

The next weeks

| | |
|----------|---|
| March 2 | §10.5 Count data and log-linear models |
| March 9 | §10.6 Overdispersion and quasi-likelihood, GEEs |
| March 16 | §10.7 Semiparametric models |
| March 23 | Generalized additive models and lasso |
| March 30 | Finishing pieces, + review |

Homework 3: due April 2, 5 pm

Final Test: April 17, 1 - 3 pm

Fitting Poisson models

- ▶ suppose we have 3 factors, each with several levels
- ▶ observe a response at each combination of factors
- ▶ linear model might be

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}, \quad k = 1, \dots, K; j = 1, \dots, J; i = 1, \dots, I$$

- ▶ or

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \epsilon_{ijk}$$

- ▶ if the y_{ijk} are positive counts, rather than continuous, then Poisson model could have

$$y_{ijk} \sim Po(\mu_{ijk}), \quad \log(\mu_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k$$

- ▶ or

$$\log(\mu_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$$

Example 10.22: Jacamar data

470

10 · Nonlinear Regression Models

| | <i>Aphrissa boisduvalli</i> N/S/E | <i>Phoebis argante</i> N/S/E | <i>Dryas iulia</i> N/S/E | <i>Pierella luna</i> N/S/E | <i>Consul fabius</i> N/S/E | <i>Siproeta stelenes</i> † 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 artificially coloured wing undersides. (N=not sampled, S = sampled and rejected, E = eaten)

- ▶ $c = 1, \dots, 8; s = 1, \dots, 6; f = 1, 2, 3$
- ▶ $y_{csf} \sim Po(\mu_{csf}); \quad \log \mu_{csf} = \mu + \alpha_c +$
- ▶ $y_{csf} \mid y_{cs+} \sim \text{Trinomial}(y_{cs+};$

... Example 10.22

▶ $y_{csf} \sim Po(\mu_{csf})$
 $\log \mu_{csf} = \mu + \alpha_c + \beta_s + \gamma_f + (\alpha\beta)_{cs} + (\alpha\gamma)_{cf} + (\beta\gamma)_{sf}$

▶ $y_{csf} \mid y_{cs+} \sim \text{Trinomial}(y_{cs+}; \frac{\mu_{csf}}{\mu_{cs+}})$

▶ $y_{cs+} \sim Po(\mu_{cs+})$

$$\mu_{cs+} =$$

▶ want to assess effect of colour and species on fate; e.g. if $(\alpha\gamma)_{cf} \equiv 0$,

▶ all models need to include $\alpha_c + \beta_s + (\alpha\beta)_{cs}$ – why?

... Example 10.22

```
> jac.long[1:6,]
  jac.y   jac.col jac.sp jac.fate
1     0 Unpainted   Ab       N
2     0 Unpainted   Ab       S
3    14 Unpainted   Ab       E
4     6 Unpainted   Pa       N
5     1 Unpainted   Pa       S
6     0 Unpainted   Pa       E
...
> jac.glm0 = glm(jac.y ~ jac.col*jac.sp + jac.fate, data = jac.long, family = poisson)
> summary(jac.glm0)
...
Null deviance: 385.55  on 143  degrees of freedom
Residual deviance: 173.86  on  94  degrees of freedom
AIC: 476.99

Number of Fisher Scoring iterations: 15
> jac.glm1 = update(jac.glm0, .~. + jac.fate:(jac.col+jac.sp))
Warning message:
glm.fit: fitted rates numerically 0 occurred
> summary(jac.glm0)
...
Null deviance: 385.554  on 143  degrees of freedom
Residual deviance:  90.659  on  70  degrees of freedom
AIC: 441.79

Number of Fisher Scoring iterations: 17
```

Example 10.23

8

1 · Introduction

| 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 |

Table 1.4 Lung cancer deaths in British male physicians (Frome, 1983). The table gives man-years at risk/number of cases of lung cancer, cross-classified by years of smoking, taken to be age minus 20 years, and number of cigarettes smoked per day.

```
> data(lung.cancer)
> lung.cancer[1:3,]
  years.smok cigarettes Time y
1    15-19           0 10366 1
2    15-19           1-9  3121 0
3    15-19          10-14  3577 0
```

... Example 10.23

```
> summary(glm(y ~ cigarettes + years.smok + offset(log(Time)),  
family = poisson, data = lung.cancer))
```

```
...
```

Coefficients:

| | Estimate | Std. Error | z value | Pr(> z) | |
|-----------------|----------|------------|---------|----------|-----|
| (Intercept) | -12.5784 | 1.1475 | -10.961 | < 2e-16 | *** |
| cigarettes1-9 | 1.2200 | 0.7073 | 1.725 | 0.084547 | . |
| cigarettes10-14 | 2.0991 | 0.6363 | 3.299 | 0.000971 | *** |
| cigarettes15-19 | 2.3089 | 0.6327 | 3.649 | 0.000263 | *** |
| cigarettes20-24 | 2.9009 | 0.5956 | 4.870 | 1.11e-06 | *** |
| cigarettes25-34 | 3.1162 | 0.5947 | 5.240 | 1.61e-07 | *** |
| cigarettes35+ | 3.6059 | 0.6048 | 5.962 | 2.49e-09 | *** |

```
...
```

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 445.099 on 62 degrees of freedom  
Residual deviance: 51.471 on 48 degrees of freedom  
AIC: 201.31
```


... log-linear models

- ▶ see also Example 10.21 for a Poisson example (y is number of goals scored in soccer match)
- ▶ with the Poisson-multinomial connection, we can also fit contingency tables with more than one response factor
- ▶ example HW 2 Q4: treat wheeze and breathlessness as responses, age as covariate
- ▶ reference Faraway: Extending the Linear Model with R ; Agresti: Analysis of Categorical Data
- ▶ skip "marginal models" (p.505) and ordinal data (§10.5.2)

Over-dispersion §10.6

- ▶ over-dispersion means $\text{Var}(Y)$ is larger than expected under the Poisson or Binomial model
- ▶ which specify $\text{Var}(Y) = \mu$, or $v(\mu) = \mu(1 - \mu)/m$
- ▶ where does over-dispersion come from? possibly multiplicative “noise”, see p. 511 for Poisson, (10.34) for Binomial
- ▶ likelihood analysis computes marginal density, averaged over noise – e.g. Poisson \rightarrow Negative Binomial (Ex. 10.26)
- ▶ alternative analysis based on “quasi-likelihood” uses analogy with least squares
- ▶ recall that if $E(Y) = X\beta$, $\text{Var}(Y) = \sigma^2 I$, then $\hat{\beta}$ is best linear unbiased estimator of β , even if Y is not normally distributed (Gauss-Markov theorem)
- ▶ there could be better nonlinear estimators of β

... overdispersion

- ▶ if $E(Y) = X\beta$ and $\text{Var}(Y) = V$, then $\hat{\beta} =$
unbiased for β
- ▶ $\text{Var}(\hat{\beta}) =$ (8.19)
- ▶ if we knew V , replace $\hat{\beta}$ by weighted least squares estimator; otherwise, use $\hat{\beta}$ and adjust confidence intervals by some estimate of V , see p.377

... overdispersion

- ▶ estimation of β in a generalized linear model depends only on the specification of the mean function
- ▶ and the variance function
- ▶ suggests using the same estimating equation for β , but allow inflation of the variance function by an unknown dispersion parameter
- ▶ e.g. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\mu_j \quad -$
- ▶ e.g. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\pi_j(1 - \pi_j)/m \quad -$
- ▶ estimating equation for β is unchanged

... overdispersion



$$\sum_{j=1}^n x_j \frac{y_j - \mu_j}{g'(\mu_j) V(\mu_j)} = 0$$

- ▶ this is an unbiased estimating function $g(y; \beta)$; satisfies $E\{g(Y; \beta)\} = 0$
- ▶ under some regularity conditions the solution of $g(y; \beta) = 0$ is consistent, asymptotically normal
- ▶ a. $\text{Var}(\tilde{\beta}) = \phi(X^T W X)^{-1}$
- ▶ from general theory on unbiased estimating functions

$$E \left\{ -\frac{\partial g(Y; \beta)}{\partial \beta} \right\}^{-1} \text{Var}\{g(Y; \beta)\} E \left\{ -\frac{\partial g(Y; \beta)}{\partial \beta} \right\}^{-1}$$

... inference

- ▶ $\hat{\phi} = \frac{1}{n-p} \sum_{j=1}^n \sum_{j=1}^n \frac{(y_j - \hat{\mu}_j)^2}{V(\hat{\mu}_j)}$
- ▶ $\widehat{\text{Var}}(\tilde{\beta}_j) = \hat{\phi} \widehat{\text{Var}}(\hat{\beta}_j)$
- ▶ comparison of models: $A \subset B \quad D_A - D_B \sim \chi_{p_B - p_A}^2$
- ▶ changes to

Example 10.28

Set 2

Table 2 (Bissell, 1972) gives the numbers of faults in rolls of textile fabric. The distribution of number of faults is of interest, especially in its relation to that expected if faults occur at random at a fixed rate per metre.

Table 2. Numbers of faults in rolls of textile fabric

| Roll No. | Roll length (metres) | No. of faults | Roll No. | Roll length (metres) | No. of faults |
|----------|----------------------|---------------|----------|----------------------|---------------|
| 1 | 551 | 6 | 17 | 543 | 8 |
| 2 | 651 | 4 | 18 | 842 | 9 |
| 3 | 832 | 17 | 19 | 905 | 23 |
| 4 | 375 | 9 | 20 | 542 | 9 |
| 5 | 715 | 14 | 21 | 522 | 6 |
| 6 | 868 | 8 | 22 | 122 | 1 |
| 7 | 271 | 5 | 23 | 657 | 9 |
| 8 | 630 | 7 | 24 | 170 | 4 |
| 9 | 491 | 7 | 25 | 738 | 9 |
| 10 | 372 | 7 | 26 | 371 | 14 |
| 11 | 645 | 6 | 27 | 735 | 17 |
| 12 | 441 | 8 | 28 | 749 | 10 |
| 13 | 895 | 28 | 29 | 495 | 7 |
| 14 | 458 | 4 | 30 | 716 | 3 |
| 15 | 642 | 10 | 31 | 952 | 9 |
| 16 | 492 | 4 | 32 | 417 | 2 |

... example 10.28

```
> data(cloth)
> cloth[1:5,]
  x y
1 1.22 1
2 1.70 4
3 2.71 5
4 3.71 14
5 3.72 7
> with(cloth,plot(x,y)) # gives Fig 10.11
> cloth.glm0 = glm(y ~ x - 1, family = poisson(link = identity), data = cloth)
> summary(cloth.glm0)
Coefficients:
  Estimate Std. Error z value Pr(>|z|)
x 1.51024    0.08962    16.85  <2e-16 ***
---

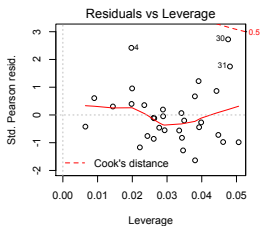
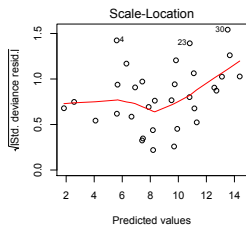
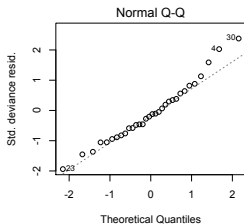
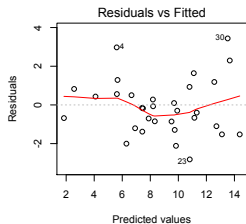
(Dispersion parameter for poisson family taken to be 1)

Null deviance:    Inf on 32 degrees of freedom
Residual deviance: 64.537 on 31 degrees of freedom
> cloth.glm1 = glm(y ~ x - 1, family = quasipoisson(link = identity), data = cloth)
> summary(cloth.glm1)
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x 1.5102    0.1328    11.38 1.35e-12 ***
---

(Dispersion parameter for quasipoisson family taken to be 2.194371)
```

negative binomial p. 515

Quasi-Poisson model fit



Example 10.29

516

10 - Nonlinear Regression Models

| City | Rain | r/m | City | Rain | r/m | City | Rain | r/m | City | Rain | r/m |
|------|------|-------|------|------|-------|------|------|-------|------|------|-------|
| 1 | 1735 | 2/4 | 11 | 2050 | 7/24 | 21 | 1756 | 2/12 | 31 | 1780 | 8/13 |
| 2 | 1936 | 3/10 | 12 | 1830 | 0/1 | 22 | 1650 | 0/1 | 32 | 1900 | 3/10 |
| 3 | 2000 | 1/5 | 13 | 1650 | 15/30 | 23 | 2250 | 8/11 | 33 | 1976 | 1/6 |
| 4 | 1973 | 3/10 | 14 | 2200 | 4/22 | 24 | 1796 | 41/77 | 34 | 2292 | 23/37 |
| 5 | 1750 | 2/2 | 15 | 2000 | 0/1 | 25 | 1890 | 24/51 | | | |
| 6 | 1800 | 3/5 | 16 | 1770 | 6/11 | 26 | 1871 | 7/16 | | | |
| 7 | 1750 | 2/8 | 17 | 1920 | 0/1 | 27 | 2063 | 46/82 | | | |
| 8 | 2077 | 7/19 | 18 | 1770 | 33/54 | 28 | 2100 | 9/13 | | | |
| 9 | 1920 | 3/6 | 19 | 2240 | 4/9 | 29 | 1918 | 23/43 | | | |
| 10 | 1800 | 8/10 | 20 | 1620 | 5/18 | 30 | 1834 | 53/75 | | | |

Table 10.19

Toxoplasmosis data: rainfall (mm) and the numbers of people testing positive for toxoplasmosis, r , out of m people tested, for 34 cities in El Salvador (Efron, 1986).

| Terms | df | Deviance |
|-----------|----|----------|
| Constant | 33 | 74.21 |
| Linear | 32 | 74.09 |
| Quadratic | 31 | 74.09 |
| Cubic | 30 | 62.63 |

Table 10.20 Analysis of deviance for polynomial logistic models fitted to the toxoplasmosis data.

- ▶ incidence of toxoplasmosis as a function of rainfall
- ▶ residual deviances approximately twice the degrees of freedom

... example 10.29

```
> data(toxo)
> toxo[1:4,]
  rain m r
1 1620 18 5
2 1650 30 15
3 1650 1 0
4 1735 4 2
> toxo.glm0 = glm(cbind(r,m-r) ~ rain + I(rain^2) + I(rain^3), data = toxo,
family = binomial)
> anova(toxo.glm0)
...
      Df Deviance Resid. Df Resid. Dev
NULL                33      74.212
rain                1    0.1244
I(rain^2)           1    0.0000
I(rain^3)           1   11.4529
> toxo.glm1 = glm(cbind(r,m-r) ~ rain + I(rain^2) + I(rain^3), data = toxo,
family = quasibinomial)
> summary(toxo.glm1)
...
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.902e+02  1.215e+02  -2.388  0.0234 *
rain         4.500e-01  1.876e-01   2.398  0.0229 *
I(rain^2)    -2.311e-04  9.616e-05  -2.404  0.0226 *
I(rain^3)     3.932e-08  1.635e-08   2.405  0.0225 *

> 74.212 - 62.635
[1] 11.577
> pchisq(.Last.value, 3, lower.tail = F)
[1] 0.008982002
> (74.212 - 62.635) / (3 * 0.008982002)
[1] 4.19175
```

Estimating functions and quasi-likelihood

- ▶ suppose we assume only that $E(Y_j) = \mu_j(\beta)$, $\text{Var}(Y_j) = \phi_j V(\mu_j)$, as in glm
- ▶ and we use the glm estimates of β , defined by the score equation

$$\sum_{j=1}^n \frac{y_j - \mu_j}{\phi_j V(\mu_j)} \frac{x_{jr}}{g'(\mu_j)} = 0$$

- ▶ n.b. Davison calls LHS $g(Y; \beta)$, different g
- ▶ using only (*), we have

$$E\{g(Y; \beta)\} = 0; \quad E\left\{-\frac{\partial g(Y; \beta)}{\partial \beta}\right\} = \text{Var}\{g(Y; \beta)\}$$

- ▶ thus g has two properties in common with the the score function from a log-likelihood

... estimating functions and quasi-likelihood



$$g(Y; \beta) = \sum_{j=1}^n \frac{y_j - \mu_j}{\phi_j V(\mu_j)} \frac{x_{jr}}{g'(\mu_j)} = 0$$

- ▶ leads to $\tilde{\beta} \sim N(\beta, X^T W X)$, where $W = \text{diag}(w_j)$
- ▶ $w_j = 1 / \{g'(\mu_j)^2 \phi_j V(\mu_j)\}$
- ▶ as usual, assume $\phi_j = \phi a_j$
- ▶ example: weighted LS – $V(\mu) = 1$, $weight_j = 1/a_j$
- ▶ example: constant coefficient of variation: $V(\mu) = \mu^2$
- ▶ example: overdispersed binomial or Poisson

... estimating functions and quasi-likelihood

- ▶ even if $V(\mu)$ is incorrectly specified, $\tilde{\beta}$ is still consistent



$$\text{a. Var}(\tilde{\beta}) = (X^T W X)^{-1} \text{Var}\{g(Y; \beta)\} (X^T W X)^{-1}$$

- ▶ often is well approximated by $(X^T W X)^{-1}$ in any case
- ▶ when extended to dependent data, called generalized estimating equation method
- ▶ assume $E(Y) = \mu(\beta)$; $\text{Var}(Y) = V(\mu)$



$$X^T V^{-1}(\mu)(Y - \mu) = 0$$

- ▶ in the dependent data case, there is no quasi-likelihood function for which (*) is the score function