Over-dispersion §10.6

- over-dispersion means Var(Y) is larger than expected under the Poisson or Binomial model
- which specify 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 $E(Y) = X\beta$, 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 $\mathrm{Var}(Y) = V$, then $\hat{\beta} = (X^TX)^{-1}X^Ty$ unbiased for β

$\mathrm{Var}(\hat{\beta}) = (X^TX)^{-1}(X^TVX)(X^TX)^{-1}$ "larger" than best

- $\mathrm{Var}(\hat{\beta}) =$ (8.10)

- 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

R will give $\hat{\sigma}^2(X^TX)^{-1}$ ("too small")

w $\hat{\sigma}^2 = SS_{res}/n-p$

"best" $\hat{\beta}_V = (X^TV^{-1}X)^{-1}(X^TV^{-1}y)$ (n+bc)

... 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

- if $E(Y) = X\beta$ and $\mathrm{Var}(Y) = V$, then $\hat{\beta} = (X^T X)^{-1} X^T y$ unbiased for β
- $\mathrm{Var}(\hat{\beta}) = (X^T X)^{-1} X^T V X (X^T X)^{-1}$ (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

OLS

... 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. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\mu_j$ –
- e.g. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\pi_j(1-\pi_j)/m$ –
- estimating equation for β is unchanged

... 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. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\mu_j$ –
- e.g. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\pi_j(1-\pi_j)/m$ –
- estimating equation for β is unchanged

... 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. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\mu_j$ –
- e.g. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\pi_j(1 - \pi_j)/m$ –
- estimating equation for β is unchanged

... 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. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\mu_j$ – quasi-Poisson
- e.g. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\pi_j(1-\pi_j)/m$ –
- estimating equation for β is unchanged

... 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. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\mu_j$ –
- e.g. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\pi_j(1 - \pi_j)/m$ – quasi binomial
- estimating equation for β is unchanged

... 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. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\mu_j$ –
- e.g. $E(y_j) = \mu_j, \quad \text{Var}(y_j) = \phi\pi_j(1 - \pi_j)/m$ –
- 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. $\text{Var}(\tilde{\beta}) = \phi(X^T \tilde{W} X)^{-1}$

now $\tilde{W}_j$ has $\phi_j = 1$ by ass

- from general theory on unbiased estimating functions

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

Example 10.29

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 Toxoplamosis data: rainfall (mm) and the numbers of people testing positive for toxoplasmosis, r, our 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

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 Toxoplamosis data: rainfall (mm) and the numbers of people testing positive for toxoplasmosis, r, our 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)
  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        32     74.087
I(rain^2)  1   0.0000        31     74.087
I(rain^3)  1  11.4529        30     62.635
> toxo.glm1 = glm(cbind(r,m-r) ~ poly(rain,3), data = toxo, family = binomial)

> summary(toxo.glm1)
...
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              0.02427    0.07693   0.315 0.752401
poly(rain, degree = 3)1 -0.08606    0.45870  -0.188 0.851172
poly(rain, degree = 3)2 -0.19269    0.46739  -0.412 0.680141
poly(rain, degree = 3)3  1.37875    0.41150   3.351 0.000806 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.212  on 33  degrees of freedom
Residual deviance: 62.635  on 30  degrees of freedom
```

... example 10.29

```
> toxo.quasi2 <- glm(cbind(r,m-r) ~ rain +I(rain^2)+I(rain^3),
+ data = toxo, family = quasibinomial)

> summary(toxo.quasi2)

Call:
glm(formula = cbind(r, m - r) ~ rain + I(rain^2) + I(rain^3),
    family = quasibinomial, data = toxo)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7620  -1.2166  -0.5079   0.3538   2.6204

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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.940446)

    Null deviance: 74.212  on 33  degrees of freedom
Residual deviance: 62.635  on 30  degrees of freedom
> (74.212-62.635)/3/1.940446
[1] 1.988718
> pf(1.988718,3,30,lower=F)
[1] 0.1368842
```