Today

- theory for generalized linear models
- examples of generalized linear models
- advice from C & D
- thoughts on Shaghayegh's study

C & D §2.6

- overall size of the investigation (*What should be my n?*)
- amount of replication at various levels
- "In those situations where resources for the investigation are limited, or, for example, access to suitable patients limited in a clinical trial, the issue will be not so much calculating the size of study desirable but with establishing whether the resources available and the number of patients likely to be accrued are sufficient to make it likely that a useful conclusion will be reached."
- comparison of two means: $m = 2\sigma^2/c$ where c is the bound desired on the comparison
- $\mathrm{Var}(\bar{y}_1 - \bar{y}_2) = 2\sigma^2/m < c \Rightarrow m > 2\sigma^2/c$
- power of a test: $m = 2\sigma^2(z_\alpha + z_\beta)^2/d^2$
- $\Pr\{\dfrac{|\bar{y}_1 - \bar{y}_2 - d|}{\sqrt{2\sigma^2/m}} > z_\alpha\} \geq 1 - \beta$

... §2.6

- "in most situations in which there are a large number of qualitatively different treatments or exposures under comparison it is reasonable to aim for exactly or approximately equal replication of the different treatments"
- "An exception is when there is a control and a number of other treatments and interest focuses on comparisons of the other treatments one at a time with the control. It is then reasonable to have approximately $\sqrt{t}$ observations on the control for each observation on the other treatments. "
  How would you prove this?

C & D, Ch. 3 Special types of study

- sampling a specific population
- experiments – experimental units, treatments under control of investigator, avoidance of systematic error by randomizaton
- "pristine simplicity of interpretation: some units randomized to T, some to C, all other aspects remaining the same. ... If there is an appreciable difference [in a measured outcome] then either it is a consequence of the play of chance or represents an effect produced by the distinction between T and C"
- potential complications
- example: non-compliance – intention-to-treat analysis; if feasible record reasons for non-compliance

C & D, randomized block designs

- $n = bt$ experimental units; units formed into b blocks, each block containing t units; t treatments assigned at random to the units in each block
- Table of comparisons

  "numbers of logically independent contrasts on an additive scale"
- $y_{ts} = \bar{y}_{..} + (\bar{y}_{t.} - \bar{y}_{..}) + (\bar{y}_{.s} - \bar{y}_{..}) + (y_{ts} - \bar{y}_{t.} - \bar{y}_{.s} + \bar{y}_{..})$
- if the final set of terms is set out in a table, all row and column sums are zero, thus the table can be reconstructed from any set of $(t-1)(b-1)$ of the entries"

Generalized linear models: theory

- model: $f(y_j; \mu_j, \phi_j) = \exp\{\dfrac{y_j\theta_j - b(\theta_j)}{\phi_j} + c(y_j; \phi_j)\}$
- $E(y_j \mid x_j) = b'(\theta_j) = \mu_j$ defines μ_j as a function of θ_j
- $g(\mu_j) = x_j^T\beta = \eta_j$ links the n observations together via covariates
- $g(\cdot)$ is the link function; η_j is the linear predictor
- $\mathrm{Var}(y_j \mid x_j) = \phi b''(\theta_j) = \phi V(\mu_j)$
- $V(\cdot)$ is the variance function

Inference

- $\ell(\beta) =$
- as in §10.2, $\dfrac{\partial \ell}{\partial \beta} = \left(\dfrac{\partial \eta}{\partial \beta}\right)^T u(\beta)$
- but now $\partial\eta/\partial\beta = X$ does not depend on β
- as in §10.2, $\beta = (X^T W X)^{-1} X^T W (X\beta + W^{-1} u)$
- but now $u_j =$
- and $w_j =$
- adjusted response is $X\beta + g'(\mu)(y - \mu)$
- distribution of $\hat{\beta}$?

What about ϕ_j?

- in most cases, either ϕ_j is known, or $\phi_j = \phi a_j$ where a_j known

- Normal distribution, $\phi =$

- Binomial distribution $\phi_j =$

- Gamma distribution, $\phi =$

- maximum likelihood estimate of ϕ may be poor (by analogy)

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

Glm Example: Jacamar data §10.2

	Aphrissa boisduvalli N/S/E	*Phoebis argante* N/S/E	*Dryas iulia* N/S/E	*Pierella luna* N/S/E	*Consul fabius* N/S/E	*Siproeta stelenes*† 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.

Table 10.2 Response of a rufous-tailed jacamar to individuals of seven species of palatable butterflies with artifically coloured wing undersides. (N=not sampled, S = sampled and rejected, E = eaten)

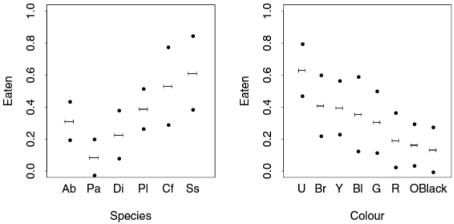


Figure 10.2 Proportion of butterflies eaten ($\pm 2SE$) for diffferent species and wing colour.

- number eaten of color c and species $s \sim$ bin(m_{cs}, π_{cs})
- model $\pi_{cs} = \frac{\exp(\alpha_c + \gamma_s)}{1 + \exp(\alpha_c + \gamma_s)}$

... jacamar data (handout)

Terms	df	Deviance
1	43	134.24
1+Species	38	114.59
1+Colour	36	108.46
1+Species+Colour	31	67.28

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

Table 10.3 Deviances and analysis of deviance for models fitted to jacamar data. The lower part of the analysis of deviance table shows results for the reduced data, without two outliers.

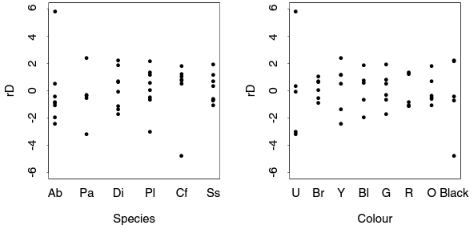
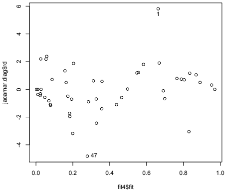


Figure 10.5 Standardized deviance residuals r_D for binomial two-way layout fitted to jacamar data.

... jacamar data

- p.485: "colour is significant at about the 0.01 level"
-
```
 > pchisq(18.03,7,lower.tail=F)
[1] 0.01183538
```
- observation 47 is an outlier; `?glm.diag` – gives deviance residuals



-
- " dropping observation 47 necessitates dropping the whole column (species)" p.485
-
```
> fit5 = glm(cbind(E,N+S) ~ colour + species, family = binomial, data = jacamar.small)
> coef(fit5)
   speciesAb    speciesPa    speciesDi    speciesPl    speciesSs
  -1.9894072   -2.2187427   -0.5596715    0.1622400    1.5018975
 colourBrown colourYellow   colourBlue  colourGreen    colourRed
   0.1588066    0.3346883   -0.5349440   -0.8330213   -1.9257494
colourOrange  colourBlack
  -1.9384921   -1.2552184
```

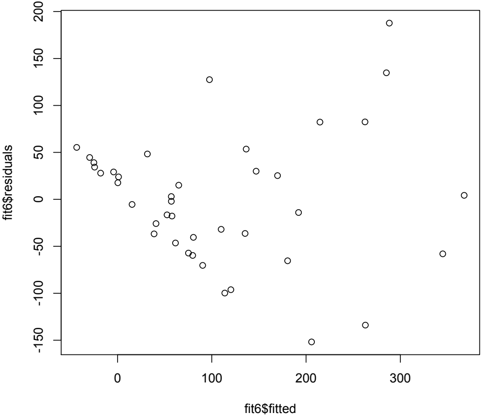
Chimp data Ex 10.16

Table 10.5 Times in minutes taken by four chimpanzees to learn ten words (Brown and Hollander, 1977, p. 257).

	Word									
Chimpanzee	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

- "when a linear model is fitted, the F-statistic for non-additivity is (8.27)" (p.485,6); (8.27) is on p.391
- linear model: $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$
- non-additivity: $y_{ij} = \mu + \alpha_i + \beta_j + \delta(\alpha_i\beta_j) + \epsilon_{ij}$
- special type of non-additivity with just 1 parameter to estimate δ

... chimp data



fit6$residuals
fit6$fitted
200
150
100
50
0
-50
-100
-150
0
100
200
300

... chimp data

- change to a model more suitable for a response that measure time
- suggestion: Gamma model with mean $\mu_{cw} = \exp(\alpha_c + \gamma_w)$
- $$f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)} y^{\nu-1} \left(\frac{\nu}{\mu}\right)^{\nu} \exp(-\nu y/\mu)$$
- $$E(y) = \mu; \qquad \mathrm{Var}(y) = \mu^2/\nu$$
- linear predictor
$$\eta_{cw} = \alpha_c + \gamma_w$$
- link function
$$\eta = \log(\mu); \qquad \mu = \exp(\eta)$$

... chimp data

486 10 · *Nonlinear Regression Models*

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

Table 10.6 Analysis of deviance for models fitted to chimpanzee data.

```
fit7 = glm(y ~ chimp + word,  family = Gamma(link = "log"), data = chimps)
> anova(fit7)
Analysis of Deviance Table

Model: Gamma, link: log

Response: y

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev
NULL                     39     60.378
chimp  3    6.948        36     53.430
word   9   38.459        27     14.972
> summary(fit7)
(Dispersion parameter for Gamma family taken to be 0.4336663)

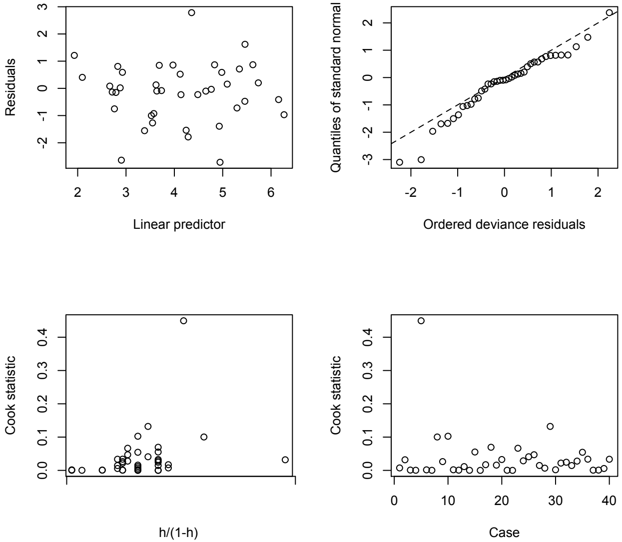
    Null deviance: 60.378  on 39  degrees of freedom
Residual deviance: 14.972  on 27  degrees of freedom
```

... chimp data

- "the signficance of the deviance reductions ... is gauged by F-tests" (p.486)
- see Eq (10.2), but note a few lines above "for now we suppress ϕ"
- see Example 10.3: $D_B - D_A = \phi^{-1} \sum\{...\} \dot{\sim} \chi^2_{p-q}$
- here we are estimating ϕ for the first time...
- p.483, 2nd paragraph: "when ϕ is unknown, the scaled deviance is replaced by the deviance"
- net result: deviance reduction for `chimp`, adjusted for `word` is 6.22 on 3 d.f.
- this is scaled by the estimate of ϕ, using (10.20), which is 0.434 from `R` code; 0.432 in text
- refer $(6.22/3)/0.433$ to $F_{3,27}$ distribution; p-value is `pf(4.788,3,27,lower.tail=F) # 0.0084`

... chimp data

`plot.glm.diag(fit7)`



... chimp data

- the canonical link is $\eta_{cw} = 1/\mu_{cw}$
- interpretation as the speed at which a word is learned
- non-additivity test for this link has p-value 0.11
- how to compare inverse link to log link?

Calcium data: Example 10.1

10.1 · Introduction

Table 10.1 Calcium uptake (nmoles/mg) of cells suspended in a solution of radioactive calcium, as a function of time suspended (minutes) (Rawlings, 1988, p. 403).

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

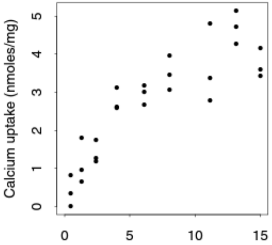


Figure 10.1 Calcium uptake (nmoles/mg) of cells suspended in a solution of radioactive calcium, as a function of time suspended (minutes).

... calcium data

- model
  $E(y_j) = \beta_0\{1 - \exp(-x_j/\beta_1)\}, \quad y_j = E(y_j) + \epsilon_j,\ \epsilon_j \sim N(0, \sigma^2)$
- fitting:

$$\min_{\beta_0,\beta_1} \sum_{j=1}^{n} (y_j - \eta_j)^2$$

- use `nls` or `nlm`; requires starting values
-
```
> library(SMPracticals); data(calcium)
> fit = nls(cal ~ b0*(1+exp(-time/b1)), data = calcium, start = list(b0=5,b1=5))
> summary(fit)
Formula: cal ~ b0 * (1 - exp(-time/b1))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
b0   4.3094     0.3029  14.226 1.73e-13 ***
b1   4.7967     0.9047   5.302 1.71e-05 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.5464 on 25 degrees of freedom

Number of iterations to convergence: 3
Achieved convergence tolerance: 9.55e-07
```

... calcium data

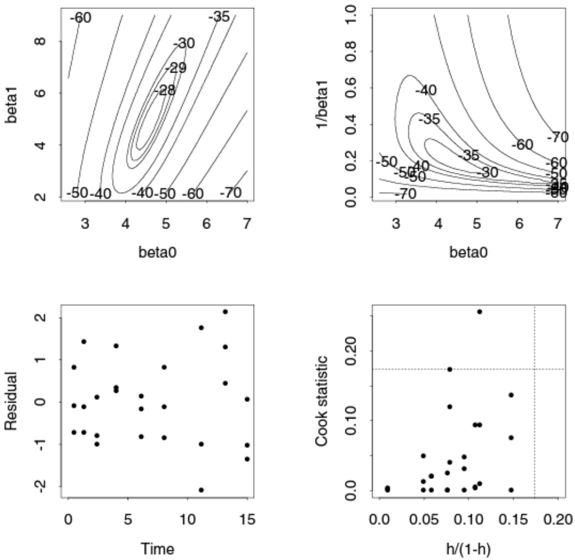


Figure 10.4 Fit of a nonlinear model to the calcium data. Upper left: contours for $\ell_p(\beta_0, \beta_1)$. Upper right: contours for $\ell_p(\beta_0, \gamma_1)$, where $\gamma_1 = 1/\beta_1$. Lower left: standardized residuals plotted against time. Lower right: plot of Cook statistics against $h/(1-h)$, where h is leverage.

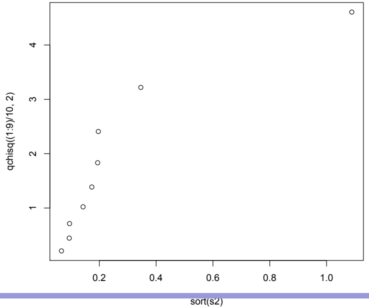
... calcium data

- there are 3 observations at each time point
- can fit a model with a different parameter for each time: $E(y_j) = \eta_j + \epsilon_j$
- the nonlinear model is nested within this; constrains η_j as above
- `anova(lm(cal ~ factor(time), data = calcium))`
- ```
  Analysis of Variance Table

  Response: cal
                Df Sum Sq Mean Sq F value    Pr(>F)
  factor(time)   8 48.437  6.0546  22.720 6.688e-08 ***
  Residuals     18  4.797  0.2665
  ```
- ```
  > deviance(fit) # 7.464514 (mistake in Davison)
  > sum(residuals(fit)^2) # 7.464514
  > (7.464514 - 4.797)/(25 - 18) # 0.3811
  > .3811/.2665
  [1] 1.429919  ## Davison has 1.53
  > > pf(1.430,7,18)
  [1] 0.7461687
  ```

... calcium data

- checking constant variance assumption
- estimates of σ^2 at each time, each with 2 degrees of freedom
- ```
  > s2 = tapply(calcium$cal, factor(calcium$time), var)
  > s2
  > s2
        0.45        1.3        2.4          4        6.1       8.05
  0.17367258 0.34616902 0.09523507 0.09422579 0.06686923 0.19656739
       11.15      13.15         15
  1.08876166 0.19415027 0.14279290
  > plot(sort(s2),qchisq((1:9)/10,2))
  ```


  ```

In the News

BONDS | FEBRUARY 7, 2012

Speaking Up Is Hard to Do: Researchers Explain Why

Article | Video | Comments (97)

Email | Print | Save | Like 636 | +1 65 | Tweet 565 | A A

Robert Murphy, an online marketing representative in San Francisco, was invited to a business meeting with his boss and six colleagues a few weeks ago. He had attended previous meetings on the subject, and he prepared with additional research. He brought a thick sheaf of notes and contracts with him to the conference room.

Ever felt like an idiot in a meeting at work or clammed up at a cocktail party? New research from Virginia Tech shows that many people are actually less intelligent in

So what did he contribute to the discussion? Absolutely nothing.

"I just sat there like a lump, fixated on the fact that I was quiet," says Mr. Murphy, 31 years old.

Have you ever clammed up at a party or found yourself tongue-tied at a meeting for fear of saying something stupid—even

http://online.wsj.com/article/SB10001424052970204136404577207020525853492.html?mod=wsj_share_tweet