Overdispersion

Definition

Overdispersion: Models with over-large deviances and residuals, but otherwise showing no systematic lack of fit. Structure in the data is obscured by additional noise, so overdispersion increases uncertainty.

Example: Count and proportion data are more variable than would be expected under the Poisson and binomial models.

Approaches to Dealing with Overdispersion

Parametric models

Basic Idea: Suppose that the response Y has a standard distribution conditional on the unobserved variable ϵ, but that ϵ induces extra variation in Y.

Model Setting: Assume ϵ is unobserved and it satisfies $\mathbb{E}(\epsilon) = 1$ and $\operatorname{var}(\epsilon) = \xi$. We further assume $Y \sim P(\mu\epsilon)$. Then, $\mathbb{E}(Y) = \mu$ and $\operatorname{var}(Y) = \mu(1 + \xi\mu) > \mu$. (Variance function is quadratic.)

In this way, we obtain a random variable whose variance is larger than Possion. If $\xi = 0$, this model reduces to a Possion model.

Note: If we let $\operatorname{var}(\epsilon) = \xi/\mu$, then $\operatorname{var}(Y) = \mu(1 + \xi)$ (Variance function is linear.)

Quasi-likelihood

Basic Idea: Modify standard methos to accomodate overdispersion and treat the generalized linear model score statistic as an estimating function $g(Y; \beta)$ for β.

Overview of Result: Estimators retain their large-sample normal distributions by fitting standard models, but with an inflated variance matrix .

Quasi-likelihood Equation

An estimator $\tilde{\beta}$ is obtained by solving *Quasi-likelihood Equation*:

$$g(Y; \beta) = X^{\mathrm{T}} u(\beta) = \sum_{j=1}^{n} x_j u_j(\beta) = \sum_{j=1}^{n} x_j \frac{Y_j - \mu_j}{g'(\mu_j)\, \phi_j V(\mu_j)} = 0$$

where $g(\mu_j) = \eta_j = x_j^{\mathrm{T}} \beta$.

If the mean structure has been chosen correctly, then $\mathbb{E}(Y_j) = \mu_j$ and the estimating function is unbiased, that is $\mathbb{E}\{g(Y;\beta)\} = 0$ for all β. Under regularity condition,

- $\mathbb{E}(\tilde{\beta}) \to \beta$
- $\tilde{\beta} \dot{\sim} \mathcal{N}(\beta, \mathrm{E}\left\{-\frac{\partial g(Y;\beta)}{\partial \beta^{\mathrm{T}}}\right\}^{-1} \mathrm{var}\{g(Y;\beta)\}\mathrm{E}\left\{-\frac{\partial g(Y;\beta)^{\mathrm{T}}}{\partial \beta}\right\}^{-1}),$
    - If the variance function specified, $\mathrm{var}(Y_j) = \phi_j V(\mu_j)$
      $$\tilde{\beta} \dot{\sim} \mathcal{N}(\beta, (X^\top W X)^{-1})$$
      where $W = \mathrm{diag}(\left\{g'(\mu_j)^2 \phi_j V(\mu_j)\right\}^{-1})$.
    - If the variance function mispecified,
      $$\tilde{\beta} \dot{\sim} \mathcal{N}(\beta, (X^\top W X)^{-1}(X^\top W' X)(X^\top W X)^{-1})$$
      where W' is a diagonal matrix involving the true and assumed variance functions.

Comment

- Under an exponential family model, the quasi-likelihood equation is the score statistics.
- $\tilde{\beta}$ is optimal among estimators based on linear combinations of the $Y_j - \mu_j$, in analogy with the Gauss–Markov theorem.
- The quasi-likelihood estimate $\tilde{\beta}$ equals the maximum likelihood estimate, but with smaller deviance if $\phi > 1$.

Quasi-likelihood Function

$g(Y;\beta)$ is the derivative with respect to β of the *quasi-likelihood function*

$$Q(\beta; Y) = \sum_{j=1}^{n} \int_{Y_j}^{\mu_j} \frac{Y_j - u}{\phi a_j V(u)} du$$

Deviance: $-2\phi Q(\beta; Y)$. This can compare nested model under overdispersion.

Reference

- Statstical Models