MONASH University

MONASH BUSINESS SCHOOL

ETC3580: Advanced Statistical Modelling

Week 4: Binomial and proportion responses

Outline

Binomial responses

Binomial distribution

Y is binomially distributed $B(m, p)$ if

$$P(Y = y) = \binom{m}{y} p^y (1 - p)^{m-y}$$

Y = number of "successes" in m independent trials, each with probability p of success.

- Likelihood

$$L = \prod_{i=1}^{n} \binom{m_i}{y_i} p_i^{y_i} (1 - p_i)^{m_i - y_i}$$

- Binomial regression also uses a logit link

$$p_i = e^{\eta_i} / (1 + e^{\eta_i})$$

$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_q x_{i,q}$$

Binomial likelihood

Likelihood

$$L = \prod_{i=1}^{n} \binom{m_i}{y_i} p_i^{y_i} (1 - p_i)^{m_i - y_i}$$

$$\begin{aligned}
\log L(\beta) &= \sum_{i=1}^{n} \left[\log \binom{m_i}{y_i} + y_i \log(p_i) + (m_i - y_i) \log(1 - p_i) \right] \\
&= \sum_{i=1}^{n} \left[\log \binom{m_i}{y_i} + y_i \eta_i - y_i \log(1 + e^{\eta_i}) - (m_i - y_i) \log(1 + e^{\eta_i}) \right] \\
&= \sum_{i=1}^{n} \left[\log \binom{m_i}{y_i} + y_i \eta_i - m_i \log(1 + e^{\eta_i}) \right]
\end{aligned}$$

Binomial responses

In R:

- `glm` needs a two-column matrix of success and failures. (So rows sum to *m*).

```
fit <- glm(cbind(successes,failures) ~
    x1 + x2,
  family=binomial, data=df)
```

- Everything else works the same as for binary regression.

Overdisperson

- If mean correctly modelled, but observed variance larger than model, we called the data "overdispersed". [Same for underdispersion.]
- Concept of overdispersion irrelevant for OLS and logistic regression because there cannot be any more variance than what is modelled.
- For binomial regression: $y_i \sim B(m_i, p_i)$, $\mathrm{E}(y_i) = m_i p_i$, $\mathrm{Var}(y_i) = m_i p_i (1 - p_i)$.
- If model correct, $D = -2 \log L \sim \chi^2_{n-q}$. So $D > n - q$ indicates overdisperson.

Overdisperson

$D > n - q$ can also be the result of:

- missing covariates or interaction terms
- negligence of non-linear effects
- large outliers
- sampling from clusters
- non-independence
- m small (χ^2 approximation fails)

Overdisperson

Solution 1: Drop strict binomial assumption and let $E(y_i) = m_i p_i$, $\mathrm{Var}(y_i) = \phi m_i p_i (1 - p_i)$.

Pearson residuals

$$r_i = \frac{y_i - m_i \hat{p}_i}{\sqrt{m_i \hat{p}_i (1 - \hat{p}_i)}}$$

Simple estimate of dispersion parameter

Estimate $$\hat{\phi} = \frac{1}{n} \sum_{i=1}^{n} r_i^2.$$

- OK for estimation and standard errors on coefficients.
- But no proper inference via deviance.

Overdisperson

Solution 2: Define a "quasi-likelihood" that behaves like the log-likelihood but allows for $\mathrm{Var}(y_i) = \phi m_i p_i (1 - p_i)$.

```
fit <- glm(cbind(successes,failures) ~
    x1 + x2,
  family=quasibinomial, data=df)
```

- Must use *F*-tests rather than χ^2 tests for comparing models. (Approximation only.)

Inference for GLMs

Model	***F*-test**	**χ^2 test**
Normal OLS	Exact	–
Binary Logistic	Approx	Better approx
Binary Probit	Approx	Better approx
Binomial Logistic	Approx	Better approx
Quasibinomial logistic	Approx	–

Outline

Proportion responses: three approaches

Suppose $y \in [0, 1]$.

logitNormal model

$$\log(y/(1-y)) \mid \boldsymbol{x} \sim \mathrm{N}(\boldsymbol{x}'\beta, \sigma^2)$$

quasiBinomial model

$$\begin{aligned} y \mid \boldsymbol{x} \quad & \text{has mean } p \text{ and variance } \phi p(1-p) \\ \text{where} \quad & \log(p/(1-p)) = \boldsymbol{x}'\beta \end{aligned}$$

Beta model

$$\begin{aligned} & y \mid \boldsymbol{x} \sim \mathrm{Beta}(a, b) \\ \text{where} \quad & \mathrm{E}(y \mid \boldsymbol{x}) = a/(a+b) = e^{\boldsymbol{x}'\beta}/(1+e^{\boldsymbol{x}'\beta}) \end{aligned}$$

logitNormal model

logitNormal model

$$\log(y/(1-y)) \mid \boldsymbol{x} \sim \mathrm{N}(\boldsymbol{x}'\beta, \sigma^2)$$

- Provided no empirical proportions are at either 0 or 1, we can compute a logit transformation of the observed proportions. $\log(y/(1-y))$
- Then just fit a Gaussian linear regression using OLS.
- Back-transform the predictions using the inverse logit. $e^y/(1+e^y)$

```
lm(log(y/(1-y)) ~ x1 + x2, data=df)
```

Quasi-binomial model

quasiBinomial model

$$y \mid \boldsymbol{x} \quad \text{has mean } p \text{ and variance } \phi p(1-p)$$
$$\text{where} \qquad \log(p/(1-p)) = \boldsymbol{x}'\beta$$

- logit link keeps predicted proportions in $(0, 1)$
- Variance function $\phi p(1-p)$ makes sense for proportions as greatest variation around $p = 0.5$ and least around $p = 0$ and $p = 1$.

```
glm(y ~ x1 + x2,
  family=quasibinomial, data=df)
```

Beta regression

Beta model

$$y \mid \boldsymbol{x} \sim \text{Beta}(a, b)$$

$$\text{where} \qquad \text{E}(y \mid \boldsymbol{x}) = a/(a+b) = e^{\boldsymbol{x}'\beta}/(1+e^{\boldsymbol{x}'\beta})$$

Beta density

$$f(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}$$

where $y \in [0, 1]$ and $\Gamma(u) = \int_0^\infty x^{u-1}e^{-x}dx$.

- $\text{E}(y) = \frac{a}{a+b}$ $\qquad$ $\text{Var}(y) = \frac{ab}{(a+b)^2(a+b+1)}$

Beta regression

Beta model

$$y \mid \boldsymbol{x} \sim \text{Beta}(a, b)$$

where $\quad \text{E}(y \mid \boldsymbol{x}) = a/(a+b) = e^{\boldsymbol{x}'\beta}/(1+e^{\boldsymbol{x}'\beta})$

Beta density

$$f(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}$$

where $y \in [0,1]$ and $\Gamma(u) = \int_0^\infty x^{u-1}e^{-x}dx$.

- $\text{E}(y) = \frac{a}{a+b}$ $\qquad$ $\text{Var}(y) = \frac{ab}{(a+b)^2(a+b+1)}$

- Reparameterize so $\mu = a/(a+b)$ and $\phi = a+b$.

Beta regression

Beta model

$$y \mid \boldsymbol{x} \sim \text{Beta}(a, b)$$

where $\quad \text{E}(y \mid \boldsymbol{x}) = a/(a+b) = e^{\boldsymbol{x}'\beta}/(1+e^{\boldsymbol{x}'\beta})$

Beta density

$$f(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}$$

where $y \in [0, 1]$ and $\Gamma(u) = \int_0^\infty x^{u-1}e^{-x}dx$.

- $\text{E}(y) = \frac{a}{a+b} \qquad \text{Var}(y) = \frac{ab}{(a+b)^2(a+b+1)}$

- Reparameterize so $\mu = a/(a+b)$ and $\phi = a+b$.
- Then $\text{E}(Y) = \mu$ and $\text{Var}(Y) = \mu(1-\mu)/(1+\phi)$.

Beta regression

Beta model

$$y \mid \boldsymbol{x} \sim \text{Beta}(a, b)$$
$$\text{where} \qquad \text{E}(y \mid \boldsymbol{x}) = a/(a+b) = e^{\mathbf{x}'\beta}/(1+e^{\mathbf{x}'\beta})$$

Beta density

$$f(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}$$

where $y \in [0, 1]$ and $\Gamma(u) = \int_0^\infty x^{u-1}e^{-x}dx$.

- $\text{E}(y) = \frac{a}{a+b}$ $\qquad \text{Var}(y) = \frac{ab}{(a+b)^2(a+b+1)}$

- Reparameterize so $\mu = a/(a+b)$ and $\phi = a+b$.
- Then $\text{E}(Y) = \mu$ and $\text{Var}(Y) = \mu(1-\mu)/(1+\phi)$.

```
mgcv::gam(y ~ x1 + x2, family=betar())
```