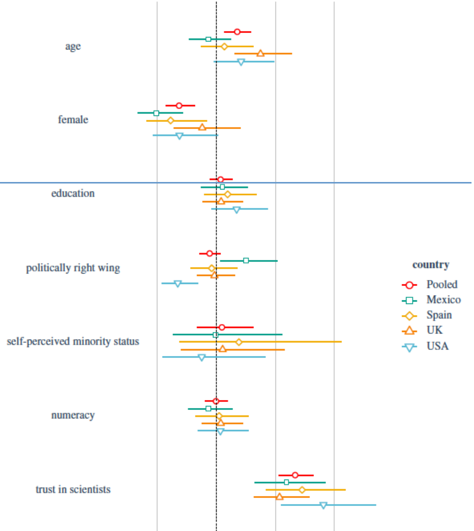
Methods of Applied Statistics I

STA2101H F LEC9101

Week 7

October 22 2020



Today

Start Recording

1. In the News
2. HW2 revised due date November 5
3. Explanation FLM-2 §5.4-5.7
4. Theory of logistic regression
5. Examples of logistic regression
6. Introduction to `tidyverse`

Syllabus update 3

- October 26 3.00 – 4.00 Kristian Lum
- `https://canssiontario.utoronto.ca/?mec-events=ares_lum_kristian`

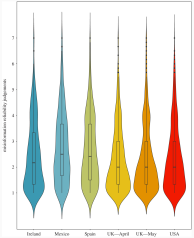
"Fairness, Accountability, and Transparency:
(Counter)-Examples from Predictive Models
in Criminal Justice"
Assistant Professor, CIS, U Penn
Previously, Lead Statistician at the Human Rights
Data Analysis Group (HRDAG)

Recap

- Preliminary Analysis: data auditing, data screening, data cleaning, preliminary summaries (tables, plots) CD Ch.5
- Explanation after linear regression FLM-2 Ch.5

- Analysis of binomial data – Challenger shuttle
- Logistic regression model $p_i = \exp(x_i^T \beta)/\{1 + \exp(x_i^T \beta)\}$
- Fitting the model with `glm`
- Maximum likelihood estimation under the hood

- Covid misinformation paper



?Logistic regression: what if it is unbalanced?

- model for logistic regression $y_i \sim Bin(n_i, p_i), \quad i = 1, \ldots, n$
- $\mathbb{E}(y_i) = n_i p_i, \quad \mathrm{Var}(y_i) = n_i p_i (1 - p_i)$
- lack of balance: e.g. $n_1 = 10, n_2 = 1000$

- estimates $\hat{p}_1 = y_1/n_1$, $\hat{p}_2 = y_2/n_2$; also $\mathrm{var}(\hat{p}_i) = p_i(1 - p_i)/n_i$
- if for example $p_1 = p_2 = 0.5$,

$$\mathrm{var}(\hat{p}_1) = 0.025, \quad \mathrm{var}(\hat{p}_2) = 0.00025$$

$$\mathrm{s.e.}(\hat{p}_1) = 0.158, \quad \mathrm{s.e.}(\hat{p}_2) = 0.0158$$

- precise information about some groups of individuals, and less precise information about others
- suggests that estimates for those covariates may have large standard errors, simply due to sample size issues

Explanation

- Note: Chapter uploaded to Quercus page, under ”Modules”
- §5.1-3: interpretation of coefficients, causal effects, designed experiments
- §5.4 observational data
    - difficult to infer causality from observational data
    - treatment not assigned at random
    - treated group could differ from untreated group in many different ways
    - in the voting example “voting system” is the “treatment” NH primary 2008
    - it appears to influence the outcome (proportion voting for Obama)
    - but Faraway uncovered a potential confounder: outcome of 2004 primary (Dean)
- §5.5. matching
    - create blocks (pairs) and “assign treatment/control” to each unit in the pair

      thought experiment
    - Faraway uses an algorithm to create pairs of wards that are similar – except that 1 ward was ‘treated’, the other was ‘control’
    - this is called propensity score matching in causal inference

On p. 70, just before exercises, Faraway mentions a “natural experiment”


STUDIES SHOW

How an Ill-Fated Fishing Voyage Helped Us Understand Covid-19

NY Times October 20

Explanation

- with observational data, we usually adjust for confounders in a regression model
- but we can never be completely sure there isn't an unmeasured confounder
- so it is nearly impossible to conclude causality from an observational study
- so how do we know that smoking causes lung cancer?
- §5.7 "Bradford-Hill criteria"
  - strength of the observed association
  - consistency of the observed association
  - specificity of the potential cause
  - the potential cause occurs earlier in time than the outcome
  - there is a dose-response relationship
  - there is subject-matter theory that makes a causal effect plausible
  - there is corroborating evidence from other types of studies (e.g. animal studies)

Four cardinal rules of statistics

- ONE: Correlation does not imply causation.

  Unless you can design your study to uncover causation, the best you can do is discover correlations

- TWO: A *p*-value is just a test of sample size. I don't agree!

  In other words, we can have STATISTICAL significance w/o PRACTICAL significance.... In many contemporary settings, sample sizes are so huge that we can get TINY p-values even when the deviation from the null hypothesis is negligible. *I do agree*

- THREE: Seek and ye shall find.

  If you look at your data for long enough, you will find something interesting, even if only by chance!

Regression modelling with binomial

- model:

$$y_i \sim Bin(n_i, p_i)$$

$n_i = 6, i = 1, \ldots, n$

- regression: link the p_i's through x_i
- for example,

$$p_i = \frac{\exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q)}{1 + \exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q))}$$

- more concisely

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

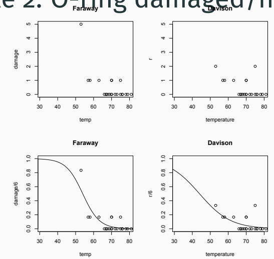
- $x_i^{\mathrm{T}} = (1, x_{i1}, \ldots, x_{iq}); \quad \beta = (\beta_0, \beta_1, \ldots, \beta_q)^{\mathrm{T}}$

all vectors are column vectors

- many examples where we would like to analyse a binary response $y_i = 0/1$
- example from last week: covid misinformation

> *To investigate the effects of susceptibility to misinformation about COVID-19 on people's willingness to (i) get vaccinated against COVID-19 (yes/no), and (ii) recommend getting vaccinated to vulnerable friends or family members (yes/no), we conducted two logistic regressions*

- example 2: O-ring damaged/not damaged



NB zeroes

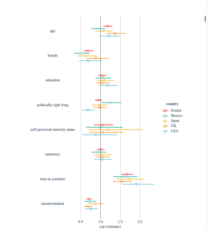


Figure 3. Logistic regression model for 'Would you get vaccinated against COVID-19 (y/n)' by country. *Note*: coefficients are exponentiated and represent odds ratios. Left of the dotted line (i.e. values less than 1) indicates *reduced* likelihood to get vaccinated.

Table 10.8 Data on nodal involvement (Brown, 1980).

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	1	1	0	1	1
1	1	1	0	1	1	1
1	1	1	0	0	1	1
1	0	1	0	1	0	0
1	1	0	1	1	1	0
1	0	0	1	1	0	0
1	1	0	1	0	1	0
1	1	0	0	1	0	1

Can we predict nodal involvment from other measurements?

... Binary responses

- suppose y_i is binary — takes values 0, 1 only
- and there are several covariates x_i associated with *i*th observation
- ?what's wrong with

$$y_i = x_i^{\mathrm{T}} \beta + \epsilon_i$$

- what's the probability distribution of y_i? — Bernoulli
- the only parameter in the distirbution is $p_i = \Pr(y_i = 1)$ — $1 - p_i = ?$
- suppose $y_1, \ldots, y_n$ are independent Bernoulli — $(y_1, \ldots, y_n) = \underline{y}$
- joint distribution

$$f(\underline{y}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{(1-y_i)}$$

… Binary responses

- joint distribution

$$f(\underline{y}) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

- log-likelihood function

$$\ell(p; y) = \sum_{i=1}^{n}\{y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\} = \sum_{i=1}^{n}\{y_i \log\{p_i/(1 - p_i)\} + \log(1 - p_i)\}$$

- logistic regression $\log\{p_i/(1 - p_i)\} = x_i^{\mathrm{T}}\beta$
- log-likelihood function

$$\ell(\beta; y) = \sum_{i=1}^{n}[y_i x_i^{\mathrm{T}}\beta - \log\{1 + \exp(x_i^{\mathrm{T}}\beta)\}]$$

… Binary responses

- where's the epsilon? There isn't one
- what's the model? It has two parts
- Regression.

$$\mathbb{E}(y_i) = p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1+\exp(x_i^{\mathrm{T}}\beta)}$$

- Probability distribution.

$$y_i \sim Bernoulli(p_i)$$

- What are these parts in linear regression?
- Regression

$$\mathbb{E}(y_i) = \mu_i = x_i^{\mathrm{T}}\beta$$

- Probability distribution

$$y_i \sim Normal(\mu_i, \sigma^2)$$

Binomial responses

- if you add a lot of Bernoulli's together, all with the same p_i, you get
- how could they have the same p_i in our model?
- $p_i = function(x_i^{\mathrm{T}}\beta)$
- different observations with the same p_i are called covariate classes
- Example 10.18 in SM – Table 10.8 has 23 rows of binomials

  sample sizes vary from 1 to 6
- `data(nodal)` in `library(SMPracticals)` has 53 rows of binary observations
- `R` expects `cbind(r, m-r)` in `glm` with binomial data, but if all observations are binary you can get away with `r` only
- see `?family` (check `Details`)
- you can also specify proportions y_i/n_i, but then you need to use `weights`

Review: Likelihood inference

- model: $y_i \sim f(y_i; \theta), i = 1, \ldots, n$ independent
- joint density: $f(\underline{y}; \theta) = \prod_{i=1}^{n} f(y_i; \theta)$
- likelihood function $L(\theta; \underline{y}) = f(\underline{y}; \theta)$

- log-likelihood function $\ell(\theta; \underline{y}) = \log L(\theta; \underline{y}) = \sum_{i=1}^{n} \log f(y_i; \theta)$
- maximum likelihood estimate $\hat{\theta} = \arg\sup \ell(\theta; \underline{y})$ $\ell'(\hat{\theta}) = 0$
- Fisher information $j(\theta) = -\ell''(\theta)$

- properties:

$$(\hat{\theta} - \theta) j^{1/2}(\hat{\theta}) \xrightarrow{d} N(0, I)$$

asymptotically normal

- likelihood ratio statistic

$$w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_p$$

p is dimension of θ

... Likelihood inference

- properties:

$$(\hat{\theta} - \theta) j^{1/2}(\hat{\theta}) \xrightarrow{d} N(0, I)$$

asymptotically normal

-

$$\hat{\theta}_k \dot{\sim} N(\{\theta_k, j^{-1}(\hat{\theta})_{kk}\}$$

`vcov(logitmodel)`

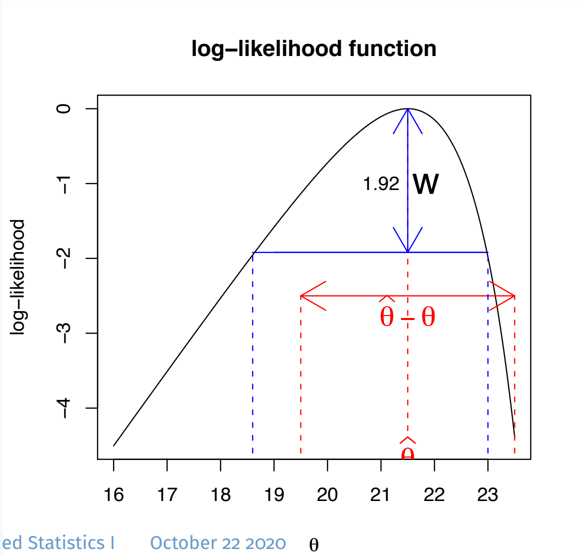
- likelihood ratio statistic

$$w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_p$$

p is dimension of θ

- compare two models using change in likelihood ratio statistic

... Likelihood inference



should be $w/2$

- fit model A get estimate $\hat{\theta}_A$
- fit model B get estimate $\hat{\theta}_B$

Model B smaller than A

- likelihood ratio test

$$LRT = 2\{\ell_A(\hat{\theta}_A) - \ell_B(\hat{\theta}_B)\}$$

- compares the maximized log-likelihood function under model A and model B
- example
  model A: $logit(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad \theta_A = (\beta_0, \beta_1, \beta_2)$
  model B: $logit(p_i) = \beta_0 + \beta_1 x_{1i}, \quad \theta_B = (\beta_0, \beta_1)$

- when model B is nested in model A, LRT is approximately χ^2_ν distributed

$$\nu = dim(A) - dim(B)$$

... Challenger data

```
> head(shuttle2)
  m r temperature pressure
1 6 0          66       50
2 6 1          70       50
3 6 0          69       50
4 6 0          68       50
5 6 0          67       50
6 6 0          72       50
> logitmodcorrect2 <- glm(cbind(r,m-r) ~ temperature + pressure, family = binomial, data = shuttle2)
> summary(logitmodcorrect2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.520195   3.486784   0.723   0.4698
temperature -0.098297   0.044890  -2.190   0.0285 *
pressure     0.008484   0.007677   1.105   0.2691
---
```

$\hat{\beta}_1 \dot\sim N(0, 0.044^2)$

$\hat{\beta}_2 \dot\sim N(0, 0.008^2)$

... Challenger data

- Model A: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$
- Model B: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i$
- nested: Model B is obtained by setting $\beta_2 = 0$
- the change in deviance is a likelihood ratio test

```
> anova(logitmodcorrect,logitmodcorrect2)
     Analysis of Deviance Table

Model 1: cbind(r, m - r) ~ temperature
Model 2: cbind(r, m - r) ~ temperature + pressure
       Resid. Df Resid. Dev Df Deviance
1         21      18.086
2         20      16.546  1   1.5407
> 1 - pchisq(1.5407, df = 1).    0.214
```

... inference

- Model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 \mathtt{temp}_i + \beta_2 \mathtt{pressure}_i$
- Model B: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 \mathtt{temp}_i$
- nested: Model B is obtained by setting $\beta_2 = 0$
- Under Model B, the change in deviance is (approximately) an observation from a χ_1^2
- $\Pr(\chi_1^2 \geq 1.5407) = 0.22$
  this is a p-value for testing $H_0 : \beta_2 = 0$
- so is $1 - \Phi\{\frac{\hat{\beta}_2}{\widehat{s.e.}(\hat{\beta}_2)}\} = 1 - \Phi(1.105) = 0.27$

ELM p.30

```
> summary(logitmodcorrect)

Call:
glm(formula = cbind(r, m - r) ~ temperature, family = binomial,
    data = shuttle2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.95227  -0.78299  -0.54117  -0.04379   2.65152

Coefficients:
...
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.230  on 22  degrees of freedom
Residual deviance: 18.086  on 21  degrees of freedom
AIC: 35.647
```



```
Number of Fisher Scoring iterations: 5
```

Special to the binomial: residual deviance and Poisson

- the logistic regression model $p_i = p_i(\beta) = \exp(x_i^{\mathrm{T}}\beta)/\{1 + \exp(x_i^{\mathrm{T}}\beta)\}, \quad \hat{p}_i = p_i(\hat{\beta})$
- is nested in the saturated model $\tilde{p}_i = y_i/n_i$

- the saturated model has one estimate of p_i for each row of the data
- residual deviance compares the regression model to the saturated model

- under the fitted model, approximately distributed as χ^2_{n-q}
  if each n_i "large" ELM p.29
- this is LRT of the regression model compared to the saturated model

```
> summary(logitmodcorrect)
  ...
   Null deviance: 24.230  on 22  degrees of freedom
   Residual deviance: 18.086  on 21  degrees of freedom
    AIC: 35.647
   Number of Fisher Scoring iterations: 5
```

Residual deviance

- Residual Deviance is log-likelihood ratio statistic for the fitted model compared to the saturated model
- saturated model is maximized at $\tilde{p}_i = y_i/n_i$

$$\ell(\tilde{p}) = \sum_{i=1}^{n} \{y_i \log(y_i/n_i) + (n_i - y_i) \log(1 - y_i/n_i)\}$$

- fitted model maximized at $\hat{\beta}$

$$\ell(\hat{\beta}) = \sum_{i=1}^{n} \{y_i \log p_i(\hat{\beta}) + (n_i - y_i) \log(1 - p_i(\hat{\beta}))\}$$

- twice the difference:

$$2 \sum_{i=1}^{n} [y_i \log\{y_i/n_i p_i(\hat{\beta})\} + (n_i - y_i) \log\{(n_i - y_i)/(n_i - n_i p_i(\hat{\beta}))\}]$$

FELM Eq.(2.1): $\hat{y}_i = n_i p_i(\hat{\beta})$

- If data is distributed as Binomial
- and each n_i is "large" >5
- Residual deviance is a test of goodness of fit of the model
- A happy quirk of logistic regression

- interpretation of parameters in terms of log odds §2.5

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
temperature -0.11560    0.04702  -2.458   0.0140 *
```

"a unit increase in temperature is associated with a decrease in log-odds of O-ring failure of 0.116"

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
temperature -0.11560    0.04702  -2.458   0.0140 *
```

"a unit increase in temperature is associated with an increase in log-odds of O-ring damage of -0.116"

"an increase in the odds of $\exp(-0.116) = 0.89$" so actually a decrease

" an increase in the probability of ?? depends on the baseline probability

go to rsos.201199.pdf

aggregated data presented in textbook

Table 10.8 Data on nodal involvement (Brown, 1980).

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	1	1	0	1	1
1	1	1	0	1	1	1
1	1	1	0	0	1	1
1	0	1	0	1	0	0
1	1	0	1	1	1	0
1	0	0	1	1	0	0

… example 10.18

- `library(SMPracticals); data(nodal); head(nodal)` all covariates 0/1
- several patients have the same value of the covariates

  covariate classes: ELM
- these can be added up to make a binomial observation
- ```
  > nodal2[1:4,]
    m r age stage grade xray acid
  1 6 5   0     1     1    1    1
  2 6 1   0     0     0    0    1
  3 4 0   1     1     1    0    0
  4 4 2   1     1     0    0    1
  ```
- ```
  > ex1018binom = glm(cbind(r,m-r) ~ ., data = nodal2, family = binomial)
  > summary(ex1018binom)  # stuff omitted
  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -3.0794     0.9868  -3.121  0.00180 **
  age          -0.2917     0.7540  -0.387  0.69881
  stage         1.3729     0.7838   1.752  0.07986 .
  grade         0.8720     0.8156   1.069  0.28500
  xray          1.8008     0.8104   2.222  0.02628 *
  acid          1.6839     0.7915   2.128  0.03337 *
  ---

  (Dispersion parameter for binomial family taken to be 1)

      Null deviance: 40.710  on 22  degrees of freedom
  Residual deviance: 18.069  on 17  degrees of freedom
  AIC: 41.693

  Number of Fisher Scoring iterations: 5
  ```

... example 10.18 variable selection

```
> step(ex1018binom)

Coefficients:
(Intercept)        stage         xray         acid
     -3.052        1.645        1.912        1.638

Degrees of Freedom: 22 Total (i.e. Null);  19 Residual
Null Deviance:     40.71
Residual Deviance: 19.64  AIC: 39.26
```

– we can drop `age` and `grade` without affecting quality of the fit

– in other words the model can be simplified by setting two regression coefficients to zero

– several mistakes in text on pp. 491,2;

– deviances in Table 10.9 are incorrect as well `http://statwww.epfl.ch/davison/SM/` has corrected version

... example 10.18: variable selection

- `step` implements stepwise regression
- evaluates each fit using AIC = $-2\ell(\hat{\beta}; y) + 2p$
- penalizes models with larger number of parameters

- we can also compare fits by comparing deviances
-

```
> update(ex1018binom, . ~ . - aged - stage)

Call:  glm(formula = cbind(r, m - r) ~ grade + xray + acid, family = binomial,
    data = nodal2)

Coefficients:
(Intercept)        grade         xray         acid
     -2.734        1.420        1.750        1.797

Degrees of Freedom: 22 Total (i.e. Null);  19 Residual
Null Deviance:      40.71
Residual Deviance: 21.28   AIC: 40.9

> deviance(ex1018binom)
[1] 18.06869
> pchisq(21.28-18.07,df=2,lower=F)
[1] 0.2008896
```

AIC

- as terms are added to the model, deviance always decreases
- because log-likelihood function always increases
- similar to residual sum of squares

- Akaike Information Criterion penalizes models with more parameters
- $$AIC = 2\{-\ell(\hat{\beta}; y) + p\}$$

SM (4.57)

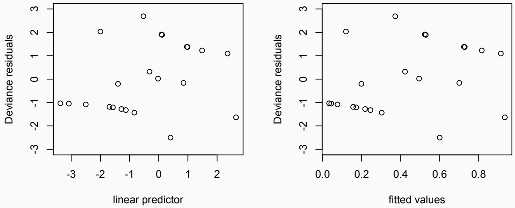
- comparison of two model fits by difference in AIC

... example 10.18

```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```



Deviance residuals


```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```

Deviance: $2\sum_{i=1}^{n}[y_i \log\{y_i/n_i p_i(\hat{\beta})\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i p_i(\hat{\beta}))\}]$

approximately χ^2_{n-q}

$$r_{Di} = \pm\sqrt{}(2[y_i \log\{y_i/n_i\hat{p}_i\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i\hat{p}_i)\}])$$

... example 10.18: residuals

```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```