Lecture 5: Overdispersion in logistic regression

Claudia Czado

TU München

Overview

- Definition of overdispersion
- Detection of overdispersion
- Modeling of overdispersion

Overdispersion in logistic regression

Collett (2003), Chapter 6

Logistic model: $Y_i \sim bin\,(n_i, p_i)$ independent

$$p_i = e^{\boldsymbol{x_i^t \beta}}/(1 + e^{\boldsymbol{x_i^t \beta}})$$

$$\Rightarrow \quad E(Y_i) = n_i p_i \qquad Var(Y_i) = n_i p_i (1 - p_i)$$

If one assumes that p_i is correctly modeled, but the observed variance is larger or smaller than the expected variance from the logistic model given by $n_i p_i (1 - p_i)$, one speaks of under or overdispersion. In application one often observes only overdispersion, so we concentrate on modeling overdispersion.

How to detect overdispersion?

If the logistic model is correct the asymptotic distribution of the residual deviance $D \sim \chi^2_{n-p}$. Therefore $D > n - p = E(\chi^2_{n-p})$ can indicate overdispersion.

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

- missing covariates and/or interaction terms;
- negligence of non linear effects;
- wrong link function;
- existance of large outliers;
- binary data or n_i small.

One has to exclude these reasons through EDA and regression diagnostics.

Residual deviance for binary logistic models

Collett (2003) shows that the residual deviance for binary logistic models can be written as

$$D = -2 \sum_{i=1}^{n} (\hat{p}_i \ln \left(\frac{\hat{p}_i}{1-\hat{p}_i} \right) + \ln(1-\hat{p}_i)),$$

where $\hat{p}_i = e^{\mathbf{x_i}^t \hat{\boldsymbol{\beta}}}/(1+e^{\mathbf{x_i}^t \hat{\boldsymbol{\beta}}})$. This is independent of Y_i, therefore not useful to assess goodness of fit.

Need to group data to use residual deviance as goodness of fit measure.

Reasons for overdispersion

Overdispersion can be explained by

- variation among the success probabilities or

- correlation between the binary responses

Both reasons are the same, since variation leads to correlation and vice versa. But for interpretative reasons one explanation might be more reasonable than the other.

Variation among the success probabilities

If groups of experimental units are observed under the same conditions, the success probabilities may vary from group to group.

Example: The default probabilities of a group of creditors with same conditions can vary from bank to bank. Reasons for this can be not measured or imprecisely measured covariates that make groups differ with respect to their default probabilities.

Correlation among binary responses

Let $Y_i = \sum_{j=1}^{n_i} R_{ij}$ $\quad R_{ij} = \begin{cases} 1 & \text{success} \\ 0 & \text{otherwise} \end{cases}$ $\quad P(R_{ij} = 1) = p_i$

$$\Rightarrow \quad Var(Y_i) \quad = \quad \sum_{j=1}^{n_i} \underbrace{Var(R_{ij})}_{p_i(1-p_i)} \quad + \quad \underbrace{\sum_{j=1}^{n_i} \sum_{k \neq j} Cov(R_{ij}, R_{ik})}_{\neq 0}$$

$$\neq n_i p_i (1 - p_i) = \text{ binomial variance}$$

Y_i has not a binomial distribution.

Examples:

- same patient is observed over time

- all units are from the same family or litter (cluster effects)

Modeling of variability among success probabilities

Williams (1982)

Y_i = Number of successes in n_i trials with random success probability v_i, $i = 1, \ldots, n$

Assume $E(v_i) = p_i$ $Var(v_i) = \phi p_i(1 - p_i)$, $\phi \geq 0$ unknown scale parameter.

Note: $Var(v_i) = 0$ if $p_i = 0$ or 1

$v_i \in (0, 1)$ is unobserved or latent random variable

Conditional expectation and variance of Y_i:

$$\begin{aligned} E(Y_i|v_i) &= n_i v_i \\ Var(Y_i|v_i) &= n_i v_i(1 - v_i) \end{aligned}$$

Since

$$\begin{aligned} E(Y) &= E_X(E(Y|X)) \\ Var(Y) &= E_X(Var(Y|X)) + Var_X(E(Y|X)), \end{aligned}$$

the unconditional expectation and variance is

$$\begin{aligned} E(Y_i) &= E_{v_i}(E(Y_i|v_i)) = E_{v_i}(n_i v_i) = n_i p_i \\ Var(Y_i) &= E_{v_i}(n_i v_i(1 - v_i)) + Var_{v_i}(n_i v_i) \\ &= n_i[E_{v_i}(v_i) - E_{v_i}(v_i^2)] + n_i^2 \phi p_i(1 - p_i) \\ &= n_i(p_i - \phi p_i(1 - p_i) - p_i^2) + n_i^2 \phi p_i(1 - p_i) \\ &= n_i p_i(1 - p_i)[1 + (n_i - 1)\phi] \end{aligned}$$

Remarks

- $\phi = 0 \Rightarrow$ no overdispersion

- $\phi > 0 \Rightarrow$ overdispersion if $n_i > 1$

- $n_i = 1$ (Bernoulli data) $\Rightarrow$ no information about ϕ available, this model is not useful

Modelling of correlation among the binary responses

$$Y_i = \sum_{j=1}^{n_i} R_{ij}, \quad R_{ij} = \begin{cases} 1 & \text{success} \\ 0 & \text{otherwise} \end{cases} \qquad P(R_{ij}=1) = p_i$$

$$\Rightarrow E(Y_i) = n_i p_i$$

but $\quad Cor(R_{ij}, R_{ik}) = \delta \quad k \neq j$

$$\Rightarrow \quad Cov(R_{ij}, R_{ik}) = \delta\sqrt{Var(R_{ij})Var(R_{ik})} = \delta p_i(1-p_i)$$

$$\Rightarrow \quad Var(Y_i) = \sum_{j=1}^{n_i} Var(R_{ij}) + \sum_{j=1}^{n_i}\sum_{k\neq j} Cov(R_{ij}, R_{ik})$$

$$= n_i p_i(1-p_i) + n_i(n_i-1)[\delta p_i(1-p_i)]$$

$$= n_i p_i(1-p_i)[1+(n_i-1)\delta]$$

Remarks

- $\delta = 0 \Rightarrow$ no overdispersion

- $\delta > 0 \Rightarrow$ overdispersion if $n_i > 1$
  $\delta < 0 \Rightarrow$ underdispersion.

- Since we need $1 + (n_i - 1)\delta > 0 \quad \delta$ cannot be too small. For $n_i \rightarrow \infty \Rightarrow \delta \geq 0$.

- Unconditional mean and variance are the same if $\delta \geq 0$ for both approaches, therefore we cannot distinguish between both approaches

Estimation of ϕ

$$Y_i|v_i \sim bin(n_i, v_i) \qquad E(v_i) = p_i \qquad Var(v_i) = \phi p_i(1-p_i) \qquad i = 1, \ldots, g$$

Special case $n_i = n \ \forall i$

$$Var(Y_i) = np_i(1-p_i)\underbrace{[1 + (n-1)\phi]}_{\sigma^2} \qquad \text{heterogenity factor}$$

One can show that

$$E(\chi^2) \approx (g-p)[1 + (n-1)\phi] = (g-p)\sigma^2$$

where $p =$ number of parameters in the largest model to be considered and $\chi^2 = \sum_{i=1}^{g} \frac{(y_i - n\hat{p}_i)^2}{n\hat{p}_i(1-\hat{p}_i)}$.

$$\Rightarrow \quad \hat{\sigma}^2 = \frac{\chi^2}{g-p} \quad \Rightarrow \quad \hat{\phi} = \frac{\hat{\sigma}^2 - 1}{n-1}$$

Estimation of $\boldsymbol{\beta}$ remains the same

Analysis of deviance when variability among the success probabilities are present

model	df	deviance	covariates
1	ν_1	D_1	$x_{i_1}, \ldots, x_{i\nu_1}$
2	ν_2	D_2	$x_{i_1}, \ldots, x_{i\nu_1}, x_{i(\nu_1+1)}, \ldots, x_{i\nu_2}$
0	ν_0	D_0	$x_{i_1}, \ldots, x_{i\nu_0}$

For $Y_i|v_i \sim bin(n_i, v_i),\ i = 1, \ldots, g.$
Since $E(\chi^2) \approx \sigma^2(g-p)$ we expect

$$\underbrace{\chi^2}_{\chi^2 \text{ Statistic}} \overset{a}{\sim} \underbrace{\sigma^2\chi^2_{g-p}}_{\text{distribution}} \quad \text{and} \quad D \overset{a}{\sim} \chi^2 \overset{a}{\sim} \sigma^2\chi^2_{g-p}$$

$$\Rightarrow \quad \frac{(D_1 - D_2)/(\nu_2 - \nu_1)}{D_0/\nu_0} \overset{a}{\sim} \frac{\chi^2_{\nu_2-\nu_1}}{\chi^2_{\nu_0}} \overset{a}{\sim} F_{\nu_2-\nu_1,\nu_0}$$

$\longrightarrow$ no change to ordinary case

Estimated standard errors in overdispersed models

$$\widehat{se}(\hat{\beta}_j) = \hat{\sigma} \cdot \widehat{se_0}(\hat{\beta}_j),$$

where

$\widehat{se_0}(\hat{\beta}_j) =$ estimated standard error in the model without overdispersion

This holds since $Var(Y_i) = \sigma^2 n_i p_i (1 - p_i)$ and in both cases we have $EY_i = p_i$.

Beta-Binomial models

$$
\begin{aligned}
v_i &= \text{latent success probability } \in (0,1) \\
v_i &\sim Beta(a_i, b_i)
\end{aligned}
$$

$$
\begin{aligned}
f(v_i) &= \tfrac{1}{B(a_i,b_i)} v_i^{a_i-1}(1-v_i)^{b_i-1},\ a_i, b_i > 0 \quad \text{density} \\
B(a,b) &= \int_0^1 x^{a-1}(1-x)^{b-1}dx \ - \ \text{Beta function}
\end{aligned}
$$

$$
\begin{aligned}
E(v_i) &= \tfrac{a_i}{a_i+b_i} =: p_i \\
Var(v_i) &= \tfrac{a_i b_i}{(a_i+b_i)^2(a_i+b_i+1)} = p_i(1-p_i)/[a_i+b_i+1] = p_i(1-p_i)\tau_i \\
\tau_i &:= \tfrac{1}{a_i+b_i+1}
\end{aligned}
$$

If $a_i > 1, b_i > 1 \quad \forall i$ we have unimodality and $Var(v_i) < p_i(1-p_i)\frac{1}{3}$.
If $\tau_i = \tau$, the beta binomial model is equivalent to the model with variability among success probabilities with $\phi = \tau < \frac{1}{3}$ ($\Rightarrow$ more restrictive).

(Marginal) likelihood

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \int_0^1 f(y_i|v_i) f(v_i) dv_i$$

$$= \prod_{i=1}^{n} \int \binom{n_i}{y_i} \frac{1}{B(a_i,b_i)} v_i^{y_i} (1-v_i)^{n_i-y_i} v_i^{a_i-1} (1-v_i)^{b_i-1} dv_i$$

where $p_i = e^{\boldsymbol{x_i^t\beta}}/(1+e^{\boldsymbol{x_i^t\beta}})$ $\qquad p_i = \frac{a_i}{a_i+b_i}$

$$= \prod_{i=1}^{n} \binom{n_i}{y_i} \frac{B(y_i+a_i, n_i-y_i+b_i)}{B(a_i,b_i)}$$

needs to be maximized to determine MLE of $\boldsymbol{\beta}$.

Remark: no standard software exists

Random effects in logistic regression

Let v_i = latent success probability with $E(v_i) = p_i$

$$\log\left(\frac{v_i}{1-v_i}\right) = \boldsymbol{x_i^t}\boldsymbol{\beta} + \delta_i \qquad \text{"random effect"}$$

δ_i measures missing or measured imprecisely covariates. When an intercept is included we can assume $E(\delta_i) = 0$. Further assume δ_i $i.i.d.$ with $Var(\delta_i) = \sigma_\delta^2$

Let Z_i $i.i.d.$ with $E(Z_i) = 0$ and $Var(Z_i) = 1$

$$\Rightarrow \delta_i \stackrel{D}{=} \gamma Z_i \qquad \text{with} \quad \gamma = \sigma_\delta^2 \geq 0$$

Therefore

$$\log\left(\frac{v_i}{1-v_i}\right) = \boldsymbol{x_i^t}\boldsymbol{\beta} + \gamma Z_i$$

Remark: this model can also be used for binary regression data

Estimation in logistic regression with random effects

If $Z_i \sim N(0,1)$ $i.i.d.$ the joint likelihood for $\boldsymbol{\beta}, \gamma, Z_i$ is given by

$$
\begin{aligned}
L(\boldsymbol{\beta}, \gamma, \mathbf{Z}) &= \prod_{i=1}^{n} \binom{n_i}{y_i} v_i^{y_i} (1 - v_i)^{n_i - y_i} \\
&= \prod_{i=1}^{n} \binom{n_i}{y_i} \frac{\exp\{\boldsymbol{x_i^t \beta} + \gamma Z_i\}^{y_i}}{[1 + \exp\{\boldsymbol{x_i^t \beta} + \gamma Z_i\}]^{n_i}} \qquad p + 1 + n \text{ parameters}
\end{aligned}
$$

Too many parameters, therefore maximize marginal likelihood

$$
\begin{aligned}
L(\boldsymbol{\beta}, \gamma) &:= \int_{\mathbb{R}^n} L(\boldsymbol{\beta}, \gamma, \mathbf{Z}) f(\mathbf{Z}) d\mathbf{Z} \\
&= \prod_{i=1}^{n} \binom{n_i}{y_i} \int_{-\infty}^{\infty} \frac{\exp\{\boldsymbol{x_i^t \beta} + \gamma Z_i\}^{y_i}}{[1 + \exp\{\boldsymbol{x_i^t \beta} + \gamma Z_i\}]^{n_i}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} Z_i^2} dZ_i
\end{aligned}
$$

This can only be determined numerically. One approach is to use a Gauss-Hermite approximation given by

$$\int_{-\infty}^{\infty} f(u)e^{-u^2}du \approx \sum_{j=1}^{m} c_j f(s_j)$$

for known c_j and s_j (see tables in Abramowitz and Stegun (1972)).
$m \approx 20$ is often sufficient.

Remarks for using random effects

- no standard software for maximization

- one can also use a non normal random effect

- extension to several random effects are possible. Maximization over high dim. integrals might require Markov Chain Monte Carlo (MCMC) methods

- random effects might be correlated in time or space, when time series or spatial data considered.

References

Abramowitz, M. and I. A. Stegun (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables. 10th printing, with corr.* John Wiley & Sons.

Collett, D. (2003). *Modelling binary data (2nd edition)*. London: Chapman & Hall.

Williams, D. (1982). Extra binomial variation in logistic linear models. *Applied Statistics 31*, 144–148.