters, MLEs $\widehat{\eta}^B = \eta(\widehat{\beta}^B)$.
- □ Model B is **nested within** model A: B can be obtained by restricting A.
- □ Likelihood ratio statistic for comparing the models is

$$2(\widehat{\ell}_A - \widehat{\ell}_B) = 2\sum_{j=1}^{n} \left\{\log f(y_j; \widehat{\eta}_j^A) - \log f(y_j; \widehat{\eta}_j^B)\right\} = D_B - D_A,$$

  and this $\dot{\sim}\ \chi^2_{p-q}$ if the models are regular.
- □ If ϕ unknown, replace it by an estimate: same distributional approximations will apply.

Example 7 (Normal linear model) *Find the difference of deviances in the normal linear model.*

Note to Example 7

□ Suppose that the y_j are normal with means η_j and known variance ϕ. Then

$$\log f(y_j; \eta_j, \phi) = -\tfrac{1}{2}\left\{\log(2\pi\phi) + (y_j - \eta_j)^2/\phi\right\}$$

is maximized with respect to η_j when $\tilde\eta_j = y_j$, giving $\log f(y_j; \tilde\eta_j, \phi) = -\frac{1}{2}\log(2\pi\phi)$. Therefore the scaled deviance for a model with fitted means $\widehat\eta_j$ is

$$D = \phi^{-1}\sum_{j=1}^{n}(y_j - \widehat\eta_j)^2,$$

which is just the residual sum of squares for the model, divided by ϕ. If $\eta_j = x_j^{\mathrm{T}}\beta$ is the correct normal linear model, the distribution of the residual sum of squares is $\phi\chi^2_{n-p}$, so values of D extreme relative to the χ^2_{n-p} distribution call the model into question.

□ The difference between deviances for nested models A and B in which β has dimensions p and $q < p$,

$$D_B - D_A = \phi^{-1}\sum_{j=1}^{n}\left\{(y_j - \widehat\eta_j^B)^2 - (y_j - \widehat\eta_j^A)^2\right\} \;\dot\sim\; \chi^2_{p-q}$$

when model B is correct. This distribution is exact for linear models.

□ If ϕ is unknown, it is replaced by an estimate. The large-sample properties of deviance differences outlined above still apply, though in small samples it may be better to replace the approximating χ^2 distribution by an F distribution with numerator degrees of freedom equal to the degrees of freedom for estimation of ϕ.

Model checking

Model checking I

□ Need to assess whether a given model fits adequately, or needs to be modified.

□ Two basic approaches:

- overall tests by **model expansion**, e.g., by adding a term in the model and testing for significance;
- **regression diagnostics** for detecting a few possibly dodgy observations.

□ Most widely used diagnostics in the linear model $y = X_{n\times p}\beta + \varepsilon$ are **residuals** $e_j = y_j - \widehat y_j$ and (much better) **standardized residuals**

$$r_j = \frac{y_j - \widehat y_j}{s(1 - h_{jj})^{1/2}}, \quad j = 1, \ldots, n,$$

where the **leverage** h_{jj} is the jth diagonal element of the hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, and the **Cook statistic**

$$C_j = \frac{1}{ps^2}(\widehat y - \widehat y_{-j})^{\mathrm{T}}(\widehat y - \widehat y_{-j}) = \frac{r_j^2 h_{jj}}{p(1 - h_{jj})},$$

which measures the effect of deleting the jth case (x_j, y_j) on the fitted model.

Diagnostics in general case

□ Linear model ideas work as approximations (2nd order Taylor series, painful expansions).

□ **Leverage** h_{jj} defined as jth diagonal element of

$$H = W^{1/2}X(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W^{1/2},$$

depends in general on $\widehat{\beta}$, unlike in linear model.

□ **Cook statistic** is change in deviance

$$C_j = 2p^{-1}\left\{\ell(\widehat{\beta}) - \ell(\widehat{\beta}_{-j})\right\} \doteq \frac{h_{jj}}{p(1-h_{jj})}r_{Pj}^2,$$

where $\widehat{\beta}_{-j}$ is MLE when jth case (x_j, y_j) is dropped, and r_{Pj} is **standardized Pearson residual** (see below).

□ There are several types of residual (see next page).

Residuals in general case

□ **Deviance residual**:

$$d_j = \mathrm{sign}(\tilde{\eta}_j - \widehat{\eta}_j)[2\{\ell_j(\tilde{\eta}_j;\phi) - \ell_j(\widehat{\eta}_j;\phi)\}]^{1/2},$$

for which $\sum d_j^2 = D$ is deviance.

□ **Pearson residual**: $u_j(\widehat{\beta})/\sqrt{w_j(\widehat{\beta})}$.

□ Standardized versions

$$r_{Dj} = \frac{d_j}{(1-h_{jj})^{1/2}}, \quad r_{Pj} = \frac{u_j(\widehat{\beta})}{\{w_j(\widehat{\beta})(1-h_{jj})\}^{1/2}},$$

and (even better)

$$r_j^* = r_{Dj} + r_{Dj}^{-1}\log(r_{Pj}/r_{Dj}) \mathrel{\dot{\sim}} N(0,1)$$

for many models.

□ These all reduce to usual standardized residual for normal linear model.

Summary

□ For regression problems with independent responses y_j dependent on parameters β through parameter $\eta_j = \eta(x_j;\beta)$, generalise least squares estimation to maximum likelihood estimation, using iterative weighted least squares algorithm: iterate to convergence

$$\widehat{\beta} = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wz, \quad z = X\beta + W^{-1}u,$$

where

$$X_{n\times p} \equiv X(\beta) = \frac{\partial\eta}{\partial\beta^{\mathrm{T}}}, \quad u_{n\times 1} \equiv u(\eta) = \frac{\partial\ell}{\partial\eta}, \quad W_{n\times n} \equiv W(\eta) = -\mathrm{E}\left\{\frac{\partial^2\ell}{\partial\eta\partial\eta^{\mathrm{T}}}\right\},$$

with ℓ the log likelihood for the data.

□ Standard likelihood theory is used for confidence intervals and model comparison.

□ Linear model diagnostics (residuals, leverage, Cook statistics, ...) generalise to this setting.

□ Next: generalized linear models (GLMs), wide class of models with exponential family-like response distributions.

Motivation

□ Need to generalise linear model beyond normal responses, e.g. to data with $y \in \{0, 1, \ldots, m\}$, or $y \in \{0, 1, \ldots\}$, or $y > 0$.

□ Consider **exponential family** response distributions (gamma, exponential, binomial, Poisson, $\ldots$), since they have an elegant unifying theory, and encompass many possibilities (in addition to the normal distribution)

□ Basic idea is to build models such that

$$\mathrm{E}(y) = \mu, \quad g(\mu) = \eta = x^{\mathrm{T}}\beta,$$

where g is suitable function, and distribution of y lies in exponential family (well, almost).

□ **Warnings**:

- **Don't** confuse Generalized Linear Model (GLM) with General Linear Model (GLM, in older books, the latter is $y = X\beta + \varepsilon$, with $\mathrm{cov}(\varepsilon) = \sigma^2 V$ not diagonal);
- **Don't** write $y = \mu + \varepsilon$, since in a GLM the distribution of ε usually depends on μ.

Construction of exponential families

□ Take a baseline density/mass function $f_0(y)$ with support $\mathcal{Y} = \{y : f_0(y) > 0\}$ and a (possibly vector) statistic $s(y)$ with $\mathrm{var}_0\{s(Y)\} > 0$.

□ Define the **natural parameter space**

$$\mathcal{N} = \left\{\theta : \kappa(\theta) = \log \int e^{s(y)^{\mathrm{T}}\theta} f_0(y)\,\mathrm{d}y < \infty\right\} \subset \mathbb{R}^k,$$

where $k = \dim s(y)$. Obviously $0 \in \mathcal{N}$, and Hölder's inequality yields that $\mathcal{N}$ is convex and $\kappa(\theta)$ strictly convex on $\mathcal{N}$.

□ This implies that the **exponentially tilted** version of f_0, i.e.,

$$f(y;\theta) = f_0(y)\exp\left\{s(y)^{\mathrm{T}}\theta - \kappa(\theta)\right\}, \quad y \in \mathcal{Y}, \theta \in \mathcal{N}, \tag{9}$$

is a well-defined density/mass function.

□ Expression (9) is called a **natural exponential family** if $s(y) = y$, and called **regular** if $\mathcal{N}$ is an open set.

□ The **cumulant-generating function** of $s(Y)$ is $K(t) = \kappa(\theta + t) - \kappa(\theta)$, so

$$\mathrm{E}_\theta\{s(Y)\} = \partial\kappa(\theta)/\partial\theta, \quad \mathrm{var}_\theta\{s(Y)\} = \partial^2\kappa(\theta)/\partial\theta\partial\theta^{\mathrm{T}};$$

since $\mathrm{var}_\theta\{s(Y)\}$ is positive definite for all θ, $\mu(\theta) = \mathrm{E}_\theta\{s(Y)\}$ is strictly increasing in θ.

Exponential families

□ Since the mean function $\mu(\theta) = \mathrm{E}_\theta\{s(Y)\}$ is strictly increasing in θ, we can reparametrise (9) in terms of μ, setting $\theta = \theta(\mu)$, and this also yields

$$\mathrm{var}_\theta\{s(Y)\} = \frac{\partial^2 \kappa(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} = \frac{\partial\mu(\theta)}{\partial\theta^{\mathrm{T}}} = V(\mu),$$

say, where $V(\mu)$ is called the **variance function** of the family. It can be shown that $V(\mu)$ and the domain $\mathcal{M}$ of μ characterise the family.

□ The usual definition eliminates the baseline density f_0 and puts

$$f(y;\omega) = \exp\left\{s(y)^{\mathrm{T}}\theta(\omega) - b(\omega) + c(y)\right\}, \quad y \in \mathcal{Y}, \omega \in \Omega,$$

which allows more flexibility in the parametrisation, but is equivalent to the constructive approach.

Example 8 (Uniform density) *Construct exponential families for which the baseline density is uniform on* $(0,1)$ *and* $s(y)$ *respectively equals*

$$y, \quad (\log y, \log(1-y)), \quad (\sin(2\pi y), \cos(2\pi y)).$$

Note to Example 8

□ Let $f_0(y) = 1$ for $y \in \mathcal{Y} = (0, 1)$. Now

$$\kappa(\theta) = \log \int e^{y\theta} f_0(y)\,\mathrm{d}y = \log \int_0^1 e^{y\theta}\,\mathrm{d}y = \log \left\{(e^\theta - 1)/\theta\right\} < \infty$$

for all $\theta \in \mathcal{N} = (-\infty, \infty)$, and the natural exponential family is

$$f(y; \theta) = \begin{cases} \theta e^{\theta y}/(e^\theta - 1), & 0 < y < 1, \\ 0, & \text{otherwise}. \end{cases} \tag{10}$$

For this or any natural exponential family with bounded $\mathcal{Y}$, $\mathcal{N} = (-\infty, \infty)$ and the family is regular.

□ If $f_0(y)$ is uniform on $(0, 1)$ and $s(y)$ equals $(\log y, \log(1 - y))^{\mathrm{T}}$, then

$$\kappa(\theta) = \log \int_0^1 \exp\left\{\theta_1 \log y + \theta_2 \log(1 - y)\right\}\,\mathrm{d}y = \log B(1 + \theta_1, 1 + \theta_2),$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ is the beta function. The resulting model is usually written in terms of $a = \theta_1 + 1$ and $b = \theta_2 + 1$, giving the **beta density**

$$f(y; a, b) = \frac{y^{a-1}(1 - y)^{b-1}}{B(a, b)}, \quad 0 < y < 1, \quad a, b > 0. \tag{11}$$

In this parametrization the natural parameter space is $\mathcal{N} = (0, \infty) \times (0, \infty)$.

□ For the third example, we have $f_0(y) = 1$ for $0 \le y < 1$, and take

$$s(y) = (\cos(2\pi y), \sin(2\pi y))^{\mathrm{T}}, \quad \theta(\omega) = (\tau \cos(2\pi\gamma), \tau \sin(2\pi\gamma))^{\mathrm{T}},$$

where $\mathcal{N} = \mathbb{R}^2$ but $\omega = (\tau, \gamma)$ lies in $\Omega = [0, \infty) \times [0, 1)$. This choice of $s(y)$ ensures that $f(y) = f(y \pm k)$ for all integer k. Now $s(y)^{\mathrm{T}}\theta(\omega) = \tau \cos\{2\pi(y - \gamma)\}$ and

$$\int e^{s(y)^{\mathrm{T}}\theta(\omega)} f_0(y)\,\mathrm{d}y = \int_0^1 e^{\tau \cos\{2\pi(y-\gamma)\}}\,\mathrm{d}y = \frac{1}{2\pi} \int_0^{2\pi} e^{\tau \cos y}\,\mathrm{d}y,$$

which is a modified Bessel function of the first kind. If we replace $2\pi y$ by y, we obtain the von Mises density

$$f(y; \tau, \gamma) = \{2\pi I_0(\tau)\}^{-1} e^{\tau \cos(y - \gamma)}, \quad 0 \le y < 2\pi, \ \tau > 0, 0 \le \gamma < 2\pi.$$

The **mean direction** γ gives the direction in which observations are concentrated, and the **precision** τ gives the strength of that concentration. Notice that $\tau = 0$ gives the uniform distribution on the circle, whatever the value of γ. Here interest focuses on Y rather than on $s(Y)$, which is introduced purely in order to generate a natural class of densities for y.

Generalized linear model (GLM)

□ Normal linear model has three key aspects:

- structure for covariates: **linear predictor**, $\eta = x^{\mathrm{T}}\beta$;
- response distribution: $y \sim N(\mu, \sigma^2)$;
- linear relation $\eta = \mu$ between $\mu = \mathrm{E}(y)$ and η.

□ GLM extends last two to

- Y has density/mass function

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\}, \quad y \in \mathcal{Y}, \theta \in \Omega_\theta, \phi > 0, \tag{12}$$

  where
  - ▷ $\mathcal{Y}$ is the support of Y,
  - ▷ Ω_θ is the parameter space of valid values for $\theta \equiv \theta(\eta)$, and
  - ▷ the **dispersion parameter** ϕ is often known;
- $\eta = g(\mu)$, where g is monotone **link function**
  - ▷ the **canonical link** function giving $\eta = \theta = b'^{-1}(\mu)$ has nice statistical properties;
  - ▷ but a range of link functions are possible for each distribution of Y.

Examples

Example 9 (GLM density) *Show that the moment-generating function of $f(y;\theta,\phi)$ is $M_Y(t) = \exp\{b(\theta + t\phi) - b(\theta)\}$, and deduce that we can write*

$$\mathrm{E}(Y) = b'(\theta) = \mu, \quad \mathrm{var}(Y) = \phi b''(\theta) = \phi b''\{b'^{-1}(\mu)\} = \phi V(\mu);$$

the function $\mu \mapsto V(\mu)$ is known as the **variance function**.

Example 10 (Poisson distribution) *Write the Poisson mass function as a GLM density, and find its canonical link function.*

Example 11 (Normal distribution) *Write the normal density function as a GLM density, and find its canonical link function.*

Example 12 (Binomial distribution) *Write the binomial mass function as a GLM density, and find its canonical link function.*

Note to Example 9

Suppose that Y has a continuous density; if not the argument below is the same, except that integral signs are replaced by summations.
Let $\Omega_\theta = \{\theta : b(\theta) < \infty\}$.
We have

$$M_Y(t) = \mathrm{E}\{\exp(tY)\} = \int e^{ty} \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\} \mathrm{d}y = \int \exp\left\{\frac{y(\theta + t\phi) - b(\theta)}{\phi} + c(y;\phi)\right\} \mathrm{d}y.$$

If $\theta + t\phi \in \Omega_\theta$, then

$$\int \exp\left\{\frac{y(\theta + t\phi) - b(\theta + t\phi)}{\phi} + c(y;\phi)\right\} \mathrm{d}y = 1,$$

so

$$M_Y(t) = \mathrm{E}\{\exp(tY)\} = \exp\left[\{b(\theta + t\phi) - b(\theta)\}/\phi\right].$$

Hence the cumulant-generating function of Y is

$$K_Y(t) = \log M_Y(t) = \{b(\theta + t\phi) - b(\theta)\}/\phi,$$

and differentiating twice with respect to t and setting $t = 0$ yields

$$K'_Y(t)\big|_{t=0} = b'(\theta), \quad K''_Y(t)\big|_{t=0} = \phi b''(\theta).$$

Since $b(\theta)$ is strictly convex on Ω_θ, $b'(\theta)$ is a monotonic increasing function of θ, so $b'^{-1}(\cdot)$ exists and is itself monotonic, so $V(\mu) = b''\{b'^{-1}(\mu)\}$ is well-defined.

Note to Example 10

The Poisson density may be written as

$$f(y;\mu) = \exp\left(y \log \mu - \mu - \log y!\right), \quad y = 0, 1, \ldots, \quad \mu > 0,$$

which has GLM form (12) with $\theta = \log \mu$, $b(\theta) = e^\theta$, $\phi = 1$, and $c(y;\phi) = -\log y!$. The mean of y is $\mu = b'(\theta) = e^\theta = \mu$, and its variance is $b''(\theta) = e^\theta = \mu$, so the variance function is linear: $V(\mu) = \mu$.

Note to Example 11

The normal density with mean μ and variance σ^2 may be written

$$f(y;\mu,\sigma^2) = \exp\left\{-\frac{(y^2 - 2y\mu + \mu^2)}{2\sigma^2} - \tfrac{1}{2}\log(2\pi\sigma^2)\right\},$$

so

$$\theta = \mu, \quad \phi = \sigma^2, \quad b(\theta) = \tfrac{1}{2}\theta^2, \quad c(y;\phi) = -\tfrac{1}{2\phi}y^2 - \tfrac{1}{2}\log(2\pi\phi).$$

As the first and second derivatives of $b(\theta)$ are θ and 1, we have $V(\mu) = 1$; the variance function is constant.

Note to Example 12

We write the binomial density

$$f(r;\pi) = \binom{m}{r}\pi^r(1-\pi)^{m-r}, \quad 0<\pi<1, \quad r=0,\ldots,m,$$

in the form

$$\exp\left[m\left\{\frac{r}{m}\log\left(\frac{\pi}{1-\pi}\right)+\log(1-\pi)\right\}+\log\binom{m}{r}\right],$$

so

$$y=\frac{r}{m},\ \phi=\frac{1}{m},\ \theta=\log\left(\frac{\pi}{1-\pi}\right),\ b(\theta)=\log(1+e^\theta),\ c(y;\phi)=\log\binom{m}{r}.$$

The mean and variance of y are

$$\mu=b'(\theta)=\frac{e^\theta}{1+e^\theta}, \quad \phi b''(\theta)=\frac{e^\theta}{m(1+e^\theta)^2};$$

the variance function is $V(\mu)=\mu(1-\mu)$.

Estimation of β

Example 13 (IWLS algorithm) *Find the components of the IWLS algorithm for a GLM.*

□ If canonical link is used then $\theta_j = x_j^{\mathrm{T}}\beta$, so if ϕ is known, then

$$\begin{aligned}\ell(\beta) &= \sum_{j=1}^n\left\{\frac{y_jx_j^{\mathrm{T}}\beta-b(x_j^{\mathrm{T}}\beta)}{\phi}+c(y_j;\phi)\right\}\\ &= \{y^{\mathrm{T}}X\beta-K(\beta)\}/\phi+C(y;\phi),\end{aligned}$$

say, which in terms of β is a linear exponential family with canonical parameter $\beta_{p\times 1}$ and canonical statistic $(X^{\mathrm{T}}y)_{p\times 1}$.

□ If X is full rank, then $\ell(\beta)$ is strictly concave and has a unique maximum in terms of β.

□ Problem: the maximum may be at infinity in certain (rare) cases—this can arise with binomial responses: beware of $\widehat{\theta}_r \approx \pm 36$.

Note to Example 13

□ To compute the quantities needed for the IWLS step $\widehat{\beta} = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W(X\beta + W^{-1}u)$, we need

$$X_{n\times p} = \frac{\partial \eta}{\partial \beta^{\mathrm{T}}}, \quad W_{n\times n} = \mathrm{diag}\{\mathrm{E}(-\partial^2 \ell_j/\partial \eta_j^2)\}, \quad u_{n\times 1} = \{\partial \ell_j/\partial \eta_j\},$$

where (with ϕ_j instead of ϕ for generality, see the next slide),

$$\ell_j(\beta) = \left\{\frac{y_j\theta_j - b(\theta_j)}{\phi_j} + c(y_j;\phi_j)\right\}, \quad b'(\theta_j) = \mu_j, \quad \eta_j = g(\mu_j) = x_j^{\mathrm{T}}\beta.$$

□ First note that $\partial\eta_j/\partial\beta_r = x_{jr}$, so $X = \partial\eta/\partial\beta^{\mathrm{T}}$ is just a matrix of constants.

□ We need the first and second derivatives of ℓ_j with respect to η_j, so we write

$$\frac{\partial \ell_j}{\partial \eta_j} = \frac{\partial \mu_j}{\partial \eta_j}\frac{\partial \theta_j}{\partial \mu_j}\frac{\partial \ell_j}{\partial \theta_j},$$

with

$$\frac{\partial \eta_j}{\partial \mu_j} = g'(\mu_j), \quad \frac{\partial \mu_j}{\partial \theta_j} = b''(\theta_j) = V(\mu_j), \quad \frac{\partial \ell_j}{\partial \theta_j} = \frac{y_j - b'(\theta_j)}{\phi_j},$$

which yields

$$u_j = \frac{\partial \ell_j}{\partial \eta_j} = \frac{y_j - b(\theta_j)}{g'(\mu_j)\phi_j V(\mu_j)} = \frac{y_j - \mu_j}{g'(\mu_j)\phi_j V(\mu_j)} = \frac{A(\theta_j)}{B(\theta_j)},$$

say, where $\mathrm{E}(A) = 0$. For the second derivative, we note that

$$\frac{\partial^2 \ell_j}{\partial \eta_j^2} = \frac{\partial}{\partial \eta_j}\frac{\partial \ell_j}{\partial \eta_j} = \left(\frac{\partial \mu_j}{\partial \eta_j}\frac{\partial \theta_j}{\partial \mu_j}\frac{\partial}{\partial \theta_j}\right)\frac{\partial \ell_j}{\partial \eta_j} = \frac{\partial \mu_j}{\partial \eta_j}\frac{\partial \theta_j}{\partial \mu_j}\left\{\frac{A'(\theta_j)}{B(\theta_j)} - \frac{A(\theta_j)B'(\theta_j)}{B(\theta_j)^2}\right\},$$

and on noting that $B(\theta_j)$ is non-random and $A'(\theta_j) = -b''(\theta_j) = -V(\mu_j)$, we obtain

$$w_j = \mathrm{E}\left(-\frac{\partial^2 \ell_j}{\partial \eta_j^2}\right) = \frac{1}{g'(\mu_j)}\frac{1}{V(\mu_j)}\frac{V(\mu_j)}{g'(\mu_j)\phi_j V(\mu_j)} = \frac{1}{g'(\mu_j)^2\phi_j V(\mu_j)}.$$

Note to Example 13, part II

□ From above we see that the components of the score statistic $u(\beta)$ and the weight matrix $W(\beta)$ may be expressed in terms of components μ_j of the mean vector μ as

$$\begin{aligned} u_j &= \frac{\partial\theta_j}{\partial\eta_j}\frac{\partial\ell_j(\theta_j)}{\partial\theta_j} = \frac{y_j-\mu_j}{g'(\mu_j)\phi_j V(\mu_j)}, \\ w_j &= \left(\frac{\partial\theta_j}{\partial\eta_j}\right)^2 \frac{\partial^2\ell_j(\theta_j)}{\partial\theta_j^2} = \frac{1}{g'(\mu_j)^2\phi_j V(\mu_j)}, \end{aligned} \tag{13}$$

where $g'(\mu_j) = dg(\mu_j)/d\mu_j$. Thus $\widehat{\beta}$ is obtained by iterative weighted least squares regression of response

$$z = X\beta + g'(\mu)(y-\mu) = \eta + g'(\mu)(y-\mu)$$

on the columns of X using weights (13).

□ By using y as an initial value for μ and $g(y)$ as an initial value for $\eta = X\beta$, we avoid needing an initial value for β.

□ It may be necessary to modify y slightly for this initial step. For example if we use the log link for Poisson data, and some y_j equal zero, then we may need to replace them with some small positive value to avoid taking $\log 0$ for some components of the initial $\eta = \log y$.

Estimation of ϕ

□ When ϕ unknown, it is often replaced by $\phi_j = \phi a_j$, with known a_j and a_j^{-1} treated as a weight. Then we replace the scaled deviance by the **deviance** ϕD.

□ If the model is correct and ϕ is known, then **Pearson's statistic**

$$P = \frac{1}{\phi}\sum_{j=1}^{n}\frac{(y_j-\widehat{\mu}_j)^2}{a_j V(\widehat{\mu}_j)} \dot{\sim} \chi^2_{n-p},$$

analogously to the sum of squares in a linear model, with $\mathrm{E}(P) \doteq n-p$.

□ The MLE of ϕ can be badly behaved, so usually we prefer the method of moments estimator

$$\widehat{\phi} = \frac{1}{n-p}\sum_{j=1}^{n}(y_j-\widehat{\mu}_j)^2/\{a_j V(\widehat{\mu}_j)\},$$

which is obtained by solving the equation $P = n-p$, based on noting that $\mathrm{E}(\chi^2_{n-p}) = n-p$.

□ If the data are sparse (e.g., many small binomial or Poisson counts), then standard asymptotic results are suspect.

Jacamar data

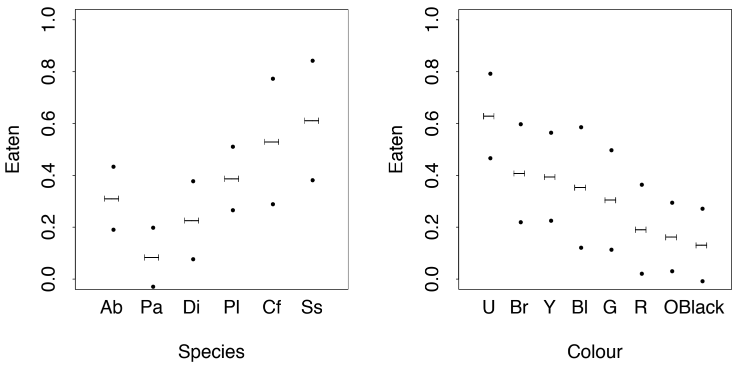
Table 6: Response (N=not sampled, S = sampled and rejected, E = eaten) of a rufous-tailed jacamar to individuals of seven species of palatable butterflies with artifically coloured wing undersides. Data from Peng Chai, University of Texas.

	Aphrissa	*Phoebis*	*Dryas*	*Pierella*	*Consul*	*Siproeta*
	boisduvalli	*argante*	*iulia*	*luna*	*fabius*	*stelenes*†
	N/S/E	N/S/E	N/S/E	N/S/E	N/S/E	N/S/E
Unpainted	0/0/14	6/1/0	1/0/2	4/1/5	0/0/0	0/0/1
Brown	7/1/2	2/1/0	1/0/1	2/2/4	0/0/3	0/0/1
Yellow	7/2/1	4/0/2	5/0/1	2/0/5	0/0/1	0/0/3
Blue	6/0/0	0/0/0	0/0/1	4/0/3	0/0/1	0/1/1
Green	3/0/1	1/1/0	5/0/0	6/0/2	0/0/1	0/0/3
Red	4/0/0	0/0/0	6/0/0	4/0/2	0/0/1	3/0/1
Orange	4/2/0	6/0/0	4/1/1	7/0/1	0/0/2	1/1/1
Black	4/0/0	0/0/0	1/0/1	4/2/2	7/1/0	0/1/0

† includes *Philaethria dido* also.

Jacamar data

Figure 5: Proportion of butterflies eaten ($\pm 2SE$) for different species and wing colour.



Jacamar data

□ How does a bird respond to the species s and wing colour c of its prey?

□ Response has 3 (ordered) categories: not attacked (N), attacked but then rejected (S), attacked and eaten (E)

□ The data form an 8×6 layout, with a 3-category response in each cell, total m_{cs}

□ Assume that the number in category E (response) is binomial:

$$R_{cs} \sim B(m_{cs}, \pi_{cs}), \quad c = 1, \ldots, 8, s = 1, \ldots, 6,$$

where c is colour and s is species, with probability that bird attacks and eats butterfly is

$$\pi_{cs} = \frac{\exp(\alpha_c + \gamma_s)}{1 + \exp(\alpha_c + \gamma_s)}, \quad c = 1, \ldots, 8, s = 1, \ldots, 6,$$

so

– large α_c corresponds to colours that the jacamar likes to eat,

– large γ_s corresponds to species that it likes.

□ This is a GLM with response $y_{cs} = r_{cs}/m_{cs}$, $\mathrm{E}(y_{cs}) = \pi_{cs}$, and canonical (logit) link function

$$\eta = \log\{\pi/(1-\pi)\}, \quad \eta_{cs} = \alpha_c + \gamma_s.$$

Jacamar data: Analysis of deviance

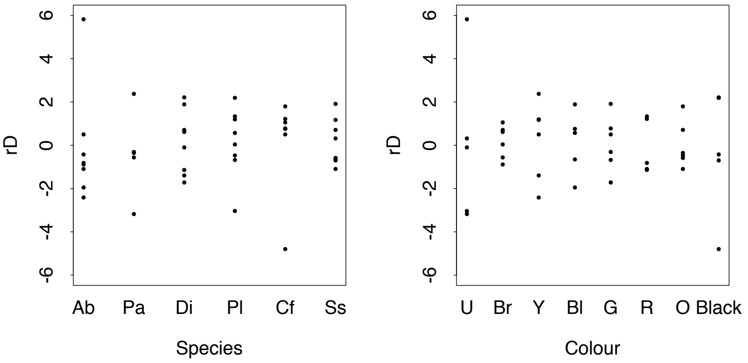
Table 7: Deviances and analysis of deviance for models fitted to jacamar data. The lower part shows results for the reduced data, without two outliers.

	Full data		Without outliers	
Terms	df	Deviance	df	Deviance
1	43	134.24	35	73.68
1+Species	38	114.59	31	46.04
1+Colour	36	108.46	28	63.20
1+Species+Colour	31	67.28	24	28.02

Terms	df	Deviance reduction	Terms	df	Deviance reduction
Species (unadj. for Colour)	5	19.64	Species (adj. for Colour)	5	41.18
Colour (adj. for Species)	7	47.31	Colour (unadj. for Species)	7	25.78
Species (unadj. for Colour)	4	27.63	Species (adj. for Colour)	4	35.18
Colour (adj. for Species)	7	18.03	Colour (unadj. for Species)	7	10.48

Jacamar data: Residuals

Figure 6: Standardized deviance residuals r_D for binomial two-way layout fitted to jacamar data.



Jacamar data: Parameter estimates

Table 8: Estimated parameters and standard errors for the jacamar data, without 2 outliers.

Aphrissa boisduvalli	*Phoebis argante*	*Dryas iulia*	*Pierella luna*	*Consul fabius*	*Siproeta stelenes*
–1.99 (0.79)	–2.22 (0.85)	–0.56 (0.67)	0.16 (0.54)	—	1.50 (0.78)

Brown	Yellow	Blue	Green	Red	Orange	Black
0.16 (0.73)	0.33 (0.68)	–0.53 (0.81)	–0.83 (0.75)	–1.93 (0.88)	–1.94 (0.85)	–1.26 (0.86)

□ Interpretation

□ Residual deviance: 28.02, with 24 df

□ Pearson statistic: 25.58, with 24 df

□ Standardized residuals in range -2.03 to 1.96: OK.

Chimpanzee data

Table 9: Times in minutes taken by four chimpanzees to learn ten words.

Chimpanzee	Word									
	1	2	3	4	5	6	7	8	9	10
1	178	60	177	36	225	345	40	2	287	14
2	78	14	80	15	10	115	10	12	129	80
3	99	18	20	25	15	54	25	10	476	55
4	297	20	195	18	24	420	40	15	372	190

□ Another two-way layout.

□ Times vary from 2 to 476 minutes — need transformation (e.g., logarithm) if use linear model.

Chimpanzee data

□ How does learning time depend on word w and chimp c?

□ Response is continuous and positive, so we try fitting the gamma distribution with mean and shape parameters μ and ν, i.e.,

$$f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)} y^{\nu-1} \left(\frac{\nu}{\mu}\right)^{\nu} \exp(-\nu y/\mu), \quad y > 0, \quad \nu, \mu > 0,$$

so dispersion parameter is $\phi = 1/\nu$ ($\phi = 1$ for exponential).

□ Possible link functions:

$$\eta = \log\mu, \text{ (log, most often used)}, \quad \eta = 1/\mu, \text{ (inverse, canonical)}$$

□ Linear model structure:

$$\eta_{cw} = \alpha_c + \gamma_w, \quad c = 1, \dots, 4, w = 1, \dots, 10,$$

but the interpretation of the α_c and γ_w will depend on the link function.

□ With the log link, the deviances for models 1, 1+Chimp, 1+Word, and 1+Chimp+Word are 60.38, 53.43, 21.19, and 14.97. How many df are there for each model?

Chimpanzee data: Analysis of deviance

Table 10: Analysis of deviance for models fitted to chimpanzee data.

Term	df	Deviance reduction	Term	df	Deviance reduction
`Chimp` (unadj. for `Word`)	3	6.95	`Chimp` (adj. for `Word`)	3	6.22
`Word` (adj. for `Chimp`)	9	38.46	`Word` (unadj. for `Chimp`)	9	39.19

□ Method of moments estimate is $\widehat{\phi} = 0.432$, so $\widehat{\nu} = 1/\widehat{\phi} = 2.31$.

□ Use F tests to assess effects of `Word` and `Chimp`, for example obtaining

$$\frac{6.22/3}{0.423} = 4.78 \dot{\sim} F_{3,27}$$

if there is no difference between the chimps. What is the corresponding statistic for testing differences between words?

□ Residuals suggest that this model, or one with the inverse link, are both adequate, and both are better than fitting a normal linear model to the log times.

Summary

□ Generalized linear models extend the classical linear model in two ways:

- the response distribution is (almost) exponential family, so includes binomial, Poisson, gamma and other distributions in addition to the normal;
- the relation between the linear predictor $\eta = x^{\mathrm{T}}\beta$ and the mean μ is determined by a wide range of possible link functions.

□ Canonical link functions give particularly simple models and are widely used.

□ Estimates of β are obtained by IWLS, which has a simple form, with no need for initial values.

□ A simple estimate of the dispersion parameter ϕ is available using the method of moments.

□ Models are compared using the analysis of deviance, which generalises the analysis of variance in the classical linear model.

□ Standard likelihood theory results are used for inference (standard errors, confidence intervals, etc.)

□ Standard diagnostics (residuals, . . .) extend in a natural way to this setting.

Binary response

□ Response Y has Bernoulli distribution with

$$\mathrm{P}(Y=1)=\pi, \quad \mathrm{P}(Y=0)=1-\pi, \quad 0<\pi<1.$$

and $\mathrm{E}(Y)=\mu=\pi$, $\mathrm{var}(Y)=\pi(1-\pi)$.

□ Linear link function $\pi=\eta=x^{\mathrm{T}}\beta$ can give $\pi\notin[0,1]$, so not usually a good idea.

□ Y can be interpreted in terms of a hidden variable/tolerance distribution: let $Z=x^{\mathrm{T}}\gamma+\sigma\varepsilon$, where $\varepsilon\sim F$. Set $Y=I(Z>0)$, and note that

$$\pi=\mathrm{P}(Y=1)=\mathrm{P}(x^{\mathrm{T}}\gamma+\sigma\varepsilon>0)=\mathrm{P}(\varepsilon>-x^{\mathrm{T}}\gamma/\sigma)=1-F(-x^{\mathrm{T}}\beta),$$

say. Note that $\beta=\gamma/\sigma$ is estimable, but γ and σ are not.

□ The corresponding link function is given by

$$\eta=x^{\mathrm{T}}\beta=-F^{-1}(1-\pi)=g(\pi),$$

so different choices of F yield different possible link functions.

Link functions

Tolerance distributions and corresponding link functions for binary data.

Distribution F		Link function	
Logistic	$e^u/(1+e^u)$	Logit	$\eta=\log\{\pi/(1-\pi)\}$
Normal	$\Phi(u)$	Probit	$\eta=\Phi^{-1}(\pi)$
Log Weibull	$1-\exp(-\exp(u)\}$	Log-log	$\eta=-\log\{-\log(\pi)\}$
Gumbel	$\exp\{-\exp(-u)\}$	Complementary log-log	$\eta=\log\{-\log(1-\pi)\}$

□ The logit and probit links are symmetric.

□ Logit (canonical link) is usual choice, good for medical studies (later), with nice interpretation, but the probit is very similar to it and may be preferred in some cases, for its relation to the normal distribution.

□ The log-log and complementary log-log links are asymmetric.

Logistic regression

□ Commonest choice of link function for proportion data is the **logit**, which gives

$$\mathrm{P}(Y=1)=\pi=\frac{\exp(x^{\mathrm{T}}\beta)}{1+\exp(x^{\mathrm{T}}\beta)},\quad \mathrm{P}(Y=0)=1-\pi=\frac{1}{1+\exp(x^{\mathrm{T}}\beta)},$$

leading to a linear model for the **log odds** of success,

$$\log\left\{\frac{\mathrm{P}(Y=1)}{\mathrm{P}(Y=0)}\right\}=\log\left(\frac{\pi}{1-\pi}\right)=x^{\mathrm{T}}\beta,\quad \beta\in\mathbb{R}^p.$$

□ The likelihood for β based on independent responses $y_1,\ldots,y_n$ with covariate vectors $x_1,\ldots,x_n$ and corresponding probabilities $\pi_1,\ldots,\pi_n$ is

$$L(\beta)=\prod_{j=1}^n \pi_j^{y_j}(1-\pi_j)^{1-y_j}=\cdots=\frac{\exp\left(\sum_{j=1}^n y_j x_j^{\mathrm{T}}\beta\right)}{\prod_{j=1}^n\left\{1+\exp\left(x_j^{\mathrm{T}}\beta\right)\right\}},$$

which is a regular exponential family with $s(y)=X^{\mathrm{T}}y$ and log likelihood

$$\ell(\beta)=(X^{\mathrm{T}}y)^{\mathrm{T}}\beta-\sum_{j=1}^n\log\left\{1+\exp\left(x_j^{\mathrm{T}}\beta\right)\right\},\quad \beta\in\mathbb{R}^p,$$

known as the **logistic regression model**.

Nodal involvement data

Data on nodal involvement: 53 patients with prostate cancer have nodal involvement (r), with five binary covariates age, stage, etc.

m	r	age	stage	grade	xray	acid
6	5	0	1	1	1	1
6	1	0	0	0	0	1
4	0	1	1	1	0	0
4	2	1	1	0	0	1
4	0	0	0	0	0	0
3	2	0	1	1	0	1
3	1	1	1	0	0	0
3	0	1	0	0	0	1
3	0	1	0	0	0	0
2	0	1	0	0	1	0
2	1	0	1	0	0	1
2	1	0	0	1	0	0
1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	
1	1	0	0	1	0	1
1	0	0	0	0	1	1
1	0	0	0	0	1	0

Deviances for nodal involvement models

Scaled deviances D for 32 logistic regression models for nodal involvement data. + denotes a term included in the model.

age	st	gr	xr	ac	df	D	age	st	gr	xr	ac	df	D
					52	40.71	+	+	+			49	29.76
+					51	39.32	+	+		+		49	23.67
	+				51	33.01	+	+			+	49	25.54
		+			51	35.13	+		+	+		49	27.50
			+		51	31.39	+		+		+	49	26.70
				+	51	33.17	+			+	+	49	24.92
+	+				50	30.90		+	+	+		49	23.98
+		+			50	34.54		+	+		+	49	23.62
+			+		50	30.48		+		+	+	49	19.64
+				+	50	32.67			+	+	+	49	21.28
	+	+			50	31.00	+	+	+	+		48	23.12
	+		+		50	24.92	+	+	+		+	48	23.38
	+			+	50	26.37	+	+		+	+	48	19.22
		+	+		50	27.91	+		+	+	+	48	21.27
		+		+	50	26.72		+	+	+	+	48	18.22
			+	+	50	25.25	+	+	+	+	+	47	18.07

Model selection

- □ We have 32 competing models, and would like to select the 'best', or a few 'near-best'.
- □ In general we have 2^p models, so automatic selection of some sort is needed.
- □ Could use likelihood ratio tests (differences of deviances) to compare competing models, but this involves many correlated tests, so may lead to spurious results.
- □ Usually minimise some measure of predictive fit, an information criterion, which accounts for the number of parameters in each model. Classical information criteria are

  $$\mathrm{AIC} \equiv D + 2p, \quad \mathrm{BIC} \equiv D + p \log n,$$

  where D is the deviance.
- □ Properties:
  - AIC tends to overfit, i.e., it has a positive probability of choosing a model that is too complex, even as $n \to \infty$;
  - BIC applies a stronger penalty as $n \to \infty$, so *if the true model is among those fitted*, it will choose it with probability one as $n \to \infty$;
  - BIC usually yields less complex models than AIC, but they may predict less well.
- □ There are many other information criteria, but these are most used in practice.

Example: Nodal involvement

□ Model with lowest AIC has stage, xray, acid:

$$x^{\mathrm{T}}\widehat{\beta} = -3.05 + 1.65 I_{\textsf{stage}} + 1.91 I_{\textsf{xray}} + 1.64 I_{\textsf{acid}},$$

where $I_{\textsf{stage}} = 1$ indicates that stage takes its higher level, etc.

□ Interpretation of this model:

– for an individual with stage, xray and acid at their lowest levels, the fitted probability of nodal involvement is $e^{-3.05}/(1+e^{-3.05}) \doteq 0.045$ (though there are no such people in the data, so this involves extrapolation);

– for someone with only $I_{\textsf{stage}} = 1$, the odds of nodal involvement are $e^{-3.05+1.65} = e^{-1.4} \doteq 0.25$, a probability of 0.2;

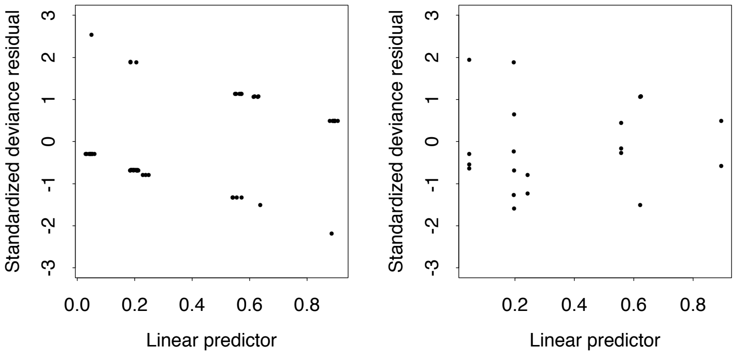
– for someone with $I_{\textsf{stage}} = I_{\textsf{xray}} = I_{\textsf{acid}} = 1$, the odds of nodal involvement are $e^{-3.05+1.65+1.91+1.64} \doteq 8.6$, a probability of 0.9;

□ Problems with interpretation of residual deviance of 19.64: how many df? — can amalgamate independent binary responses with same covariates.

□ Likewise problems with residuals . . .

Nodal involvement residuals

Figure 7: Standardized deviance residuals for nodal involvement data, for ungrouped responses (left) and grouped responses (right).



Summary

□ Proportion data are often modelled using the Bernoulli/binomial response distributions.

□ Link functions (logit, probit, . . .) have interpretations in terms of underlying continuous variables that have been dichotomized.

□ The canonical and most commonly-used link is the logit, and fitting using this yields logistic regression, in which the canonical parameter is the log odds.

□ The deviance can be used to compare models (so can AIC, BIC, . . .), but using its absolute value to assess fit can be dangerous (exercise).

□ Residuals for binary data are not very informative.

□ Standard data setups, such as the 2×2 table, can be represented using binomial response models, and have nice representations in terms of the canonical parameter of a logistic regression model.

Count data slide 69

Types of count data

- □ $y \in \{0, 1, 2, \ldots\}$, perhaps with upper bound m, depending on sampling scheme:
  - counts, with no fixed total;
  - m individuals, subdivided into various categories:
    - ▷ **nominal response**—unordered categories (gender, nationality, ...)
    - ▷ **ordinal response**—ordered categories (pain level, spiciness of curry, ...)
- □ Simplest models:
  - single unbounded response, or Poisson approximation to binomial, takes $Y \sim \mathrm{Pois}(\mu)$;
  - group of responses $(Y_1, \ldots, Y_d)$ with fixed total $\sum Y_j = m$ has multinomial distribution, probabilities $(\pi_1, \ldots, \pi_d)$ and denominator m.
- □ Previous examples:
  - Doll and Hill data on smoking had response y Poisson with $\mu = T\lambda(x; \beta)$;
  - Jacamar data had ordinal (?) response N/S/E with total N+S+E fixed—multinomial with $d = 3$

Poisson and multinomial distributions

- □ $Y \sim \mathrm{Pois}(\mu)$ implies that

$$f(y; \mu) = \frac{\mu^y}{y!} e^{-\mu}, \quad y = 0, 1, 2, \ldots, \quad \mu > 0.$$

- □ Exponential family with natural parameter $\theta = \log \mu$, GLM with canonical logarithmic link, $x^{\mathrm{T}}\beta = \eta = \log \mu$.
- □ If Y is number of events in Poisson process of rate λ observed for period of length T, then $\mu = \lambda T$ and we set $\eta = x^{\mathrm{T}}\beta + \log T$
  - **offset** $\log T$ is fixed part of linear predictor η
- □ If $Y_r \overset{\mathrm{ind}}{\sim} \mathrm{Pois}(\mu_r)$, $r = 1, \ldots, d$, then the joint distribution of $Y_1, \ldots, Y_d$ given $Y_1 + \cdots + Y_d = m$ is **multinomial,** with denominator m, and probabilities

$$\pi_1 = \frac{\mu_1}{\sum_{r=1}^d \mu_r}, \quad \ldots, \quad \pi_d = \frac{\mu_d}{\sum_{r=1}^d \mu_r}.$$

- □ If $(Y_1, \ldots, Y_d) \sim \mathrm{Mult}(m; \pi_1, \ldots, \pi_d)$, then marginal and conditional distributions, e.g., of

$$(Y_1 + Y_2, Y_3 + Y_4 + Y_5, Y_6, \ldots, Y_d), \quad (Y_1, Y_2, Y_4) \mid (Y_3, Y_5, \ldots, Y_d),$$

are also multinomial.

Log-linear and logistic regressions

□ Special case: if $d = 2$, then

$$Y_2 \mid Y_1 + Y_2 = m \quad \sim \quad B\left(m, \pi = \frac{\mu_2}{\mu_1 + \mu_2}\right)$$

□ If $\mu_1 = \exp(\gamma + x_1^{\mathrm{T}}\beta)$, $\mu_2 = \exp(\gamma + x_2^{\mathrm{T}}\beta)$, then

$$\pi = \frac{\exp(\gamma + x_2^{\mathrm{T}}\beta)}{\exp(\gamma + x_1^{\mathrm{T}}\beta) + \exp(\gamma + x_2^{\mathrm{T}}\beta)} = \frac{\exp\{(x_2 - x_1)^{\mathrm{T}}\beta\}}{1 + \exp\{(x_2 - x_1)^{\mathrm{T}}\beta\}},$$

which corresponds to a logistic regression model for Y_2 with denominator m and probability π.

□ Can estimate β using log linear model or logistic model—but can't estimate γ from logistic model.

Poisson regression slide 73

Premier League data

```
> soccer
    month day year       team1       team2 score1 score2
1     Aug  19 2000    Charlton ManchesterC      4      0
2     Aug  19 2000     Chelsea     WestHam      4      2
3     Aug  19 2000    Coventry   Middlesbr      1      3
4     Aug  19 2000       Derby Southampton      2      2
5     Aug  19 2000       Leeds     Everton      2      0
6     Aug  19 2000   Leicester  AstonVilla      0      0
7     Aug  19 2000   Liverpool    Bradford      1      0
8     Aug  19 2000  Sunderland     Arsenal      1      0
9     Aug  19 2000   Tottenham     Ipswich      3      1
10    Aug  20 2000 ManchesterU   Newcastle      2      0
11    Aug  21 2000     Arsenal   Liverpool      2      0
12    Aug  22 2000    Bradford     Chelsea      2      0
13    Aug  22 2000     Ipswich ManchesterU      1      1
14    Aug  22 2000   Middlesbr   Tottenham      1      1
15    Aug  23 2000     Everton    Charlton      3      0
16    Aug  23 2000 ManchesterC  Sunderland      4      2
17    Aug  23 2000   Newcastle       Derby      3      2
18    Aug  23 2000 Southampton    Coventry      1      2
19    Aug  23 2000     WestHam   Leicester      0      1
20    Aug  26 2000     Arsenal    Charlton      5      3
 ...
```

Premier League data

□ 380 soccer matches in English Premier League in 2000–2001 season.

□ Data: home score y^h_{ij} and away score y^a_{ij} when team i is at home to team j, for $i, j, = 1, \dots, 20$, $i \neq j$.

□ Treat these as Poisson counts with means

$$\mu^h_{ij} = \exp(\Delta + \alpha_i - \beta_j), \quad \mu^a_{ij} = \exp(\alpha_j - \beta_i)$$

where

- Δ represents the home advantage;
- α_i and β_i represent the offensive and defensive strengths of team i.

□ Two possibilities for fitting:

- Poisson GLM, with 39 parameters;
- binomial GLM, with 20 parameters.

Premier League data: Analysis of deviance

Poisson model			Binomial model		
Terms	df	Deviance reduction	Terms	df	Deviance reduction
Home	1	33.58	Home	1	33.58
Defence	19	39.21	Team	19	79.63
Offence	19	58.85			
Residual	720	801.08	Residual	332	410.65

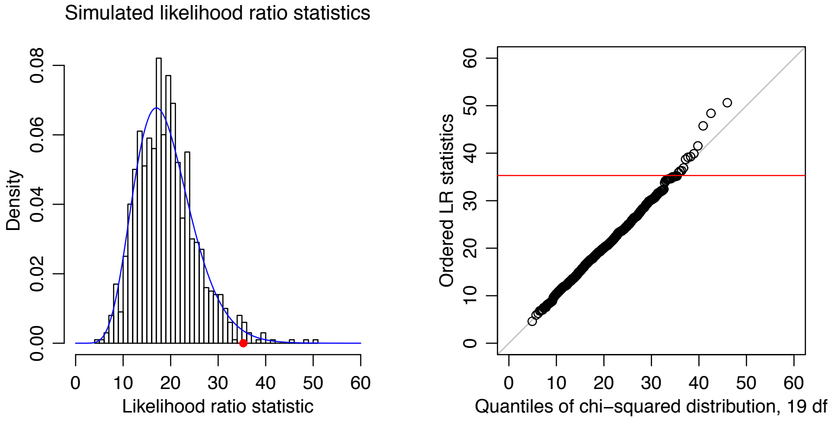
□ There's a strong effect of playing at home, and lots of evidence of differences among the teams—more in offence than defence.

□ Both residual deviances are a little large, but since the counts are small, we don't expect the large-sample χ^2 distribution to apply well to the residual deviance.

□ Simulations from the fitted model suggest that the residual deviances are not unusually large, so there's no evidence of a lack of fit.

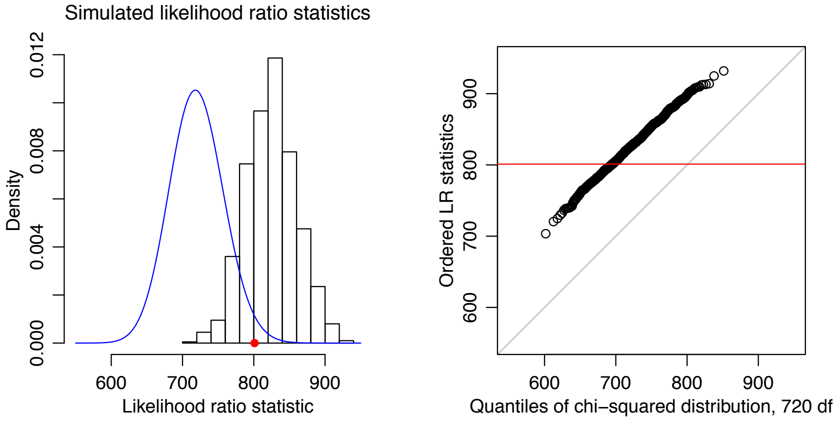
Premier League data: Null deviance for defence effect

Defence effect deviance (in red) for the Poisson model is large(ish) relative to χ^2_{19} distribution, but the asymptotics seem OK, based on simulations from a model without this effect (i.e., `Home + Offence`). It seems we can trust asymptotic distributions for differences of deviances, even though the counts are small.



Premier League data: Residual deviance

Residual deviance of 801 (in red) for the Poisson model seems large(ish) relative to χ^2_{720} distribution, but the asymptotics are suspect because most of the counts are small. Comparison of observed deviance with χ^2_{720} distribution shows that 801 is in fact somewhat smaller than average for datasets simulated from the fitted model.



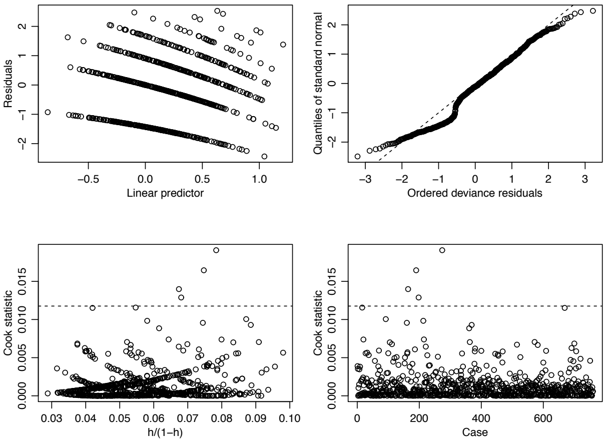
Premier League data: Estimates

	Overall (δ)	Offensive (α)	Defensive (β)
Manchester United	0.39	0.22	0.15
Liverpool	0.13	0.12	-0.08
Arsenal	—	0.04	—
Chelsea	-0.09	0.08	-0.22
Leeds	-0.10	0.02	-0.17
Ipswich	-0.16	-0.10	-0.13
Sunderland	-0.33	-0.31	-0.10
Aston Villa	-0.48	-0.31	-0.15
West Ham	-0.53	-0.33	-0.30
Middlesborough	-0.53	-0.35	-0.17
Charlton	-0.55	-0.21	-0.43
Tottenham	-0.58	-0.28	-0.38
Newcastle	-0.59	-0.35	-0.30
Southampton	-0.60	-0.45	-0.25
Everton	-0.75	-0.32	-0.46
Leicester	-0.77	-0.47	-0.31
Manchester City	-0.90	-0.40	-0.56
Coventry	-0.93	-0.53	-0.52
Derby	-0.93	-0.51	-0.45
Bradford	-1.29	-0.71	-0.62
SEs	0.29	0.20	0.20

Home advantage: $\widehat{\Delta} = 0.37\ (0.07)$, $\exp(\widehat{\Delta}) = 1.45$.

Premier League data: Assessment of fit

Diagnostic plots for fitted model: residuals against $\widehat{\eta}$ (top left); normal QQ-plot of residuals (top right); Cook statistic C_j against leverage ratio $h_j/(1-h_j)$ (lower left); Cook statistic C_j against case number (lower right).



Sampling schemes

□ A **contingency table** contains individuals (sampling units) cross-classified by various categorical variables.

– Example: the jacamar data cross-classify butterflies by

6 species × 8 colours × 3 fates

for a total of 144 categories, each with its number of butterflies $0, 1, \ldots, 14$.

□ The sampling scheme underlying a table may fix certain totals. Suppose a pollster wants to find out how people will vote in the coming US election. She might

– wait in the high street for a morning, and get opinions from those people willing to talk to her;

– wait until she has the views of a fixed number, say m, of people;

– wait until she has the views of fixed numbers of men and women.

Example 14 *Find the likelihoods for each of these sampling schemes, under (unrealistic!) assumptions of independence of voters.*

Note to Example 14

□ An $R \times C$ table arises by randomly sampling a population over a fixed period and then classifying the resulting individuals.

□ In the first scheme there are no constraints on the row and column totals, and a simple model is that the count in the (r, c) cell, y_{rc}, has a Poisson distribution with mean μ_{rc}. The resulting likelihood is

$$\prod_{r,c} \left\{ \frac{\mu_{rc}^{y_{rc}}}{y_{rc}!} e^{-\mu_{rc}} \right\};$$

this is simply the Poisson likelihood for the counts in the RC groups.

□ The pollster may set out with the intention of interviewing a fixed number m of individuals, stopping only when $\sum_{rc} y_{rc} = m$. In this case the data are multinomially distributed, with likelihood

$$\frac{m!}{\prod_{r,c} y_{rc}!} \prod_{r,c} \pi_{rc}^{y_{rc}}, \quad \sum_{r,c} \pi_{rc} = 1,$$

with $\pi_{rc} = \mu_{rc} / \sum_{s,t} \mu_{st}$ the probability of falling into the (r, c) cell.

□ A third scheme is to interview fixed numbers of men and of women, thus fixing the row totals $m_r = \sum_c y_{rc}$ in advance. In effect this treats the row categories as subpopulations, and the column categories as the response. This yields independent multinomial distributions for each row, and product multinomial likelihood

$$\prod_r \left\{ \frac{m_r!}{\prod_c y_{rc}!} \prod_c \pi_{rc}^{y_{rc}} \right\}, \quad \sum_c \pi_{1c} = \cdots = \sum_c \pi_{Rc} = 1,$$

in which $\pi_{rc} = \mu_{rc} / \sum_t \mu_{rt}$.

Contingency tables and Poisson response models

□ Multinomial models can be fitted using Poisson errors, provided the appropriate baseline terms are always included in the linear predictor.

□ Write the data as two-way layout, with C columns and R rows with fixed totals (e.g., $6 \times 8 = 48$ rows each with 3 columns for the jacamar data).

□ Consider Poisson model with means $\mu_{rc} = \exp(\gamma_r + x_{rc}^{\mathrm{T}}\beta)$:

- the row parameters $\gamma_1, \dots \gamma_R$ are **nuisance parameters**, not of interest;
- we want inference for the **parameter of interest**, β.

□ Corresponding multinomial model has fixed row totals m_r and probabilities

$$\pi_{rc} = \frac{\mu_{rc}}{\sum_{d=1}^{C} \mu_{rd}} = \frac{\exp(\gamma_r + x_{rc}^{\mathrm{T}}\beta)}{\sum_{d=1}^{C} \exp(\gamma_r + x_{rd}^{\mathrm{T}}\beta)} = \frac{\exp(x_{rc}^{\mathrm{T}}\beta)}{\sum_{d=1}^{C} \exp(x_{rd}^{\mathrm{T}}\beta)},$$

for $r = 1, \dots, R$, $c = 1, \dots, C$; i.e., one multinomial variable for each row.

□ The resulting multinomial log likelihood is

$$\begin{aligned}\ell_{\text{Mult}}(\beta; y \mid m) &\equiv \sum_{r=1}^{R}\sum_{c=1}^{C} y_{rc} \log \pi_{rc} \\ &= \sum_{r=1}^{R}\left\{\sum_{c=1}^{C} y_{rc} x_{rc}^{\mathrm{T}}\beta - m_r \log\left(\sum_{d=1}^{C} e^{x_{rd}^{\mathrm{T}}\beta}\right)\right\}.\end{aligned}$$

Contingency tables and Poisson response models, II

Lemma 15 *Show that if parameters τ_r for the row margins are included in the above setup, then we can write*

$$\ell_{\text{Poiss}}(\beta, \tau) = \ell_{\text{Poiss}}(\tau; m) + \ell_{\text{Mult}}(\beta; y \mid m).$$

□ Implications:

- the MLEs of β and τ based on the LHS are the same as those from separate maximisations of the terms on the right:
  - ▷ $\widehat{\beta}$ equals the MLE for the multinomial model,
  - ▷ $\widehat{\tau}_r = m_r$
- the observed and expected information matrices for β, τ are block diagonal.
- SEs based on the multinomial and Poisson models are equal (exercise).

□ General conclusion: inferences on β are the same for multinomial and Poisson models,

> *provided the parameters associated to the margins fixed under the multinomial model, i.e., the γ_r, are included in the Poisson fit.*

Note to Lemma 15

□ The Poisson model has no conditioning, so the log likelihood is

$$\ell_{\text{Poiss}}(\beta,\gamma) \equiv \sum_{r,c}(y_{rc}\log\mu_{rc}-\mu_{rc}) = \sum_{r=1}^{R}\left(m_r\gamma_r + \sum_{c=1}^{C} y_{rc}x_{rc}^{\mathrm{T}}\beta - e^{\gamma_r}\sum_{c=1}^{C} e^{x_{rc}^{\mathrm{T}}\beta}\right),$$

where we use the fact that $\log\mu_{rc} = \gamma_r + x_{rc}^{\mathrm{T}}\beta$.

□ Now we reparametrise in terms of the row totals $\tau_r = \sum_c \mu_{rc}$, noting that

$$\tau_r = e^{\gamma_r}\sum_{c=1}^{C} e^{x_{rc}^{\mathrm{T}}\beta}, \quad \gamma_r = \log\tau_r - \log\left\{\sum_{c=1}^{C}\exp(x_{rc}^{\mathrm{T}}\beta)\right\},$$

so

$$\begin{aligned}\ell_{\text{Poiss}}(\beta,\tau) &\equiv \sum_{r=1}^{R}(m_r\log\tau_r-\tau_r) + \sum_{r=1}^{R}\left\{\sum_{c=1}^{C} y_{rc}x_{rc}^{\mathrm{T}}\beta - m_r\log\left(\sum_{c=1}^{C} e^{x_{rc}^{\mathrm{T}}\beta}\right)\right\},\\ &= \ell_{\text{Poiss}}(\tau;m) + \ell_{\text{Mult}}(\beta;y \mid m),\end{aligned}$$

which is the log likelihood corresponding to

– independent Poisson row totals m_r with means τ_r, and, independent of this,

– the multinomial log likelihood for the contingency table.

Jacamar data

Response (N=not sampled, S = sampled and rejected, E = eaten) of a rufous-tailed jacamar to individuals of seven species of palatable butterflies with artifically coloured wing undersides. Data from Peng Chai, University of Texas.

	Aphrissa boisduvalli	*Phoebis argante*	*Dryas iulia*	*Pierella luna*	*Consul fabius*	*Siproeta stelenes*†
	N/S/E	N/S/E	N/S/E	N/S/E	N/S/E	N/S/E
Unpainted	0/0/14	6/1/0	1/0/2	4/1/5	0/0/0	0/0/1
Brown	7/1/2	2/1/0	1/0/1	2/2/4	0/0/3	0/0/1
Yellow	7/2/1	4/0/2	5/0/1	2/0/5	0/0/1	0/0/3
Blue	6/0/0	0/0/0	0/0/1	4/0/3	0/0/1	0/1/1
Green	3/0/1	1/1/0	5/0/0	6/0/2	0/0/1	0/0/3
Red	4/0/0	0/0/0	6/0/0	4/0/2	0/0/1	3/0/1
Orange	4/2/0	6/0/0	4/1/1	7/0/1	0/0/2	1/1/1
Black	4/0/0	0/0/0	1/0/1	4/2/2	7/1/0	0/1/0

† includes *Philaethria dido* also.

Jacamar data: Models

□ Let factors F, S, C represent the 3 fates, the 6 species, and the 8 colours.

□ The models $C * S$, $C * S + F$, and $C * S + C * F$ mean we set

$$\log \mu_{csf} = \alpha_{cs}, \quad \log \mu_{csf} = \alpha_{cs} + \gamma_f, \quad \log \mu_{csf} = \alpha_{cs} + \gamma_{cf}.$$

□ The vector of probabilities corresponding to the model with terms $C * S$ is

$$(\pi_{cs1}, \pi_{cs2}, \pi_{cs3}) = \left(\frac{\mu_{cs1}}{\sum_{f=1}^{3} \mu_{csf}}, \frac{\mu_{cs2}}{\sum_{f=1}^{3} \mu_{csf}}, \frac{\mu_{cs3}}{\sum_{f=1}^{3} \mu_{csf}} \right) = (\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}),$$

and that corresponding to the model with terms $C * S + F$ is

$$\begin{aligned}(\pi_{cs1}, \pi_{cs2}, \pi_{cs3}) &= \left(\frac{\mu_{cs1}}{\sum_{f=1}^{3} \mu_{csf}}, \frac{\mu_{cs2}}{\sum_{f=1}^{3} \mu_{csf}}, \frac{\mu_{cs3}}{\sum_{f=1}^{3} \mu_{csf}} \right) \\ &= \frac{1}{e^{\gamma_1} + e^{\gamma_2} + e^{\gamma_3}} \left(e^{\gamma_1}, e^{\gamma_2}, e^{\gamma_3} \right).\end{aligned}$$

□ Exercise: similar computations for $C * S + C * F$ and $C * S + C * F + S * F$.

Jacamar data: Analysis of deviance

Deviances for log-linear models fitted to jacamar data.

Terms	df	Deviance
$C * S$	88	259.42
$C * S + F$	86	173.86
$C * S + C * F$	72	139.62
$C * S + S * F$	76	148.23
$C * S + C * F + S * F$	62	90.66
$C * S * F$	0	0

□ The null model $C * S$ is not of interest.

□ The first model it is sensible to fit is $C * S + F$.

□ The best model seems to be $C * S + C * F + S * F$, corresponding to independent effects of species and colour, though its deviance is high (but remember the two outlying cells!)

Smoking data

Lung cancer deaths in British male physicians. The table gives man-years at risk T/number of cases y of lung cancer, cross-classified by years of smoking t, taken to be age minus 20 years, and number of cigarettes smoked per day, d.

Years of smoking t	Daily cigarette consumption d						
	Nonsmokers	1–9	10–14	15–19	20–24	25–34	35+
15–19	10366/1	3121	3577	4317	5683	3042	670
20–24	8162	2937	3286/1	4214	6385/1	4050/1	1166
25–29	5969	2288	2546/1	3185	5483/1	4290/4	1482
30–34	4496	2015	2219/2	2560/4	4687/6	4268/9	1580/4
35–39	3512	1648/1	1826	1893	3646/5	3529/9	1336/6
40–44	2201	1310/2	1386/1	1334/2	2411/12	2424/11	924/10
45–49	1421	927	988/2	849/2	1567/9	1409/10	556/7
50–54	1121	710/3	684/4	470/2	857/7	663/5	255/4
55–59	826/2	606	449/3	280/5	416/7	284/3	104/1

Smoking data: Models

□ Suppose number of deaths y has Poisson distribution, mean $T\lambda(d,t)$, where T is man-years at risk, d is number of cigarettes smoked daily and t is time smoking (years).

□ Log-linear model:

- $\lambda_{rc} = \exp(\gamma_r + \beta_c)$, $r = 1, \ldots, 9$, $c = 1, \ldots, 7$;
- one parameter for each row and column, 15 paras in all;
- deviance 51.47 on 48 df (AIC is 81.47).

□ Substantive model (not log-linear):

- $\lambda(d,t) = (\beta_0 + \beta_1 d^{\beta_2})t^{\beta_3}$, so
  - ▷ background rate of lung cancer is $\beta_0 t^{\beta_3}$ for non-smoker;
  - ▷ additional risk due to smoking d cigarettes/day is $\beta_1 d^{\beta_2}$;
- just 4 parameters;
- deviance is 59.58 on 59 df (AIC is 67.48).

□ Substantive model is better, more parsimonious, and has a simpler interpretation.

Smoking data: Substantive model

□ Likelihood ratio test of $\beta_1 = 0$ or $\beta_2 = 0$ would be non-regular: why?

□ Reparametrize to avoid constraints $\beta_0, \beta_1 > 0$ in maximisation: set

$$\lambda(d, t) = \{e^{\gamma_0} + \exp(\gamma_1 + \beta_2 \log d)\} \exp(\beta_3 \log t),$$

with $t = 1$ for age 62.5 years.

□ Parameter estimates (standard errors):

	γ_0	γ_1	β_2	β_3
Smokers only	0.96 (25.4)	2.15 (1.45)	1.20 (0.40)	4.50 (0.34)
All data	2.94 (0.58)	1.82 (0.66)	1.29 (0.20)	4.46 (0.33)
All data ($\beta_2 = 1$)	2.75 (0.56)	2.72 (0.09)	—	4.43 (0.33)

□ Precision of $\widehat{\gamma}_0$ depends on data for non-smokers—their death-rate at age 62.5 is $e^{\widehat{\gamma}_0} = 18.9$ per 100,000 years at risk.

□ Parameter estimates suggest $\beta_2 = 1$, then get deviance of 61.84 on 60 df, an increase of $61.84 - 59.58 = 2.26 \stackrel{\cdot}{\sim} \chi^2_1$, if the simpler model is OK.

□ Beware small counts:

– χ^2 approximation to distribution of deviance unreliable — but simulation shows that the models fit well;

– residuals not *very* useful, because too discrete.

Problems with log-linear models

□ Log-linear models are mathematically elegant, but have some statistical drawbacks.

□ Consider the data below, which contain 2×2 tables for visual impairment for 4 age groups $\times$ 2 race groups.

□ A natural initial analysis would be to fit logistic models for impairment of the left eye and the right eye separately, i.e., fitting models of the form

$$\begin{aligned} y_{01} + y_{11} &\sim B\{y_{00} + y_{10} + y_{01} + y_{11}, \pi(x^{\mathrm{T}}\beta_R)\}, \\ y_{10} + y_{11} &\sim B\{y_{00} + y_{10} + y_{01} + y_{11}, \pi(x^{\mathrm{T}}\beta_L)\}, \end{aligned}$$

to each of the eight tables, resulting in estimates of β_R and β_L for the two eyes separately.

Eye		Prevalence for whites aged				Prevalence for blacks aged			
Left	Right	40–50	51–60	61–70	70+	40–50	51–60	61–70	70+
0	0	602	541	752	606	729	551	452	307
1	0	11	15	31	60	19	24	22	29
0	1	15	16	37	67	21	23	21	37
1	1	4	9	11	79	10	14	28	56

Joint distribution of visual impairment on both eyes by race and age combinations. Combination $(0, 0)$ means neither eye is visually impaired.

Problems with log-linear models, II

- □ Each of these analyses is individually valid, and should explain how the probability of visual impairment depends on age and race.
- □ If we formulate a joint log-linear model, we can write the values in a 2×2 table as independent Poisson variables with means

$$\exp(\gamma), \quad \exp(\gamma+\gamma_L), \quad \exp(\gamma+\gamma_R), \quad \exp(\gamma+\gamma_L+\gamma_R+\gamma_{LR}),$$

which gives

$$(\pi_{00}, \pi_{01}; \pi_{10}, \pi_{11}) = \frac{1}{1+e^{\gamma_R}+e^{\gamma_L}+e^{\gamma_R+\gamma_L+\gamma_{LR}}} \left(1, e^{\gamma_R}, e^{\gamma_L}, e^{\gamma_R+\gamma_L+\gamma_{LR}}\right),$$

where $\gamma_L = x^{\mathrm{T}}\delta_L$, $\gamma_R = x^{\mathrm{T}}\delta_R$, $\gamma_{LR} = x^{\mathrm{T}}\delta_{LR}$.

- □ The marginal probability of an impaired left eye is

$$\pi_L' = \frac{e^{\gamma_L}+e^{\gamma_R+\gamma_L+\gamma_{LR}}}{1+e^{\gamma_R}+e^{\gamma_L}+e^{\gamma_R+\gamma_L+\gamma_{LR}}},$$

which equals $e^{\gamma_L}/(1+e^{\gamma_L})$ only when $\gamma_{LR}=0$, so visual impairment occurs independently in each eye. Otherwise the marginal probability of an impaired left eye depends on γ_R and γ_{LR}, implying that the initial logistic fits shed no light on γ_L.

Marginal models

- □ Maybe more natural to write for the left-eye binomial probability that

$$y_{10}+y_{11} \sim B\left\{y_{00}+y_{10}+y_{01}+y_{11}, \pi_{10}+\pi_{11} = \pi_L = \frac{\exp(x^{\mathrm{T}}\beta_L)}{1+\exp(x^{\mathrm{T}}\beta_L)}\right\},$$

say.

- □ Then augment π_L and π_R by adding further parameters, e.g., setting

$$\frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}} = \frac{\pi_{11}(1-\pi_L-\pi_R+\pi_{11})}{(\pi_L-\pi_{11})(\pi_R-\pi_{11})} = \exp(x^{\mathrm{T}}\beta_{LR}).$$

If $x^{\mathrm{T}}\beta_{LR} = \gamma$ was independent of x, then there would be constant association between the eyes after adjusting for marginal effects of age and race, with more complicated models indicating more complex patterns of association.

- □ As $0 < \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11} < 1$, we find that

$$\pi_{11} \in (\max(0, \pi_L+\pi_R-1), \min(\pi_L, \pi_R)),$$

and π_{11} is the root of a quadratic equation whose coefficients depends on π_L, π_R, and $x^{\mathrm{T}}\beta_{LR}$, thereby enabling us to express the probabilities in each 2×2 table in terms of the marginal probabilities and the odds ratio.

- □ Easier to interpret than a log-linear model, but much less elegant.

Pneumoconiosis data

Period of exposure x and prevalence of pneumoconiosis amongst coalminers.

	Period of exposure (years)							
	5.8	15	21.5	27.5	33.5	39.5	46	51.5
Normal	98	51	34	35	32	23	12	4
Present	0	2	6	5	10	7	6	2
Severe	0	1	3	8	9	8	10	5

□ Here

$$\text{Normal} \quad < \quad \text{Present} \quad < \quad \text{Severe},$$

so these are ordinal responses with $d = 3$ categories and the total in each group (corresponding to each period of exposure) fixed.

□ It probably is reasonable to imagine that the choice of category stems from an underlying continuous variable, even if this cannot be quantified very well.

Models

□ Assume we have n independent individuals whose responses $I_1, \ldots, I_n$ fall into the set $\{1, \ldots, d\}$, corresponding to d ordered categories, and that

$$\gamma_l = \mathrm{P}(I_j \le l) = \pi_1 + \cdots + \pi_l, \quad l = 1, \ldots, d, \quad \gamma_d = 1,$$

□ The corresponding likelihood is $\prod_{j=1}^n \pi_{I_j}$, where usually the contribution $\pi_{I_j} \equiv \pi_{I_j}(\eta_j)$ for individual j will depend on covariates x_j through a linear predictor $\eta_j = x_j^{\mathrm{T}}\beta$.

□ We often want the interpretation of the parameters not to change if we merge adjacent categories, and we can do this using an underlying tolerance distribution, with

$$I_j = l \quad \Leftrightarrow \quad x_j^{\mathrm{T}}\beta + \varepsilon_j \in (\zeta_{l-1}, \zeta_l], \quad \zeta_0 = -\infty < \zeta_1 < \cdots < \zeta_{d-1} < \zeta_d = \infty,$$

where the tolerance distribution F of ε_j is often taken to be logistic, giving the **proportional odds model**, and

$$\pi_l(x_j^{\mathrm{T}}\beta) = \mathrm{P}(\zeta_{l-1} < x_j^{\mathrm{T}}\beta + \varepsilon \le \zeta_l) = F(\zeta_l - x_j^{\mathrm{T}}\beta) - F(\zeta_{l-1} - x_j^{\mathrm{T}}\beta), \quad l = 1, \ldots, d;$$

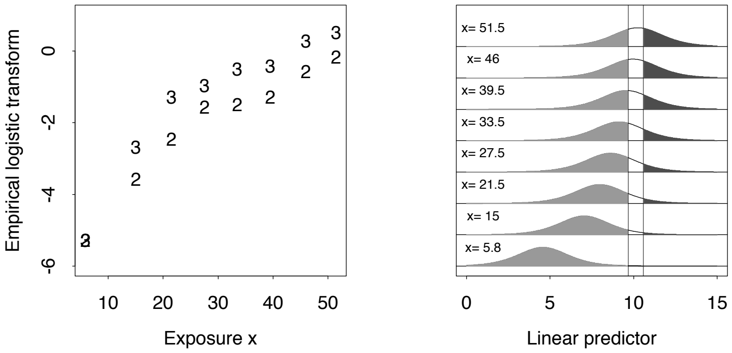
here $\zeta_1, \ldots, \zeta_{d-1}$ are aliased with an intercept β_0 and are not usually of interest.

□ Another standard choice is $F(u) = 1 - \exp\{-\exp(u)\}$.

□ To fit, we just apply IWLS to the multinomial likelihood $\prod_{j=1}^n \pi_{I_j}$.

Pneumoconiosis data

Pneumoconiosis data analysis, showing how the implied fitted logistic distributions depend on x. Left: plots of empirical logistic transforms for comparing categories 1 with $2+3$ and $1+2$ with 3; the nonlinearity suggests using $\log x$ as covariate. Right: fitted model, showing probabilities for the three groups with an underlying logistic distribution.



Final comments

- □ Log-linear models are mathematically elegant and useful defaults for count data, with close links to logistic regression, based on the relation between the Poisson and multinomial distributions.
- □ Interpretation of log-linear models can be difficult, especially for contingency tables, because marginal and conditional parameters cannot be disentangled.
- □ Marginal models less elegant mathematically, but have better interpretations in practice.
- □ Also possible to fit models for ordinal data, using multinomial models and tolerance distribution interpretation used for binomial data.

Overdispersion

□ Often find that discrete response data are more variable than might be expected from a simple Poisson or binomial model, so we see

- residual deviances larger than expected
- residuals more variable than expected under the model

but otherwise no evidence of systematic lack of fit

□ This is **overdispersion**, perhaps due to effect of unmeasured explanatory variables on the responses.

UK monthly AIDS reports 1983–1992

Diagnosis period		Reporting-delay interval (quarters):									Total reports to end of 1992
Year	Quarter	$0^\dagger$	1	2	3	4	5	6	···	≥14	
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1988	1	31	80	16	9	3	2	8	···	6	174
	2	26	99	27	9	8	11	3	···	3	211
	3	31	95	35	13	18	4	6	···	3	224
	4	36	77	20	26	11	3	8	···	2	205
1989	1	32	92	32	10	12	19	12	···	2	224
	2	15	92	14	27	22	21	12	···	1	219
	3	34	104	29	31	18	8	6	···		253
	4	38	101	34	18	9	15	6	···		233
1990	1	31	124	47	24	11	15	8	···		281
	2	32	132	36	10	9	7	6	···		245
	3	49	107	51	17	15	8	9	···		260
	4	44	153	41	16	11	6	5	···		285
1991	1	41	137	29	33	7	11	6	···		271
	2	56	124	39	14	12	7	10	···		263
	3	53	175	35	17	13	11	2			306
	4	63	135	24	23	12	1				258
1992	1	71	161	48	25	5					310
	2	95	178	39	6						318
	3	76	181	16							273
	4	67	66								133

AIDS data

□ UK monthly reports of AIDS diagnoses 1983–1992, with reporting delay up to several years!

□ Example of incomplete contingency table (very common in insurance)

□ Simple (chain-ladder) model: number of reports in row j and column k is Poisson, with mean

$$\mu_{jk} = \exp(\alpha_j + \beta_k).$$

□ Analysis of Deviance:

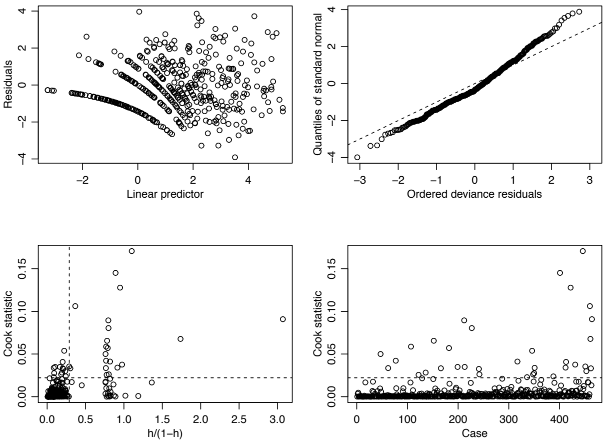
Model	df	Deviance reduction	df	Deviance
			464	14184.3
Time (rows)	37	6114.8	427	8069.5
Delay (cols)	14	7353.0	413	716.5

□ Residual deviance is obviously far too large for a Poisson model to be OK, but the model is also too complex, since we expect smooth variation in the α_j.

□ Next page shows residual analysis: no obvious problems, just generic overdispersion.

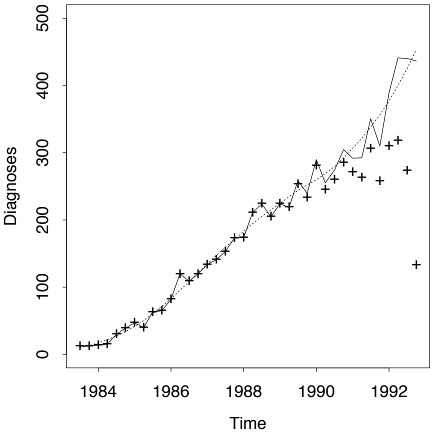
AIDS data: Assessment of fit

Diagnostic plots for fitted model: residuals against $\widehat{\eta}$ (top left); normal QQ-plot of residuals (top right); Cook statistic C_j against leverage ratio $h_j/(1-h_j)$ (lower left); Cook statistic C_j against case number (lower right).



AIDS data

- □ Data (+) and predicted true numbers based on simple Poisson model (solid) and GAM (dots).
- □ The Poisson model and data agree up to where data start to be missing.



Dealing with overdispersion

- □ Two basic approaches to accommodating overdispersion:
  - parametric modelling
  - quasi-likelihood estimation, based only on the variance function

Example 16 (Linear and quadratic variance functions) *Suppose that, conditional on $\varepsilon > 0$, $Y \sim \mathrm{Pois}(\mu\varepsilon)$, where $\mathrm{E}(\varepsilon) = 1$ and $\mathrm{var}(\varepsilon) = \xi$. Show that this can lead to either linear or quadratic variance functions, but a lot of data may be needed to distinguish them.*

Comparison of variance functions for overdispersed count data. The linear and quadratic variance functions are $V_L(\mu) = (1 + \xi_L)\mu$ and $V_Q(\mu) = \mu(1 + \xi_Q\mu)$, with $\xi_L = 0.5$ and ξ_Q chosen so that $V_L(15) = V_Q(15)$.

μ	1	2	5	10	15	20	30	40	60
Linear	1.5	3.0	7.5	15.0	22.5	30	45	60	90
Quadratic	1.0	2.1	5.8	13.3	22.5	33	60	93	180

Note to Example 16

Let ε have unit mean and variance $\xi > 0$, and to be concrete suppose that conditional on ε, Y has the Poisson distribution with mean $\mu\varepsilon$. Then

$$\mathrm{E}(Y) = \mathrm{E}_\varepsilon\left\{\mathrm{E}(Y \mid \varepsilon)\right\}, \quad \mathrm{var}(Y) = \mathrm{var}_\varepsilon\left\{\mathrm{E}(Y \mid \varepsilon)\right\} + \mathrm{E}_\varepsilon\left\{\mathrm{var}(Y \mid \varepsilon)\right\},$$

so the response has mean and variance

$$\mathrm{E}(Y) = \mathrm{E}_\varepsilon(\mu\varepsilon) = \mu, \quad \mathrm{var}(Y) = \mathrm{var}_\varepsilon(\mu\varepsilon) + \mathrm{E}_\varepsilon(\mu\varepsilon) = \mu(1+\xi\mu).$$

If on the other hand the variance of ε is ξ/μ, then $\mathrm{var}(Y) = (1+\xi)\mu$. In both cases the variance of Y is greater than its value under the standard Poisson model, for which $\xi = 0$. In the first case the variance function is quadratic, and in the second it is linear.

Negative binomial model

Example 17 (Negative binomial) *In Example 16, if ε is gamma with shape parameter $1/\nu$, show that*

$$f(y;\mu,\nu) = \frac{\Gamma(y+\nu)}{\Gamma(\nu)y!}\frac{\nu^\nu\mu^y}{(\nu+\mu)^{\nu+y}}, \quad y = 0, 1, \ldots, \quad \mu, \nu > 0,$$

and that quadratic and linear variance functions are obtained on setting $\nu = 1/\xi$ and $\nu = \mu/\xi$ respectively.
The log link function $\log\mu = x^{\mathrm{T}}\beta$ is most natural.
ξ is estimated by maximum likelihood or through Pearson's statistic.

Example 18 (AIDS data)

□ *MLE $\widehat{\xi}_Q = 22.7\ (5.5)$*

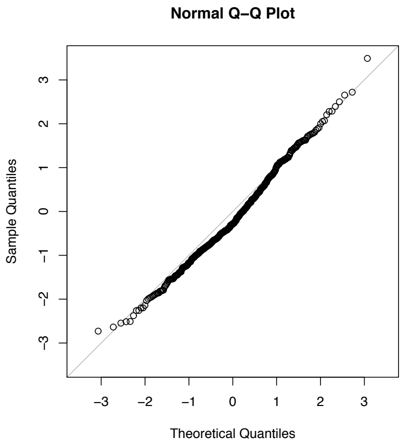
□ *Analysis of Deviance (with $\widehat{\xi}_Q$ fixed):*

Model	*df*	*Deviance reduction*	*df*	*Deviance*
			464	*7998.3*
Time (rows)	*37*	*3582.5*	*427*	*4415.8*
Delay (cols)	*14*	*3892.2*	*413*	*523.6*

□ *Still somewhat overdispersed?*

AIDS data: Deviance residuals for NB model

Clear improvement over previous plots, even if not perfect.



Quasi-likelihood

□ Recall two basic assumptions for the linear model:

– the responses are uncorrelated with means $\mu_j = x_j^{\mathrm{T}}\beta$ and equal variances σ^2;

– in addition to this, the responses are normally distributed.

□ To avoid parametric modelling, we generalise the second-order assumptions, to

$$\mathrm{E}(Y_j) = \mu_j, \quad \mathrm{var}(Y_j) = \phi_j V(\mu_j), \quad g(\mu_j) = \eta_j = x_j^{\mathrm{T}}\beta,$$

where the variance function $V(\cdot)$ and the link function are taken as known.

□ We obtain estimates $\tilde\beta$ by solving the estimating equation

$$h(\beta; Y) = X^{\mathrm{T}}u(\beta) = \sum_{j=1}^{n} x_j u_j(\beta) = \sum_{j=1}^{n} x_j \frac{Y_j - \mu_j}{g'(\mu_j)\phi_j V(\mu_j)} = 0.$$

□ If the mean structure is correct, then $\mathrm{E}(Y_j) = \mu_j$, so $\mathrm{E}\{h(\beta; Y)\} = 0$, and under mild conditions $\tilde\beta$ is consistent (but maybe not efficient) as $n \to \infty$.

Quasi-likelihood II

Recall that the general variance of an estimator $\tilde\beta$ defined by an estimating equation $h(\beta;Y)=0$ has sandwich form

$$\mathrm{E}\left\{-\frac{\partial h(\beta;Y)}{\partial\beta^{\mathrm{T}}}\right\}^{-1}\mathrm{var}\left\{h(\beta;Y)\right\}\mathrm{E}\left\{-\frac{\partial h(\beta;Y)^{\mathrm{T}}}{\partial\beta}\right\}^{-1}.$$

Lemma 19 *If $V(\mu)$ is correctly specified, then $\mathrm{var}\{\tilde\beta\}\doteq(X^{\mathrm{T}}WX)^{-1}$, where W is diagonal with (j,j) element $\{g'(\mu_j)^2\phi_jV(\mu_j)\}^{-1}$.*

□ If $\phi_j=\phi a_j$, with known $a_j>0$ and unknown $\phi>0$, then we obtain
  - $\tilde\beta$ by fitting the GLM with variance function $V(\mu)$ and link $g(\mu)$;
  - standard errors by multiplying the standard errors for this fit by $\widehat\phi^{1/2}$, where

$$\widehat\phi=\frac{1}{n-p}\sum_{j=1}^{n}\frac{(y_j-\widehat\mu_j)^2}{a_jg'(\mu_j)^2V(\widehat\mu_j)}.$$

Note to Lemma 19

We require $\mathrm{E}\{-\partial h(\beta;Y)/\partial\beta^{\mathrm{T}}\}$ and $\mathrm{var}\{h(\beta;Y)\}$. Now

$$\begin{aligned}\frac{\partial u_j(\beta)}{\partial\beta^{\mathrm{T}}} &= \frac{\partial\eta_j}{\partial\beta^{\mathrm{T}}}\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial u_j(\beta)}{\partial\mu_j}\\ &= x_j^{\mathrm{T}}\frac{1}{g'(\mu_j)}\left\{-\frac{g''(\mu_j)}{g'(\mu_j)}u_j(\beta)-\frac{V'(\mu_j)}{V(\mu_j)}u_j(\beta)-\frac{1}{g'(\mu_j)\phi_jV(\mu_j)}\right\},\end{aligned}$$

and as $\mathrm{E}\{u_j(\beta)\}=0$, it follows that

$$\begin{aligned}\mathrm{E}\left\{-\frac{\partial h(\beta;Y)}{\partial\beta^{\mathrm{T}}}\right\} &= -\sum_{j=1}^{n}x_j\mathrm{E}\left\{\frac{\partial u_j(\beta)}{\partial\beta^{\mathrm{T}}}\right\}\\ &= \sum_{j=1}^{n}x_jx_j^{\mathrm{T}}\frac{1}{g'(\mu_j)^2\phi_jV(\mu_j)}=X^{\mathrm{T}}WX,\end{aligned}$$

where W is the $n\times n$ diagonal matrix with $j\mathrm{th}$ element $\{g'(\mu_j)^2\phi_jV(\mu_j)\}^{-1}$. Moreover if in addition the variance function has been correctly specified, then $\mathrm{var}(Y_j)=\phi_jV(\mu_j)$, and hence

$$\mathrm{var}\{h(\beta;Y)\}=X^{\mathrm{T}}\mathrm{var}\{u(\beta)\}X=\sum_{j=1}^{n}x_jx_j^{\mathrm{T}}\frac{\mathrm{var}(Y_j)}{g'(\mu_j)^2\phi_j^2V(\mu_j)^2}=X^{\mathrm{T}}WX.$$

Thus the sandwich equals $(X^{\mathrm{T}}WX)^{-1}$. Had the variance function been wrongly specified, the variance matrix of $\tilde\beta$ would have been of form $(X^{\mathrm{T}}WX)^{-1}(X^{\mathrm{T}}W'X)(X^{\mathrm{T}}WX)^{-1}$, where W' is a diagonal matrix involving the true and assumed variance functions. Only if the variance function has been chosen very badly will this sandwich matrix differ greatly from $(X^{\mathrm{T}}WX)^{-1}$, which therefore provides useful standard errors unless a plot of absolute residuals against fitted means is markedly non-random. In that case the choice of variance function should be reconsidered.

Quasi-likelihood III

- Under an exponential family model, $h(\beta; Y)$ is the score statistic, so $\tilde{\beta}$ is the MLE and is efficient (i.e., it has the smallest possible variance in large samples).
- If not, inference is valid provided g and V are correctly chosen, and $\tilde{\beta}$ is optimal among estimators based on linear combinations of the $Y_j - \mu_j$, by extending the Gauss–Markov theorem.
- In fact we can define a **quasi-likelihood** Q and its score through

  $$Q(\beta; Y) = \sum_{j=1}^{n} \int_{Y_j}^{\mu_j} \frac{Y_j - u}{\phi a_j V(u)} \, \mathrm{d}u, \quad h(\beta; Y) = \frac{\partial}{\partial \beta} Q(\beta; Y),$$

  and a (quasi-)deviance as $D = -2\phi Q(\beta; Y)$.
- To compare models A, B with numbers of parameters $p_B < p_A$ and deviances $D_B > D_A$, we use the fact that

  $$\frac{(D_B - D_A)/(p_A - p_B)}{\widehat{\phi}_A} \quad \dot{\sim} \quad F_{p_A - p_B, n - p_A},$$

  if the simpler model B is adequate. This is easy in R.

AIDS example

```
> aids.ql <- glm(y~factor(time)+factor(delay),family=quasipoisson,data=aids.in)
> anova(aids.ql,test="F")
Analysis of Deviance Table

Model: quasipoisson, link: log

Response: y

Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                            464    14184.3
factor(time)  37   6114.8       427     8069.5  92.638 < 2.2e-16 ***
factor(delay) 14   7353.0       413      716.5 294.402 < 2.2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Summary

□ Overdispersion is widespread in count and proportion data.

□ We deal with it either by

- parametric modelling, or
- quasi-likelihood (QL) estimation, which involves assumptions only on the mean-variance relationship.

□ QL estimators equal the ML ones, but SEs are inflated by $\widehat{\phi}^{1/2}$.

□ (Quasi-)deviance can also be defined, and used for model comparison, with F tests replacing χ^2 tests.