北京航空航天大學
经济管理学院
BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

Generalized Linear Models

Lecture 4: Binomial and proportion responses

Outline

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.

$$E(Y) = mp.$$

$$\text{var}(Y) = mp(1-p).$$

Covariate class

- Suppose Y_i for $i = 1, 2, \cdots, n$ is binomially distributed $B(m_i, p_i)$ and Y_i are independent.
- The individual outcomes or trials that compose Y_i are all subject to the q predictors $(x_{i1}, x_{i2}, \cdots, x_{iq})$. Observations are available on m_i individuals.
- m_i individuals share the covariate vector $(x_{i1}, x_{i2}, \cdots, x_{iq})$.
- They are said to form a *covariate class*.
- The total number of individuals are $N = m_1 + m_2 + \cdots + m_n$.

Alternative ways of presenting the same data

(a) *Data listed by subject No.*

Subject No.	*Covariate* (x_1, x_2)	*Response* Y
1	1, 1	0
2	1, 2	1
3	1, 2	0
4	2, 1	0
5	2, 2	1
6	1, 2	1
7	1, 1	1

(b) *Data listed by covariate class*

Covariate (x_1, x_2)	*Class size* m	*Response* Y
1, 1	2	1
1, 2	3	2
2, 1	1	0
2, 2	1	1

Binomial responses

- Binomial regression also uses a logit link

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

$$\eta_i = log(\frac{p_i}{1 - p_i})$$

- Construct a linear predictor:

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

- The log-likelihood is then:

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

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.

Outline

Inference

- The binomial deviance:

$$
\begin{aligned}
D &= 2 \log L_L - 2 \log L_S \\
&= 2 \sum_{i=1}^{n} \log \frac{y_i}{\hat{y}_i} + (m_i - y_i) \log \frac{m_i - y_i}{m_i - \hat{y}_i}
\end{aligned}
$$

- Provided Y is truly binomial and m_i are relatively large, D is approximately χ^2_{n-q-1} if the model is correct.
- Then we can use the deviance to test whether the model is an adequate fit.
- If the deviance is far in excess of the df, the null hypothesis can be rejected (think about **why**).

Comparison two models

- $D_S - D_L \sim \chi^2_{l-s}$, where l and s are the number of parameters, assuming
  1. smaller model is correct
  2. models are nested
  3. distributional assumptions true
- Null deviance is for model with only an intercept.

Confidence intervals for coefficients

- Constructed using normal approximations for the parameter estimates.
- A 100$(1-\alpha)$% confidence interval for β_i would be :

$$\hat{\beta}_i \pm z^{\alpha/2} se(\hat{\beta}_i).$$

- Implemented in R using `confint`.

What if ungrouped?

What if all the cases that form a covariate class have not been grouped together?

Outline

Overdispersion

- If mean correctly modelled, but the residual deviance is much larger than expected, we called the data "overdispersed". [Same for underdispersion.]
- Concept of overdispersion irrelevant for logistic regression because there cannot be any more variance than what is modelled (we will show this later).
- For binomial regression: $y_i \sim B(m_i, p_i)$, $\mathrm{E}(y_i) = m_i p_i$, $\mathrm{V}(y_i) = m_i p_i(1 - p_i)$.
- Provided Y is truly binomial and m_i are relatively large, if the model is correct, $D \sim \chi^2_{n-q-1}$.
  So $D > n - q - 1$ indicates overdispersion.

Overdispersion

$D > n - q - 1$ can also be the result of:

- Wrong structure form for the model, for example:
  1. missing covariates or interaction terms
  2. negligence of non-linear effects
- Presence of large outliers
- Sparse data: m small (χ^2 approximation fails)
- p **is non-identical: sampling from clusters**
- p **is non-independent: correlation in responses**

Sampling from clusters

- Let the sample size be m, the cluster size be k and the number of clusters be $l = m/k$.
- Let the number of successes in cluster i be $Z_i \sim B(k, p_i)$.
- Now suppose that p_i is a random variable such that $Ep_i = p$ and $\mathrm{var}p_i = \tau^2 p(1-p)$.
- Let the total number of successes be $Y = Z_1 + \cdots + Z_l$. Then:

$$EY = \sum_{i=1}^{l} EZ_i = \sum_{i=1}^{l} E(E(Z_i|p_i)) = \sum_{i=1}^{l} kp = mp.$$

Sampling from clusters

$$\begin{aligned}\text{var}(Y) = \sum_{i=1}^{I} \text{var}(Z_i) &= \sum_{i=1}^{I} \{E(\text{var}(Z_i|p_i)) + \text{var}(E(Z_i|p_i))\} \\ &= \sum_{i=1}^{I} \{E(kp_i(1-p_i)) + k^2\tau^2 p(1-p)\} \\ &= (1 + (k-1)\tau^2)mp(1-p)\end{aligned}$$

- Y is overdispersed.
- Now think about logistic regression.

Correlation in responses

- Suppose $Y = \sum_{j=1}^{m} R_j$ and $R_j = \begin{cases} 1 & \text{sucess} \\ 0 & \text{otherwise} \end{cases}$
- With correlation, which means $\text{cor}(R_j, R_k) = \delta,\ k \neq j$, then we have:

$$\begin{aligned} \text{var}Y &= \sum_{j=1}^{m} \text{var}R_j + \sum_{j=1}^{m} \sum_{k \neq j} \text{cov}(R_j, R_k) \\ &= mp(1-p) + m(m-1)[\delta p(1-p)] \\ &= (1 + (m-1)\delta)mp(1-p) \end{aligned}$$

- Think about positive and negative correlation.

Overdispersion

Solution 1: Introduce an additional dispersion parameter, so that $\text{var}(Y_i) = \sigma^2 m_i p_i (1 - p_i)$.

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 $$\hat{\sigma}^2 = \frac{1}{n - p} \sum_{i=1}^{n} r_i^2.$$

- $\hat{\beta}$ is unaffected since σ^2 does not change the mean response.
- $\text{var}(\beta)$ will be scaled up by a factor of $\hat{\sigma}$.
- $F = \frac{(D_{small} - D_{large})/(df_{small} - df_{large})}{\hat{\sigma}^2}$ can be used when comparing models.

Overdispersion

Solution 2: Use a "quasi-likelihood".

- We need: inference $\Rightarrow$ distributions $\Rightarrow$ deviance $\Rightarrow$ likelihood.
- To construct a likelihood, we need distributions.
- But often where is no theory available on the random mechanism by which data were generated.
- We may only be able to specify some characteristics of the data, such as
  1. how the mean response changes with external variables.
  2. how the variation of the response changes with the average response.
  3. whether the responses are independent.
  4. whether the response distribution is skewed or symmetric.
- Quasi-likelihood is used to make inferences from experiments in which there is insufficient information to construct a likelihood function.

Quasi-likelihood

- Let Y_i have mean μ_i and variance $\phi V(\mu_i)$. Assume Y_i are independent. We define a score $U_i = \frac{Y_i - \mu_i}{\phi V(\mu_i)}$. Now:

$$EU_i = 0$$

$$\text{var } U_i = \frac{1}{\phi V(\mu_i)}$$

$$-E\frac{\partial U_i}{\partial \mu_i} = -E\frac{-\phi V(\mu_i) - (Y_i - \mu_i)\phi V'(\mu_i)}{[\phi V(\mu_i)]^2} = \frac{1}{\phi V(\mu_i)}$$

- These properties are shared by the derivative of the log-likelihood, l'. We can use U in place of l'. So we define:

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt$$

- Then we define the quasi-likelihood for all observations:

$$Q = \sum_{i=1}^{n} Q_i$$

Quasi-likelihood

- Quasi-likelihood depends directly only on the variance function.
- The choice of distribution also determines only the variance function.
- β is obtained by maximizing Q.
- The quasi-deviance is $-2\phi Q = -2\sum_i \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)} dt$.

Quasi-Binomial

- Variance function form: $V(y_i) = \phi m_i p_i (1 - p_i)$.

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

Outline

Proportion responses

Suppose $y_i \in [0, 1]$ are proportions. We can use a quasi-binomial model:

- 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 density:

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

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

- $\mathrm{E}(y) = \frac{a}{a+b}$ $\quad$ $\mathrm{V}(y) = \frac{ab}{(a+b)^2(a+b+1)}$

Beta regression

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 $\Gamma(u) = \int_0^\infty x^{u-1}e^{-x}dx$.

- $\mathrm{E}(y) = \frac{a}{a+b}$ $\quad$ $\mathrm{V}(y) = \frac{ab}{(a+b)^2(a+b+1)}$

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

Beta regression

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 $\Gamma(u) = \int_0^\infty x^{u-1}e^{-x}dx$.

- $\mathrm{E}(y) = \frac{a}{a+b}$ $\quad$ $\mathrm{V}(y) = \frac{ab}{(a+b)^2(a+b+1)}$

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

Beta regression

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 $\Gamma(u) = \int_0^\infty x^{u-1}e^{-x}dx$.

- $\mathrm{E}(y) = \frac{a}{a+b}$ $\quad$ $\mathrm{V}(y) = \frac{ab}{(a+b)^2(a+b+1)}$

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

$$\mu_i = e^{\eta_i}/(1+e^{\eta_i})$$
$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_q x_{i,q}$$

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

Outline

Latent variable interpretation

- Suppose z is a latent (unobserved) random variable:

$$y = \begin{cases} 1 & z = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon > 0 \\ 0 & \text{otherwise} \end{cases}$$

  where ε has cdf F.
- If F is "standard logistic", then $F(w) = 1/[1 + e^{-w}]$.
- So $\text{logit}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$.

That is, we can think of logistic regression as an ordinary regression with logistic noise, and we observe only if it is above or below 0.

Latent variable interpretation

- Suppose z is a latent (unobserved) random variable:

$$y = \begin{cases} 1 & z = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon > 0 \\ 0 & \text{otherwise} \end{cases}$$

  where ε has cdf F.
- If F is "standard normal", then $F(w) = \Phi(w)$.
- So $\Phi^{-1}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$.

Latent variable interpretation

- Suppose z is a latent (unobserved) random variable:

$$y = \begin{cases} 1 & z = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon > 0 \\ 0 & \text{otherwise} \end{cases}$$

  where ε has cdf F.
- If F is "standard normal", then $F(w) = \Phi(w)$.
- So $\Phi^{-1}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$.
- Here Φ^{-1} is the link function.

Latent variable interetation

General binary model

$$p_i = g(\eta_i) = P(Y_i = 1)$$

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

where g maps $\mathbb{R} \to (0, 1)$.

- $g(\eta) = e^{\eta}/(1 + e^{\eta})$: logit link, logistic regression
- $g(\eta) = \Phi(\eta)$: normal cdf link, probit regression
- $g(\eta) = 1 - \exp(-\exp(\eta))$: log-log link

Latent variable interetation

General binary model

$$p_i = g(\eta_i) = P(Y_i = 1)$$

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

where g maps $\mathbb{R} \to (0, 1)$.

- $g(\eta) = e^{\eta}/(1 + e^{\eta})$: logit link, logistic regression
- $g(\eta) = \Phi(\eta)$: normal cdf link, probit regression
- $g(\eta) = 1 - \exp(-\exp(\eta))$: log-log link

```
fit <- glm(y ~ x1 + x2,
  family=binomial(link=probit), data=df)
```

Outline

Assignment Two

Deadline: 8 April, 2018.

1. Derive the log-likelihood of binomial regression model.
2. Exercise 3 on Page 66 of our textbook.