Modélisation des données de comptage et de la surdispersion

Jean-François DUPUY
Institut de Recherche Mathématique de Rennes

Journées de Statistique Mathématique et Data Science
Hammamet, 01-04 novembre 2019

Course material

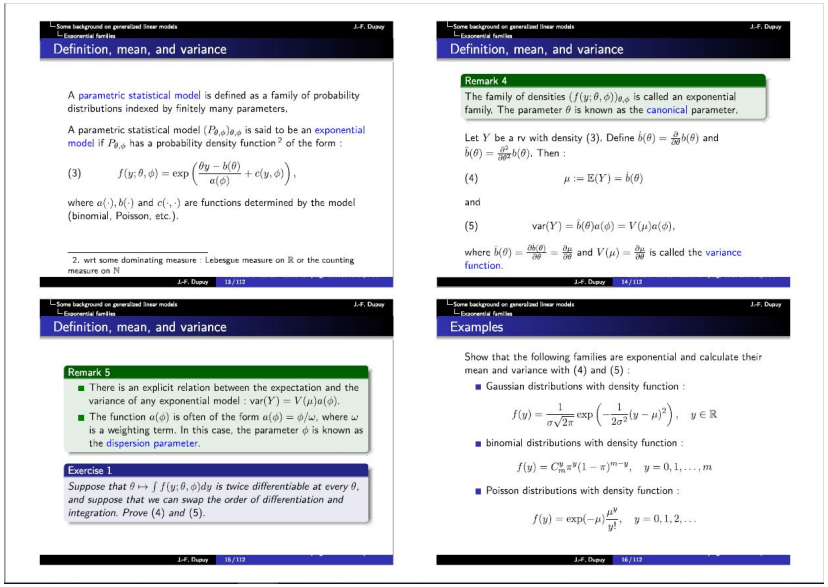
Lecture notes available

- in slide format :

  `http://dupuy.perso.math.cnrs.fr/JSMDS2019/snotes.pdf`

- in printable format :

  `http://dupuy.perso.math.cnrs.fr/JSMDS2019/pnotes.pdf`



Email : `Jean-Francois.Dupuy@insa-rennes.fr`

Generalized linear models (GLM)

GLM extend the familiar ANOVA and linear regression models for quantitative data to include count and frequencies data, for which the assumption of normal errors is no longer reasonable.

Most well-known examples include the :

- logistic regression model for binary [1] (or dichotomous) and binomial outcomes.

  Extensions allow polytomous outcomes (e.g. : improvement in a therapy with categories "none", "some" and "marked").
- Poisson regression model for count data (e.g., number of deaths per month due to some disease).

These models describe the relationship between the outcome (or "response") and a set of explanatory variables (or "predictors", "covariates"). Covariates can be continous or discrete.

1. e.g. : improvement in a therapy with categories "none" and "some".

Logistic regression model

In the logistic regression model for a binary response $Y_i \in \{0, 1\}$:

$$Y_i|\mathbf{X}_i \sim \mathcal{B}(\pi(\mathbf{X}_i)), \tag{1}$$

with probability of "success" $\pi(\mathbf{X}_i) := \mathbb{P}(Y_i = 1|\mathbf{X}_i)$ modeled as :

$$\pi(\mathbf{X}_i) = \frac{e^{\beta^\top \mathbf{X}_i}}{1 + e^{\beta^\top \mathbf{X}_i}} = \text{logit}^{-1}\left(\beta^\top \mathbf{X}_i\right), \tag{2}$$

or equivalently,

$$\text{logit}\left(\pi(\mathbf{X}_i)\right) = \log\left(\frac{\pi(\mathbf{X}_i)}{1 - \pi(\mathbf{X}_i)}\right) = \beta^\top \mathbf{X}_i,$$

where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a vector of unknown regression coefficients. Note that if $\beta_j > 0$, then $\pi(\mathbf{X}_i)$ increases as X_{ij} increases.

Logistic regression model

Remark 1

$\pi(\mathbf{X}_i)$ can be viewed as a model for the conditional expectation $\mu_i := \mathbb{E}[Y_i|\mathbf{X}_i]$ since for $Y_i|\mathbf{X}_i \sim \mathcal{B}(\pi(\mathbf{X}_i))$,

$$\pi(\mathbf{X}_i) = \mathbb{P}(Y_i = 1|\mathbf{X}_i) = \mathbb{E}[Y_i|\mathbf{X}_i].$$

Remark 2

(1) and (2) can be extended to the case of a binomial response :

$$Y_i|\mathbf{X}_i \sim \mathcal{B}(n_i, \pi(\mathbf{X}_i)),$$

with $\mathbb{P}(Y_i = y_i|\mathbf{X}_i) = \binom{n_i}{y_i}\pi(\mathbf{X}_i)^{y_i}(1 - \pi(\mathbf{X}_i))^{n_i - y_i}$.

Logistic regression model

Definition 1

- *the quantities* $\frac{\pi(\mathbf{X}_i)}{1-\pi(\mathbf{X}_i))}$ *and* $\text{logit}\left(\pi(\mathbf{X}_i)\right) = \log\left(\frac{\pi(\mathbf{X}_i)}{1-\pi(\mathbf{X}_i)}\right)$ *are called "odds" and "log odds" of success respectively.* β_j *can be viewed as the change in the log-odds associated with a unit increase in* X_{ij}*.*
- *if* i_1 *and* i_2 *have same covariates except the* j*-th one with* $X_{i_1 j} - X_{i_2 j} = 1$*, then the odds-ratio is*

$$\frac{\pi(\mathbf{X}_{i_1})}{1-\pi(\mathbf{X}_{i_1}))} \Big/ \frac{\pi(\mathbf{X}_{i_2})}{1-\pi(\mathbf{X}_{i_2}))} = e^{\beta_j}.$$

Poisson regression model

Poisson regression model specifies the conditional distribution of the count $Y_i \in \mathbb{N}$ as :

$$Y_i|\mathbf{X}_i \sim \mathcal{P}(\lambda(\mathbf{X}_i)),$$

which has density

$$f(y_i|\mathbf{X}_i) = \frac{e^{-\lambda(\mathbf{X}_i)}\lambda(\mathbf{X}_i)^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \dots.$$

The mean parameter $\mu_i := \mathbb{E}[Y_i|\mathbf{X}_i] = \lambda(\mathbf{X}_i)$ is usually modeled as :

$$\mu_i = e^{\beta^\top \mathbf{X}_i},$$

or equivalently $\ln(\mu_i) = \beta^\top \mathbf{X}_i$ (hence the name *log-linear model*).

Poisson regression model

Remark 3

- If t_i denotes the time length in which events occur (t_i known), then the conditional count distribution is modeled as $Y_i|\mathbf{X}_i \sim \mathcal{P}(t_i \cdot \lambda(\mathbf{X}_i))$.
- Under the assumption that $\lambda(\mathbf{X}_i) = e^{\beta^\top \mathbf{X}_i}$, $t_i \cdot \lambda(\mathbf{X}_i)$ writes as $e^{\beta^\top \mathbf{X}_i + \mathbf{1} \ln t_i}$. The term $\ln t_i$ is called *offset* and can be treated as a covariate with coefficient forced to 1.

Overdispersion in Poisson regression

- One property of Poisson regression model is mean-variance equality (or "equidispersion") conditional on explanatory variables :

$$\mathbb{E}[Y_i|\mathbf{X}_i] = \text{var}(Y_i|\mathbf{X}_i).$$

- If $\text{var}(Y_i|\mathbf{X}_i) > \mathbb{E}[Y_i|\mathbf{X}_i]$, data are said to be overdispersed (common causes of overdispersion : unobserved heterogeneity, excess zeros or *zero-inflation*)
- Intuitively, overdispersion arises when the variance of the data is underestimated. One consequence is that Poisson standard errors are too small, rendering the Wald tests $z_j = \widehat{\beta}_j / se(\widehat{\beta}_j)$ too large.

Definition, mean, and variance

A parametric statistical model is defined as a family of probability distributions indexed by finitely many parameters.

A parametric statistical model $(P_{\theta,\phi})_{\theta,\phi}$ is said to be an exponential model if $P_{\theta,\phi}$ has a probability density function [2] of the form :

$$f(y;\theta,\phi) = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{3}$$

where $a(\cdot), b(\cdot)$ and $c(\cdot,\cdot)$ are functions determined by the model (binomial, Poisson, etc.).

2. wrt some dominating measure : Lebesgue measure on $\mathbb{R}$ or the counting measure on $\mathbb{N}$

Definition, mean, and variance

Remark 4

The family of densities $(f(y;\theta,\phi))_{\theta,\phi}$ is called an exponential family. The parameter θ is known as the canonical parameter.

Let Y be a rv with density (3). Define $\dot{b}(\theta) = \frac{\partial}{\partial\theta} b(\theta)$ and $\ddot{b}(\theta) = \frac{\partial^2}{\partial\theta^2} b(\theta)$. Then :

$$\mu := \mathbb{E}(Y) = \dot{b}(\theta) \tag{4}$$

and

$$\text{var}(Y) = \ddot{b}(\theta)a(\phi) = V(\mu)a(\phi), \tag{5}$$

where $\ddot{b}(\theta) = \frac{\partial \dot{b}(\theta)}{\partial\theta} = \frac{\partial\mu}{\partial\theta}$ and $V(\mu) = \frac{\partial\mu}{\partial\theta}$ is called the variance function.

Definition, mean, and variance

Remark 5

- There is an explicit relation between the expectation and the variance of any exponential model : $\text{var}(Y) = V(\mu)a(\phi)$.
- The function $a(\phi)$ is often of the form $a(\phi) = \phi/\omega$, where ω is a weighting term. In this case, the parameter ϕ is known as the dispersion parameter.

Exercise. Suppose that $\theta \mapsto \int f(y;\theta,\phi)dy$ is twice differentiable at every θ and that we can swap the order of differentiation and integration. Prove (4) and (5).

Examples

Exercise. Show that the following families are exponential and calculate their mean and variance with (4) and (5) :

- Gaussian distributions with density function :

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right), \quad y \in \mathbb{R}$$

- binomial distributions with density function :

$$f(y) = C_m^y \pi^y (1-\pi)^{m-y}, \quad y = 0, 1, \ldots, m$$

- Poisson distributions with density function :

$$f(y) = \exp(-\mu)\frac{\mu^y}{y!}, \quad y = 0, 1, 2, \ldots$$

Examples

- gamma distributions with density function :

$$f(y) = y^{-1} \exp\left(-\frac{y\nu}{\mu}\right) \left(\frac{y\nu}{\mu}\right)^{\nu} \frac{1}{\Gamma(\nu)}, \quad y > 0$$

- negative binomial distributions with density function :

$$f(y) = C_{y+k-1}^{y} \pi^{k} (1-\pi)^{y}, \quad y = 0, 1, 2, \ldots$$

Hint : Let $\kappa = 1/k$ and $\mu = k\left(\frac{1-\pi}{\pi}\right)$ and recall that for any $n \in \mathbb{N}$, $\Gamma(n+1) = n!$

Canonical link

Remark 6

For each example above, the canonical parameter θ is a function of the expected value $\mu := \mathbb{E}(Y)$:

$$\theta = g(\mu).$$

This function g is known as the *canonical link function*. Note that $\mu = \mathbb{E}(Y) = \dot{b}(\theta)$, so $\theta = g(\dot{b}(\theta))$, and hence $g = \dot{b}^{-1}$.

- Gaussian $Y \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = \mathrm{id}(\mu)$ (*identity* link)
- binomial $Y \sim \mathcal{B}(m, \pi)$, $\theta = \ln(\frac{\pi}{1-\pi}) = \ln(\frac{\mu}{m-\mu})$ (*logit* link)
- Poisson $Y \sim \mathcal{P}(\mu)$, $\theta = \ln \mu$ (*log* link)
- gamma $Y \sim G(\mu, \nu)$, $\theta = -\frac{1}{\mu}$
- negative binomial, $\theta = \ln(\frac{\kappa\mu}{1+\kappa\mu})$

Three components

To define a generalized linear model, we need three elements :

1. a random component : the Z_i, $i = 1, \ldots, n$ are assumed to be independent with density belonging to the exponential family :

$$f(z_i; \theta_i, \phi) = \exp\left(\frac{\theta_i z_i - b(\theta_i)}{a(\phi)} + c(z_i, \phi)\right), \quad i = 1, \ldots, n,$$

   where $a(\cdot), b(\cdot)$, and $c(\cdot, \cdot)$ specify the model and ϕ is assumed to be identical for every observation.

2. a systematic component defined by a linear predictor :

$$\eta_i = \beta^\top \mathbf{X}_i,$$

Three components

3. a link function $g(\cdot)$ that relates the linear predictor η_i with the mean $\mu_i := \mathbb{E}[Z_i|\mathbf{X}_i]$ of Y_i, by :

$$g(\mu_i) = \eta_i,$$

with g a monotonic and differentiable function, known as the link function.

The canonical link function $g = \dot{b}^{-1}$ is often used. In this case, $\theta_i = g(\mu_i) = \beta^\top \mathbf{X}_i$.

Linear regression as a GLM

Consider the linear model :

$$Y_i = \beta^\top \mathbf{X}_i + \epsilon_i = \mu_i + \epsilon_i \quad \text{with } \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

The conditional density of Y_i has an exponential family form since :

$$\begin{aligned} f(y_i; \theta_i, \phi) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right), \\ &= \exp\left(\frac{\mu_i y_i - \frac{\mu_i^2}{2}}{\sigma^2} - \frac{1}{2}\left[\ln(2\pi\sigma^2) + \frac{y_i^2}{\sigma^2}\right]\right). \end{aligned}$$

Here : $\theta_i = \mu_i$, $a(\phi) = \sigma^2$, $b(\theta_i) = \frac{\mu_i^2}{2}$, $c(y_i, \phi) = -\frac{1}{2}\left[\ln(2\pi\sigma^2) + \frac{y_i^2}{\sigma^2}\right]$,
with $\mathbb{E}[Y_i|\mathbf{X}_i] = \dot{b}(\theta_i) = \mu_i$ and $\text{var}(Y_i|\mathbf{X}_i) = a(\phi) \cdot \ddot{b}(\theta_i) = \sigma^2$.

Further,

$$\mu_i = \mathbb{E}[Y_i|\mathbf{X}_i] = \beta^\top \mathbf{X}_i = \eta_i,$$

thus $g(\mu_i) = \eta_i$ with the identity $g(\cdot) = \text{id}(\cdot)$ as link function.

Poisson regression as a GLM

Let $Y_i|\mathbf{X}_i \sim \mathcal{P}(t_i\lambda(\mathbf{X}_i))$. The conditional density of Y_i has an exponential family form since :

$$\begin{aligned} f(y_i;\theta_i,\phi) &= \frac{e^{-t_i\lambda(\mathbf{X}_i)}(t_i\lambda(\mathbf{X}_i))^{y_i}}{y_i!}, \\ &= \exp\left(y_i \ln(t_i\lambda(\mathbf{X}_i)) - t_i\lambda(\mathbf{X}_i) - \ln(y_i!)\right). \end{aligned}$$

Here : $\theta_i = \ln(t_i\lambda(\mathbf{X}_i))$, $\mathbf{a(\phi) = 1}$, $b(\theta_i) = t_i\lambda(\mathbf{X}_i) = e^{\theta_i}$ and $c(y_i,\phi) = -\ln(y_i!)$.

Moreover,

$$\mathbb{E}[Y_i|\mathbf{X}_i] = \dot{b}(\theta_i) = e^{\theta_i} = t_i\lambda(\mathbf{X}_i),$$

and

$$\text{var}(Y_i|\mathbf{X}_i) = a(\phi)\cdot\ddot{b}(\theta_i) = e^{\theta_i} = t_i\lambda(\mathbf{X}_i).$$

Poisson regression as a GLM

The usual choice for the link function $g(\cdot)$ is to take :

$$\mu_i = \mathbb{E}[Y_i|\mathbf{X}_i] = t_i\lambda(\mathbf{X}_i) = t_i e^{\beta^\top \mathbf{X}_i},$$

hence

$$\eta_i = \beta^\top \mathbf{X}_i = \ln \mu_i - \ln t_i,$$

and thus, $\eta_i = g(\mu_i)$ with $g(\mu_i) = \ln \mu_i - \ln t_i$ (with $\ln t_i$ a known offset term). If $t_i = 1$ for every i, then $g(\mu_i) = \ln \mu_i$ (logarithmic link).

Likelihood equations

Suppose we have a sample $Z_1, \ldots, Z_n$ of independent observations with density functions :

$$f(z_i; \theta_i, \phi) = \exp\left(\frac{\theta_i z_i - b(\theta_i)}{a(\phi)} + c(z_i, \phi)\right), \quad i = 1, \ldots, n.$$

For each individual, a vector $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top$ of explanatory variables(quantitative and/or qualitative) is also observed.

Likelihood of (β, ϕ) wrt the sample $(Z_1, \mathbf{X}_1), \ldots, (Z_n, \mathbf{X}_n)$:

$$L_n(\beta, \phi) = \prod_{i=1}^{n} \exp\left(\frac{\theta_i Z_i - b(\theta_i)}{a(\phi)} + c(Z_i, \phi)\right).$$

Likelihood equations

Log-likelihood $\ell_n(\beta, \phi) = \ln L_n(\beta, \phi)$:

$$\ell_n(\beta, \phi) = \sum_{i=1}^{n} \left\{ \frac{\theta_i Z_i - b(\theta_i)}{a(\phi)} + c(Z_i, \phi) \right\} := \sum_{i=1}^{n} \ell_{n,i}(\beta, \phi).$$

The maximum likelihood estimator (MLE) $\hat{\beta}_n$ of β is obtained by solving the system of p equations :

$$\frac{\partial}{\partial \beta} \ell_n(\beta, \phi) \Big|_{\beta = \hat{\beta}_n} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \ell_{n,i}(\beta, \phi) \Big|_{\beta = \hat{\beta}_n} = 0.$$

Likelihood equations

Remark 7 (Leibniz notation)

If $z = f(y)$ and $y = g(x)$, the derivative $\frac{\partial f(g(x))}{\partial x} = \dot{f}(g(x))\dot{g}(x)$ is written $\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$.

Then for each $j = 1, \ldots, p$:

$$\frac{\partial \ell_{n,i}}{\partial \beta_j} = \frac{\partial \ell_{n,i}}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j},$$

with :

$$\frac{\partial \ell_{n,i}}{\partial \theta_i} = \frac{Z_i - \dot{b}(\theta_i)}{a(\phi)} = \frac{Z_i - \mu_i}{a(\phi)}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = V(\mu_i) = \frac{\text{var}(Z_i)}{a(\phi)} \quad \text{and} \quad \frac{\partial \eta_i}{\partial \beta_j} = X_{ij}.$$

Likelihood equations

By contrast, $\partial \mu_i / \partial \eta_i$ depends on the choice of the link $g(\mu_i) = \eta_i$:

$$\frac{\partial}{\partial \beta_j} \ell_n(\beta, \phi) = \sum_{i=1}^{n} X_{ij} \frac{(Z_i - \mu_i)}{V(\mu_i) a(\phi)} \frac{\partial \mu_i}{\partial \eta_i}, \quad j = 1, \dots, p.$$

This gives the likelihood equations for the β_j :

$$\sum_{i=1}^{n} X_{ij} \frac{(Z_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, \dots, p. \tag{6}$$

Remark 8

These equations are non-linear in β (except in certain special cases) $\Rightarrow$ iterative algorithms (e.g. Newton-Raphson, Fisher-scoring).

Dispersion parameter

Remark 9

Likelihood equations do not depend on ϕ thus neither does the MLE of $\beta \Rightarrow$ dispersion parameter ϕ can be estimated separately

Recall that $\text{var}(Y) = V(\mu)a(\phi)$ (consider $a(\phi) = \phi$). Then :

$$\phi = \frac{\text{var}(Y)}{V(\mu)}$$

with $g(\mu) = \beta^\top \mathbf{X}$. Pearson's χ^2 estimator is defined by :

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

with $\hat{\mu}_i = g^{-1}(\hat{\beta}_n^\top \mathbf{X}_i)$.

Canonical link function

If we take g as the canonical link, then $\theta_i = g(\mu_i) = \beta^\top \mathbf{X}_i = \eta_i$ and thus :

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = V(\mu_i).$$

Equation (6) reduces to :

$$\sum_{i=1}^{n} X_{ij}(Z_i - \mu_i) = 0, \quad j = 1, \dots, p,$$

Remark 10

These equations can be written in matrix form (with usual notations), as : $\mathbb{X}^\top(\mathbf{Z} - \mu) = 0$.

Example 1 : Gaussian linear model

Let $Z_i = \beta^\top \mathbf{X}_i + \epsilon_i = \mu_i + \epsilon_i$. We have $\mu_i = \beta^\top \mathbf{X}_i$ and $\mu = \mathbb{X}\beta$. The likelihood equation becomes :

$$\mathbb{X}^\top (\mathbf{Z} - \mathbb{X}\beta) = 0,$$

with solution

$$\hat{\beta}_n = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{Z}$$

whenever $\mathbb{X}^\top \mathbb{X}$ is invertible.

Example 2 : binomial model

Let $Z_i \sim \mathcal{B}(m_i, p(\mathbf{X}_i))$, then $\mu_i = \mathbb{E}(Z_i|\mathbf{X}_i) = m_i p(\mathbf{X}_i)$. If g is the canonical link function, then :

$$g(\mu_i) = \ln\left(\frac{\mu_i}{m_i - \mu_i}\right) = \ln\left(\frac{p(\mathbf{X}_i)}{1 - p(\mathbf{X}_i)}\right) = \text{logit}(p(\mathbf{X}_i)) = \beta^\top \mathbf{X}_i,$$

or $p(\mathbf{X}_i) = e^{\beta^\top \mathbf{X}_i}/(1 + e^{\beta^\top \mathbf{X}_i})$. Likelihood equations are :

$$\sum_{i=1}^{n} X_{ij}(Z_i - m_i p(\mathbf{X}_i)) = 0, \quad j = 1, \ldots, p.$$

Remark 11

The MLE has no explicit expression and needs to be approximated numerically.

Example 3 : Poisson model

Let $Z_i \sim \mathcal{P}(\lambda(\mathbf{X}_i))$, then $\mu_i = \mathbb{E}(Z_i|\mathbf{X}_i) = \lambda(\mathbf{X}_i)$. If g is the canonical link function, then :

$$g(\mu_i) = \ln \mu_i = \ln \lambda(\mathbf{X}_i) = \beta^\top \mathbf{X}_i,$$

or $\mu_i = \lambda(\mathbf{X}_i) = e^{\beta^\top \mathbf{X}_i}$. Likelihood equations are :

$$\sum_{i=1}^{n} X_{ij}(Z_i - e^{\beta^\top \mathbf{X}_i}) = 0, \quad j = 1, \dots, p.$$

In matrix form, $\mathbb{X}^\top(\mathbf{Z} - \mu) = 0$, where $\mu = (e^{\beta^\top \mathbf{X}_1}, \dots, e^{\beta^\top \mathbf{X}_n})^\top$.

Remark 12

The MLE has no explicit expression and needs to be approximated numerically.

Example 4 : gamma model

Let $Z_i \sim G(\mu(\mathbf{X}_i), \nu)$, then $\mu_i = \mathbb{E}(Z_i|\mathbf{X}_i) = \mu(\mathbf{X}_i)$. If g is the canonical link function, then :

$$g(\mu_i) = -\frac{1}{\mu_i} = \beta^\top \mathbf{X}_i,$$

or $\mu_i = -1/\beta^\top \mathbf{X}_i$.

Remark 13

Support of the gamma distribution is $]0, \infty[$, so $\mu_i > 0$. Therefore, $\beta^\top \mathbf{X}_i < 0$, which restricts the estimates of β.

$\Rightarrow$ the log link is usually preferred :

$$\ln \mu_i = \beta^\top \mathbf{X}_i.$$

In other words, $\mu_i = e^{\beta^\top \mathbf{X}_i}$.

Example 4 : gamma model

Remark 14

- In R, three possible links for the gamma distribution : the identity, log and inverse function $g(\mu_i) = 1/\mu_i$.
- The "inverse" link and canonical link are equivalent. To see this, note that if

$$-\frac{1}{\mu_i} = \beta^\top \mathbf{X}_i,$$

then

$$\frac{1}{\mu_i} = \beta^{*\top} \mathbf{X}_i,$$

with $\beta^* = -\beta$.

Asymptotic properties

We give sufficient conditions for consistency and asymptotic normality of the MLE in a glm with canonical link. We write $\mathcal{I}_n(\beta) = -\partial^2 \ell_n(\beta)/\partial\beta\partial\beta^\top$.

Theorem 1

Suppose that covariates $X_{i1}, X_{i2}, \ldots, X_{ip}, i = 1, 2, \ldots$ are bounded and $\lambda_{\min}(\mathbb{X}^\top\mathbb{X}) \to \infty$ as $n \to \infty$. Then $\hat{\beta}_n \xrightarrow{p} \beta$ and $\mathcal{I}_n(\hat{\beta}_n)^{\frac{1}{2}}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, I_p)$.

These results[3] are useful for finding confidence intervals or for testing significance of one or several regressors.

3. Gourieroux and Monfort (1981) ; Fahrmeir and Kaufmann (1985)

Confidence intervals

Let $\hat{\sigma}_j^2$ be the j-th term on the diagonal of $\mathcal{I}_n(\hat{\beta}_n)^{-1}$. We have :

$$\hat{\beta}_{n,j} \approx N(\beta_j, \hat{\sigma}_j^2).$$

Hence an asymptotic $(1-\alpha)$-level confidence interval for β_j :

$$\left[\hat{\beta}_{n,j} - u_{1-\alpha/2}\hat{\sigma}_j \; ; \hat{\beta}_{n,j} + u_{1-\alpha/2}\hat{\sigma}_j\right],$$

with $u_{1-\alpha/2}$ the quantile of order $1-\alpha/2$ of $\mathcal{N}(0,1)$ [4].

Remark 15

The quantity $\hat{\sigma}_j$ is called the *standard error* of $\hat{\beta}_{n,j}$.

4. defined by $\mathbb{P}(\mathcal{N}(0,1) \leq u_{1-\alpha/2}) = 1-\alpha/2$

Test for a single component of β (Wald test)

We wish to test non-significance of the j-th covariate in $\beta^\top \mathbf{X}_i$:

$$H_0 : \beta_j = 0 \text{ against } H_1 : \beta_j \neq 0$$

Under H_0, the Wald statistic $\hat{\beta}_{n,j}/\hat{\sigma}_j \approx \mathcal{N}(0,1)$. The following region rejects H_0 at the level α :

$$\mathcal{R}_\alpha = \left\{ \left| \frac{\hat{\beta}_{n,j}}{\hat{\sigma}_j} \right| \geq u_{1-\alpha/2} \right\}.$$

Likelihood-ratio test

We wish to test (without loss of generality) :

$$H_0 : \beta_1 = \ldots = \beta_q = 0$$

against

$$H_1 : \text{ there exists } i \in \{1, \ldots, q\} \text{ such that } \beta_i \neq 0.$$

The LRT compares the likelihoods under H_0 and H_1 and accepts H_0 if they are "close." Define :

$$\begin{aligned} D_n &= 2 \ln \left(\frac{L_n(\hat{\beta}_n)}{L_n(\hat{\beta}_{n,H_0})} \right) \\ &= 2(\ell_n(\hat{\beta}_n) - \ell_n(\hat{\beta}_{n,H_0})) \end{aligned}$$

Likelihood-ratio test

Under H_0, $D_n \approx \chi^2_q$ when n is large, hence a region of rejection of H_0 at the level α :

$$\mathcal{R}^D_\alpha = \{D_n \geq c_q(1-\alpha)\},$$

with $c_q(1-\alpha)$ the $(1-\alpha)$-quantile of χ^2_q.

Equidispersion vs overdispersion

Poisson distribution is equidispersed : its mean is equal to its variance.

In a Poisson regression model $Z|\mathbf{X} \sim \mathcal{P}(\lambda(\mathbf{X}))$, equidispersion can be stated as :

$$\mathbb{E}(Z|\mathbf{X}) = \text{var}(Z|\mathbf{X}).$$

One empirical way of checking whether $Z|\mathbf{X}$ is equidispersed is to :

- estimate $\mathbb{E}(Z|\mathbf{X} = \mathbf{x})$ and $\text{var}(Z|\mathbf{X} = \mathbf{x})$ for each $\mathbf{x}$ by the empirical mean and variance of values Z_i such that $\{\mathbf{X}_i = \mathbf{x}\}$
- plot pairs (estimated mean, estimated variance) in the plane.

If equidispersion holds, the scatterplot should coincide approximately with the line $y = x$.

Demand for medical care by the elderly

US National Medical Expenditure Survey (NMES) data

- a large cross-sectional carried out in 1987-1988 to assess the demand for medical care
- based upon $n = 4406$ individuals covered by Medicare [5]

Dataset originally taken from :

- Deb P., Trivedi P.K. *Demand for medical care by the elderly : a finite mixture approach*. Journal of Applied Econometrics 12, 313-336, 1997.

Now available as `NMES1988` in `R` package `AER` which comes with :

- Kleiber C., Zeileis A. *Applied Econometrics with R*. Springer, 2008.

5. a public insurance program providing protection against health-care costs for Americans aged 65 and older

Demand for medical care by the elderly

Objective : create a descriptive and predictive model demand for medical care - here defined as number of physician office visits - based on the covariates available for the patients.

Potential response variables : a 2×2 set of the combinations of *place of visits* (office vs. hospital/outpatient) and kind of practitioner (doctor vs. non-doctor, e.g. nurse, optician. . .)

Covariates

- self-perceived health level (poor, average, excellent)
- number of chronic conditions (cancer, arthritis, diabete. . .)
- socio-economic variables : gender, age, marital status, educational level, income
- private insurance indicator

Equidispersion vs overdispersion

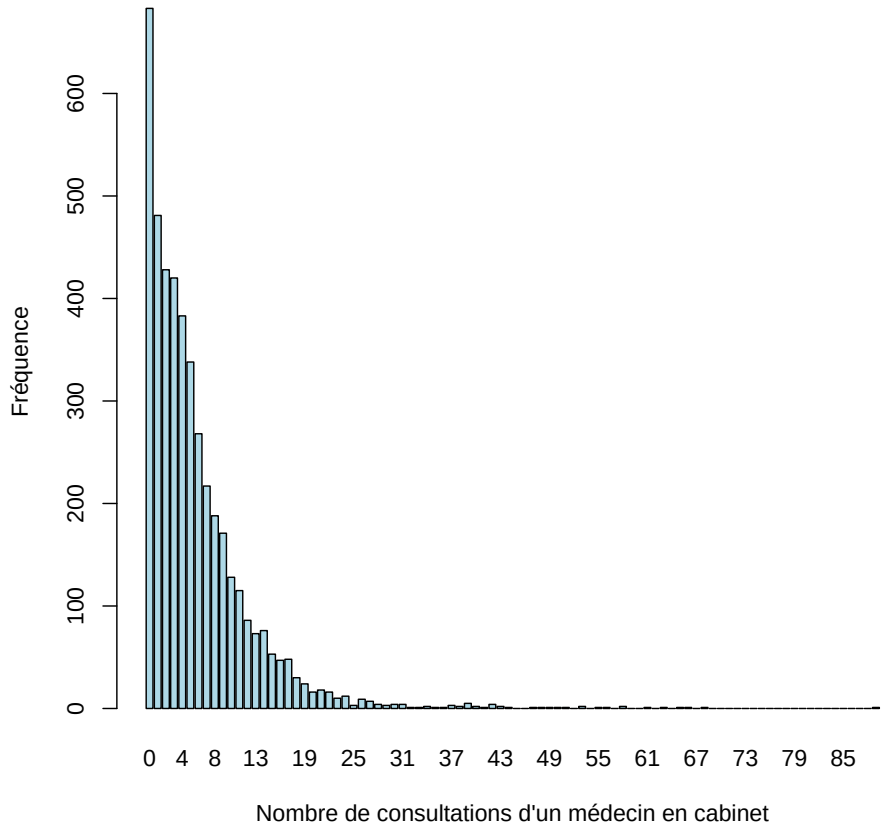


FIGURE 1 – Frequency distribution of the number of office visits to a physician (`ofp`).

Equidispersion vs overdispersion

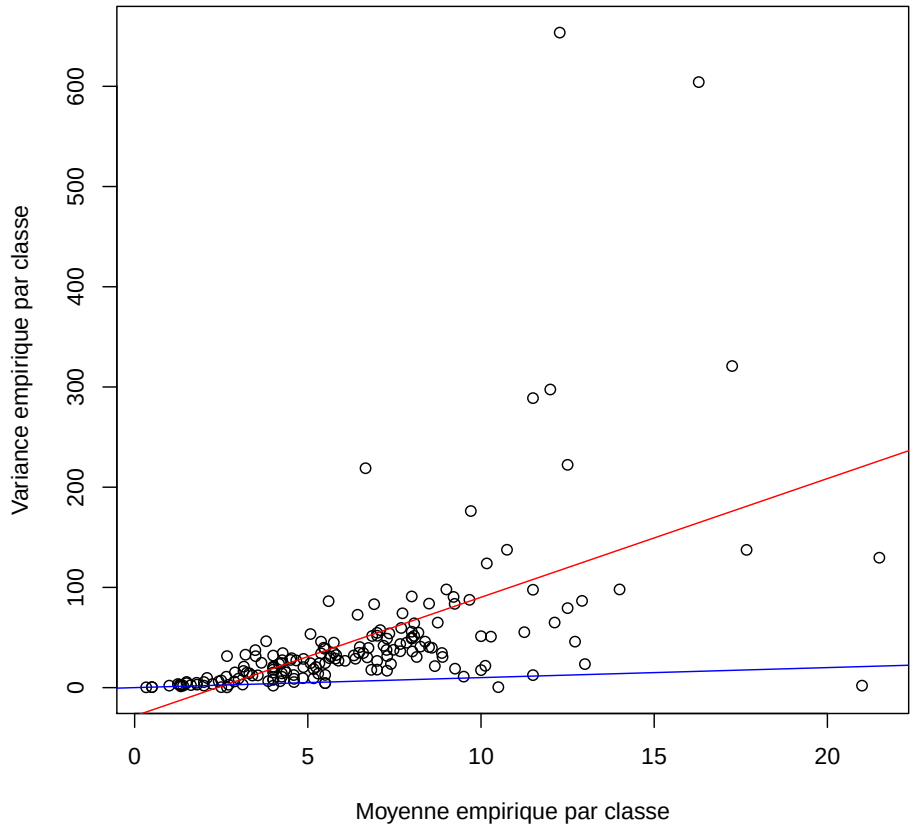


FIGURE 2 – Empirical mean and variance of `ofp` within each class `x` obtained after segmentation of `health`, `med`, `age`, and `numchron`.

Overdispersion

Overdispersion can arise for various reasons, e.g. :

- presence of unobserved heterogeneity in the data
- zero-inflation

For example, consider the model :

$$Z_i \sim \begin{cases} \mathcal{P}(\lambda_0) \text{ if } X_i = 0 \\ \mathcal{P}(\lambda_1) \text{ if } X_i = 1 \end{cases}$$

where $X_i \sim \mathcal{B}(\pi_i)$ and $\lambda_0 \neq \lambda_1$. Equivalently, $Z_i|X_i = 0 \sim \mathcal{P}(\lambda_0)$ and $Z_i|X_i = 1 \sim \mathcal{P}(\lambda_1)$.

Overdispersion

Suppose that we only observe Z_i (or that X_i is observed but its effect is not modeled). Then :

$$\begin{aligned}\mathbb{E}(Z_i) &= \mathbb{E}[\mathbb{E}[Z_i|X_i]] \\ &= \mathbb{E}(\lambda_0(1-X_i)+\lambda_1 X_i) \\ &= \pi_i\lambda_1 + (1-\pi_i)\lambda_0\end{aligned}$$

and

$$\begin{aligned}\mathsf{var}(Z_i) &= \mathbb{E}[\mathsf{var}(Z_i|X_i)] + \mathsf{var}(\mathbb{E}[Z_i|X_i]) \\ &= \mathbb{E}(Z_i) + \mathsf{var}(\lambda_0 + (\lambda_1-\lambda_0)X_i) \\ &= \mathbb{E}(Z_i) + (\lambda_1-\lambda_0)^2\pi_i(1-\pi_i).\end{aligned}$$

If $\pi_i \neq 0, 1$, then :

$$\mathsf{var}(Z_i) > \mathbb{E}(Z_i)$$

and the distribution of Z_i is overdispersed.

Overdispersion

Omission of one or several key explanatory variables from the linear predictor can also yield overdispersion :

A numerical experiment :

1. simulate 5000 indep. realizations of $X_1, X_2 \sim \mathcal{U}[0,1]$ and $X_3 \sim \mathcal{N}(0,1)$
2. simulate n indep. counts $Z_i \sim \mathcal{P}(e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}})$
3. discretize each X_k into 8 categories, yielding $N = 8^3 = 512$ classes $\mathbf{x}_1, \ldots, \mathbf{x}_N$
4. estimate $\mathbb{E}(Z|\mathbf{X} = \mathbf{x}_j)$ and $\text{var}(Z|\mathbf{X} = \mathbf{x}_j)$, $j = 1, \ldots, N$ and plot the N estimated pairs (blue circles)
5. regress the N empirical variances on the N empirical means (dashed line)
6. redo steps 4-5 while omitting X_3 (red dots and dashed line).

Overdispersion

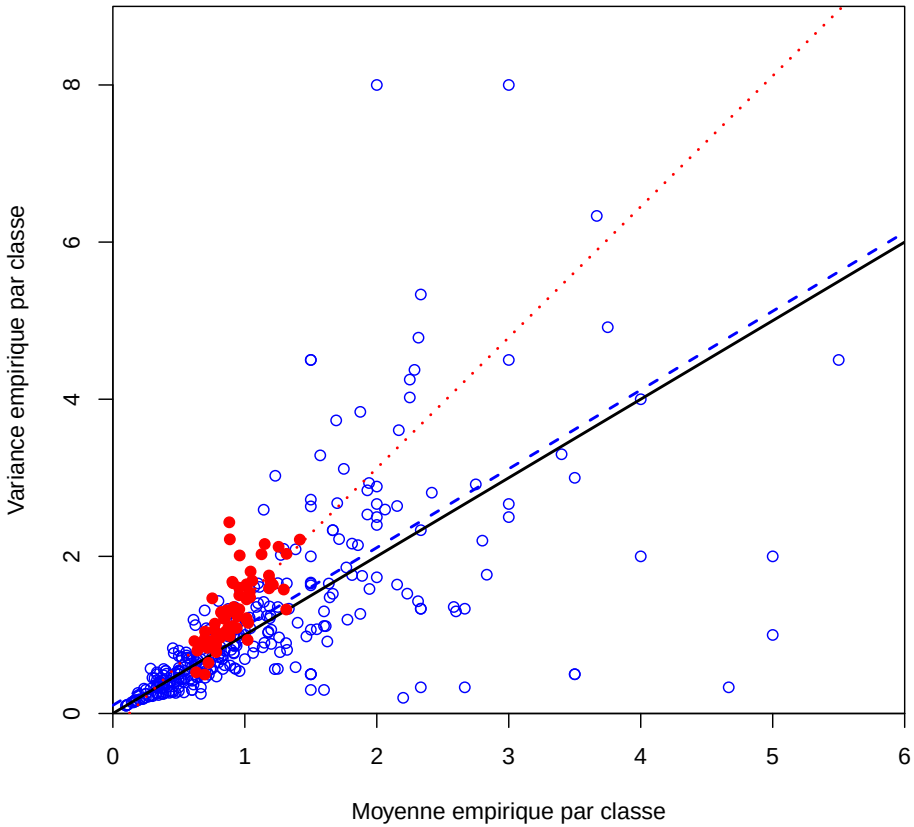


FIGURE 3 – Empirical estimates of the relation between $\mathbb{E}(Z|\mathbf{X} = \mathbf{x})$ and $\text{var}(Z|\mathbf{X} = \mathbf{x})$ (solid line is $y = x$, i.e. equidispersed case).

Overdispersion

Remark 16

Why is overdispersion a problem ?

If we fit a Poisson model to overdispersed data, we will underestimate the variance of the estimators of the β_j

$\Rightarrow$ underestimate "standard errors" that appear in the denominator of the Wald test and in confidence interval formulas

$\Rightarrow$ non-significant explanatory variables may appear to be significant, incorrect coverage probabilities

Poisson vs quasi-Poisson

Recall Poisson regression model :

$$Z_i|\mathbf{X}_i \sim \mathcal{P}(e^{\beta^\top \mathbf{X}_i}) \tag{7}$$

Then

$$\mathbb{P}(Z_i = z|\mathbf{X}_i) = \exp(-e^{\beta^\top \mathbf{X}_i})\frac{e^{z\beta^\top \mathbf{X}_i}}{z!}.$$

The MLE $\hat{\beta}_n$ of β is consistent and asymptotically normal.

Remark 17

In fact, $\hat{\beta}_n$ has these nice properties **even when the distribution of $Z_i|\mathbf{X}_i$ is NOT Poisson** provided that $\mathbb{E}(Z_i|\mathbf{X}_i) = e^{\beta^\top \mathbf{X}_i}$.

Poisson vs quasi-Poisson

Recall that the MLE $\widehat{\beta}_n$ in model (7) solves the equation :

$$S_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i(Z_i - e^{\beta^\top \mathbf{X}_i}) = 0.$$

By MLE theory, $\widehat{\beta}_n$ converges to the value β^* such that

$$S(\beta^*) = \lim_{n\to\infty} S_n(\beta^*) = 0,$$

with $S(\beta) = \mathbb{E}\left[\mathbf{X}_i(Z_i - e^{\beta^\top \mathbf{X}_i})\right]$. Now,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X}_i(Z_i - e^{\beta^\top \mathbf{X}_i})\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{X}_i(Z_i - e^{\beta^\top \mathbf{X}_i})|\mathbf{X}_i\right]\right], \\
&= \mathbb{E}\left[\mathbf{X}_i \mathbb{E}[Z_i - e^{\beta^\top \mathbf{X}_i}|\mathbf{X}_i]\right], \\
&= \mathbb{E}\left[\mathbf{X}_i \left\{\mathbb{E}\left[Z_i|\mathbf{X}_i\right] - e^{\beta^\top \mathbf{X}_i}\right\}\right],
\end{aligned}$$

which is 0 if $\mathbb{E}\left[Z_i|\mathbf{X}_i\right] = e^{\beta^\top \mathbf{X}_i}$.

Poisson vs quasi-Poisson

Conclusion :

- consistency of the MLE holds provided the conditional mean $\mathbb{E}\left[Z_i|\mathbf{X}_i\right]$ is correctly specified as $e^{\beta^\top \mathbf{X}_i}$,
- equivalently said : suppose that $Z_i|\mathbf{X}_i \sim G_i$; then as long as $\mathbb{E}_{G_i}[Z_i|\mathbf{X}_i] = e^{\beta^\top \mathbf{X}_i}$, $\widehat{\beta}_n$ will be consistent,
- given this robustness to distributional assumption, we can continue to use the MLE even if the true distribution of Z_i is not Poisson, as long as $\mathbb{E}(Z_i|\mathbf{X}_i) = e^{\beta^\top \mathbf{X}_i}$.

Quasi-Poisson model

Quasi-Poisson model is defined by the first two moments of the distribution of $Z_i|\mathbf{X}_i$:

$$\forall i = 1, \dots, n, \quad \begin{cases} Z_i|\mathbf{X}_i \sim G_i \\ \mathbb{E}(Z_i|\mathbf{X}_i) = e^{\beta^\top \mathbf{X}_i} \\ \text{var}(Z_i|\mathbf{X}_i) = \omega_i \end{cases}$$

Remark 18

We do NOT assume equidispersion, i.e. we do not assume that $\omega_i = \mathbb{E}(Z_i|\mathbf{X}_i)$.

Quasi-maximum likelihood

Parameter β is estimated by solving the same (likelihood) equation as in Poisson regression (even if G_i is left unspecified) :

$$\sum_{i=1}^{n} \mathbf{X}_i (Z_i - e^{\beta^\top \mathbf{X}_i}) = 0. \tag{8}$$

Remark 19 (Terminology)

The resulting estimator $\hat{\beta}_n^Q$ is called quasi-MLE [a] : $\hat{\beta}_n^Q$ is like the Poisson MLE in that Poisson model motivates equation (8), but is unlike the Poisson MLE in that the underlying density of Z_i need not be Poisson.

a. Wedderburn, 1974

Quasi-maximum likelihood

Suppose that $\mu_i = \mathbb{E}(Z_i|\mathbf{X}_i) = e^{\beta^\top \mathbf{X}_i}$. Then

$$\widehat{\beta}_n^Q \overset{asymp.}{\sim} \mathcal{N}(\beta, \mathsf{var}(\widehat{\beta}_n^Q))$$

where

$$\mathsf{var}(\widehat{\beta}_n^Q) = \left(\sum_{i=1}^n \mu_i \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1} \left(\sum_{i=1}^n \omega_i \mathbf{X}_i \mathbf{X}_i^\top\right) \left(\sum_{i=1}^n \mu_i \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1}.$$

Remark 20

In the special case where $\mu_i = \mathbb{E}\left[Z_i|\mathbf{X}_i\right] = \mathsf{var}(Z_i|\mathbf{X}_i) = \omega_i$ (i.e., Poisson case), $\mathsf{var}(\hat{\beta}_n^Q)$ becomes :

$$\mathsf{var}(\hat{\beta}_n^Q) = \left(\sum_{i=1}^n \mu_i \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1},$$

which is the covariance matrix of the MLE in a Poisson model.

Quasi-maximum likelihood

To make inference on β, we need to estimate $\text{var}(\hat{\beta}_n^Q)$, which depends on $\mu_i = \mathbb{E}(Z_i|\mathbf{X}_i)$ and

$$\omega_i = \text{var}(Z_i|\mathbf{X}_i) = \mathbb{E}[(Z_i - \mu_i)^2|\mathbf{X}_i]$$

We can estimate μ_i by $\hat{\mu}_i = e^{\hat{\beta}_n^{Q\top}\mathbf{X}_i}$ and ω_i by $(Z_i - \hat{\mu}_i)^2$ and finally, $\text{var}(\hat{\beta}_n^Q)$ by :

$$\widehat{\text{var}(\hat{\beta}_n^Q)} = \left(\sum_{i=1}^n \hat{\mu}_i \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1} \left(\sum_{i=1}^n (Z_i - \hat{\mu}_i)^2 \mathbf{X}_i \mathbf{X}_i^\top\right) \left(\sum_{i=1}^n \hat{\mu}_i \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1}$$

called the sandwich [6] or robust estimator.

$\hookrightarrow$ bootstrap as an alternative

6. since it is of the form $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, with $\mathbf{B}$ "sandwiched" between $\mathbf{A}^{-1}$

Quasi-maximum likelihood

Another approach is to assume that ω_i is of the specific form :

$$\omega_i = \text{var}(Z_i|\mathbf{X}_i) = \phi\mathbb{E}(Z_i|\mathbf{X}_i) = \phi\mu_i, \tag{9}$$

with ϕ an unknown dispersion parameter.

Remark 21

- Overdispersion corresponds to the case where $\phi > 1$. The opposite case of underdispersion, $\phi < 1$, is also possible, though seemingly rare in practice.
- The advantage of (9) is that it provides a numerical value that quantifies the overdispersion.

Quasi-maximum likelihood

If $\omega_i = \phi\mu_i$, then $\text{var}(\widehat{\beta}_n^Q)$ simplifies to

$$\text{var}(\widehat{\beta}_n^Q) = \phi \left(\sum_{i=1}^{n} \mu_i \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1} = \phi \, \text{var}(\widehat{\beta}_n^P).$$

The variance of $\hat{\beta}_n^Q$ is increased by a factor ϕ relative to the variance of $\hat{\beta}_n$ in the Poisson model $\Rightarrow$ a correction factor of $\sqrt{\phi}$ should be introduced in the usual Poisson ML inference.

Failure to do so may yield

- wrong conclusions on statistical significance of regressors
- confidence intervals with wrong coverage probability

However, ϕ is usually unknown and has to be estimated.

Estimating the dispersion parameter ϕ

Since

$$\phi = \frac{\text{var}(Z_i|\mathbf{X}_i)}{\mathbb{E}(Z_i|\mathbf{X}_i)} = \mathbb{E}\left[\frac{(Z_i - \mu_i)^2}{\mu_i}|\mathbf{X}_i\right],$$

it is natural to estimate ϕ by :

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Z_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, \tag{10}$$

and $\text{var}(\hat{\beta}_n^Q)$ by $\widehat{\text{var}(\hat{\beta}_n^Q)} = \hat{\phi}\left(\sum_{i=1}^n \hat{\mu}_i \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1}$.

Remark 22

Fitting a quasi-Poisson model is equivalent to : i) fitting a Poisson model ii) then correcting the variance and standard error terms by ϕ and $\sqrt{\phi}$ respectively.

A likelihood view of quasi-Poisson model

Consider the model defined by the density

$$f(z_i;\beta,\phi)=\exp\left(\frac{z_i\beta^\top\mathbf{X}_i-e^{\beta^\top\mathbf{X}_i}}{\phi}+c(z_i,\phi)\right). \tag{11}$$

Then $\mu_i=\mathbb{E}(Z_i|\mathbf{X}_i)=e^{\beta^\top\mathbf{X}_i}$ and $\omega_i=\text{var}(Z_i|\mathbf{X}_i)=\phi\mu_i$, which is the variance in quasi-Poisson model.

The MLE $\widehat{\beta}_n^{MLE}$ of β in (11) solves :

$$\sum_{i=1}^{n}\mathbf{X}_i(Z_i-e^{\beta^\top\mathbf{X}_i})=0,$$

which coincides with the estimating equation in quasi-Poisson model. Thus :

$$\widehat{\beta}_n^{MLE}=\widehat{\beta}_n^{Q}$$

A likelihood view of quasi-Poisson model

We can define a density for quasi-Poisson model $\Rightarrow$ one may wish to estimate also ϕ by ML. But differentiating w.r.t. ϕ requires an expression for $c(y_i, \phi)$, and the condition

$$\sum_{z_i=0}^{\infty} f(z_i; \beta, \phi) = \sum_{z_i=0}^{\infty} \exp\left(\frac{z_i \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i}}{\phi} + c(z_i, \phi)\right) = 1$$

entails no explicit expression for $c(y_i, \phi)$.

$\Rightarrow$ we must return to $\widehat{\phi}$ given by (10), whose construction only requires assuming $\omega_i = \phi \mu_i$

$\Rightarrow$ defining a density for quasi-Poisson model is ultimately not useful

Generalized Poisson distribution

An alternative remedy to overdispersion is to consider distributions that do *not* impose mean/variance equality.

E.g., Consul and Jain (1970) introduce the generalized Poisson distribution $GP(\lambda, \varphi)$ with pmf :

$$\mathbb{P}(Z = z) = \frac{\lambda(\lambda + (\varphi - 1)z)^{z-1}}{z!} \varphi^{-z} e^{-\frac{1}{\varphi}(\lambda + (\varphi - 1)z)}, \quad z = 0, 1, \ldots$$

Remark 23

- The $GP(\lambda, \varphi)$ reduces to Poisson $\mathcal{P}(\lambda)$ when $\varphi = 1$.
- We have : $\text{var}(Z) > \mathbb{E}(Z)$ when $\varphi > 1$ and $\text{var}(Z) < \mathbb{E}(Z)$ when $\varphi < 1 \Rightarrow$ allows for over- and underdispersion.
- Implemented in various `R` packages (under various parameterizations) : `VGAM`, `HMMpa`,...

Negative binomial model

But the most widely used overdispersed distribution is the negative binomial (NB), which can be seen as a Poisson-gamma mixture.

Let $U_i \sim G(1, \nu)$ and suppose

$$Z_i | U_i \sim \mathcal{P}(\mu_i U_i) \text{ with } \mu_i > 0.$$

Observe that

$$\mathbb{E}(Z_i) = \mathbb{E}[\mathbb{E}[Z_i|U_i]] = \mathbb{E}[\mu_i U_i] = \mu_i \mathbb{E}[U_i] = \mu_i.$$

Remark 24

In practice, U_i is unobserved. It allows to introduce heterogeneity between individuals by adding variability to the mean value of their response.

Negative binomial model

$$\begin{aligned}
\mathbb{P}(Z_i = z) &= \int_0^\infty e^{-\mu_i u} \frac{(\mu_i u)^z}{z!} \frac{\nu^\nu}{\Gamma(\nu)} u^{\nu-1} e^{-u\nu}\, du \\
&= \frac{\nu^\nu \mu_i^z}{\Gamma(\nu) z!} \int_0^\infty e^{-(\mu_i+\nu)u} u^{z+\nu-1}\, du \\
&= \frac{\nu^\nu \mu_i^z}{\Gamma(\nu) z!} \frac{\Gamma(z+\nu)}{(\mu_i+\nu)^{z+\nu}} \\
&= \frac{\Gamma(z+\nu)}{\Gamma(\nu) z!} \left(\frac{\mu_i}{\nu+\mu_i}\right)^z \left(\frac{\nu}{\nu+\mu_i}\right)^\nu .
\end{aligned}$$

Setting $\kappa = 1/\nu$ and $\mu = \mu_i$, we recognize the density of the NB distribution, and :

$$\mathbb{E}(Z_i) = \mu_i$$

and

$$\text{var}(Z_i) = \mu_i(1 + \kappa\mu_i) = \mu_i + \kappa\mu_i^2 > \mathbb{E}(Z_i).$$

Negative binomial model

Remark 25

- A regression model for Z_i is obtained by setting

$$g(\mu_i) = \beta^\top \mathbf{X}_i$$

  Since $\mu_i > 0$, we typically choose $\ln(\mu_i) = \beta^\top \mathbf{X}_i$.
- Log is not the canonical link in NB. If we used the canonical link, the constraint $\mu_i > 0$ would impose $\beta^\top \mathbf{X}_i < 0$, which restricts the estimates of β.

  $\Rightarrow$ log link is preferred, as in Poisson regression.

β and κ are estimated by MLE (no closed-form estimates). MLEs are consistent and asymptotically Gaussian.

NB1 and NB2 variance functions

The variance function $\text{var}(Z_i) = \mu_i + \kappa\mu_i^2$ is quadratic in μ_i (hence the name "NB2" model).

Another common parametrization is obtained with $U_i \sim G(\mu_i, \phi\mu_i)$ and $Z_i|U_i \sim \mathcal{P}(U_i)$, with $\phi > 0 \Rightarrow$ NB distribution with

$$\mathbb{E}(Z_i) = \mu_i \text{ and } \text{var}(Z_i) = \mu_i \left(1 + \frac{1}{\phi}\right).$$

The variance function is linear in μ_i (hence the name "NB1")[7].

Remark 26

In NB1, the ratio $\text{var}(Z_i)/\mathbb{E}(Z_i)$ is $1 + \frac{1}{\phi}$ for all $i \Rightarrow$ NB1 describes a form of overdispersion that is constant over individuals. In NB2, the ratio depends on μ_i.

7. a regression model is constructed by setting $\mu_i = e^{\beta^\top \mathbf{x}_i}$

Overdispersion tests

Equidispersion holds when $\text{var}(Y_i|\mathbf{X}_i) = \mu_i$, where $\mu_i = \mathbb{E}[Y_i|\mathbf{X}_i]$.

Several tests available to test the null hypothesis of equidispersion against the alternative of overdispersion :

- tests based on NB1 and NB2 variance functions
- a LRT test (view Poisson model as a special case of NB2 when $\kappa = 0$). Define :

$$LRT = 2(\ell_n(\hat{\beta}_n, \hat{\kappa}_n) - \ell_n(\hat{\beta}_n))$$

where $\ell_n(\hat{\beta}_n, \hat{\kappa}_n)$ and $\ell_n(\hat{\beta}_n)$ are the max log-likelihoods of NB2 and Poisson models respectively.

Overdispersion tests

Under $H_0 : \kappa = 0$ (i.e. assuming a Poisson model),

$$LRT \overset{asymp.}{\sim} \frac{1}{2}\delta_0 + \frac{1}{2}\chi^2_1.$$

Remark 27

This distribution arises because $\kappa \geq 0$ (H_0 lies on the boundary of the parameter space).

$\Rightarrow$ to test H_0 at the level α, use $\chi^2_1(1-2\alpha)$ as critical value. Why ?

Overdispersion tests

Let $\mathcal{W} := \frac{1}{2}\delta_0 + \frac{1}{2}\chi^2_1$ and $c > 0$. Then :

$$\begin{aligned}
\mathbb{P}(\mathcal{W} > c) &= \mathbb{P}(\mathcal{W} > c|\mathcal{W} = 0)\mathbb{P}(\mathcal{W} = 0) \\
&\quad + \mathbb{P}(\mathcal{W} > c|\mathcal{W} > 0)\mathbb{P}(\mathcal{W} > 0) \\
&= 0 \times \frac{1}{2} + \mathbb{P}(\mathcal{W} > c|\mathcal{W} > 0) \times \frac{1}{2} \\
&= \frac{1}{2}\mathbb{P}(\chi^2_1 > c).
\end{aligned}$$

Let $c^{\mathcal{W}}_{1-\alpha}$ be the $(1-\alpha)$-quantile of $\mathcal{W}$, i.e., $\mathbb{P}(\mathcal{W} > c^{\mathcal{W}}_{1-\alpha}) = \alpha$. Then

$$\mathbb{P}(\chi^2_1 > c^{\mathcal{W}}_{1-\alpha}) = 2\mathbb{P}(\mathcal{W} > c^{\mathcal{W}}_{1-\alpha}) = 2\alpha,$$

and thus, $c^{\mathcal{W}}_{1-\alpha}$ is the $(1-2\alpha)$-quantile of χ^2_1, namely $\chi^2_1(1-2\alpha)$.

Overdispersion tests based on NB variance functions

Under the alternative $\omega_i = \mu_i + \kappa\mu_i$, we have :

$$\frac{\omega_i - \mu_i}{\mu_i} = \kappa.$$

Basic idea is to :

- fit a Poisson model to the data, estimate μ_i by $\hat{\mu}_i = e^{\hat{\beta}_n^\top \mathbf{X}_i}$ and $(\omega_i - \mu_i)/\mu_i$ by :

$$W_i = \frac{(Z_i - \hat{\mu}_i)^2 - Z_i}{\hat{\mu}_i}$$

- estimate the linear model $W_i = \kappa + \epsilon_i$
- test $H_0 : \kappa = 0$ (Wald test)

Remark 28

Under the alternative $\omega_i = \mu_i + \kappa\mu_i^2$, we have : $(\omega_i - \mu_i)/\mu_i = \kappa\mu_i$.
Same idea as above : test $\kappa = 0$ in $W_i = \kappa\mu_i + \epsilon_i$.

Overdispersion tests

Remark 29

- In `R`, tests based on the NB variance functions are available in the `dispersiontest` function (`AER` package). LRT is implemented by the `odTest` function (`pscl` package).
- One advantage of NB model relative to quasi-Poisson is that it is associated with a formal likelihood so that information criteria (such as AIC) are readily available.

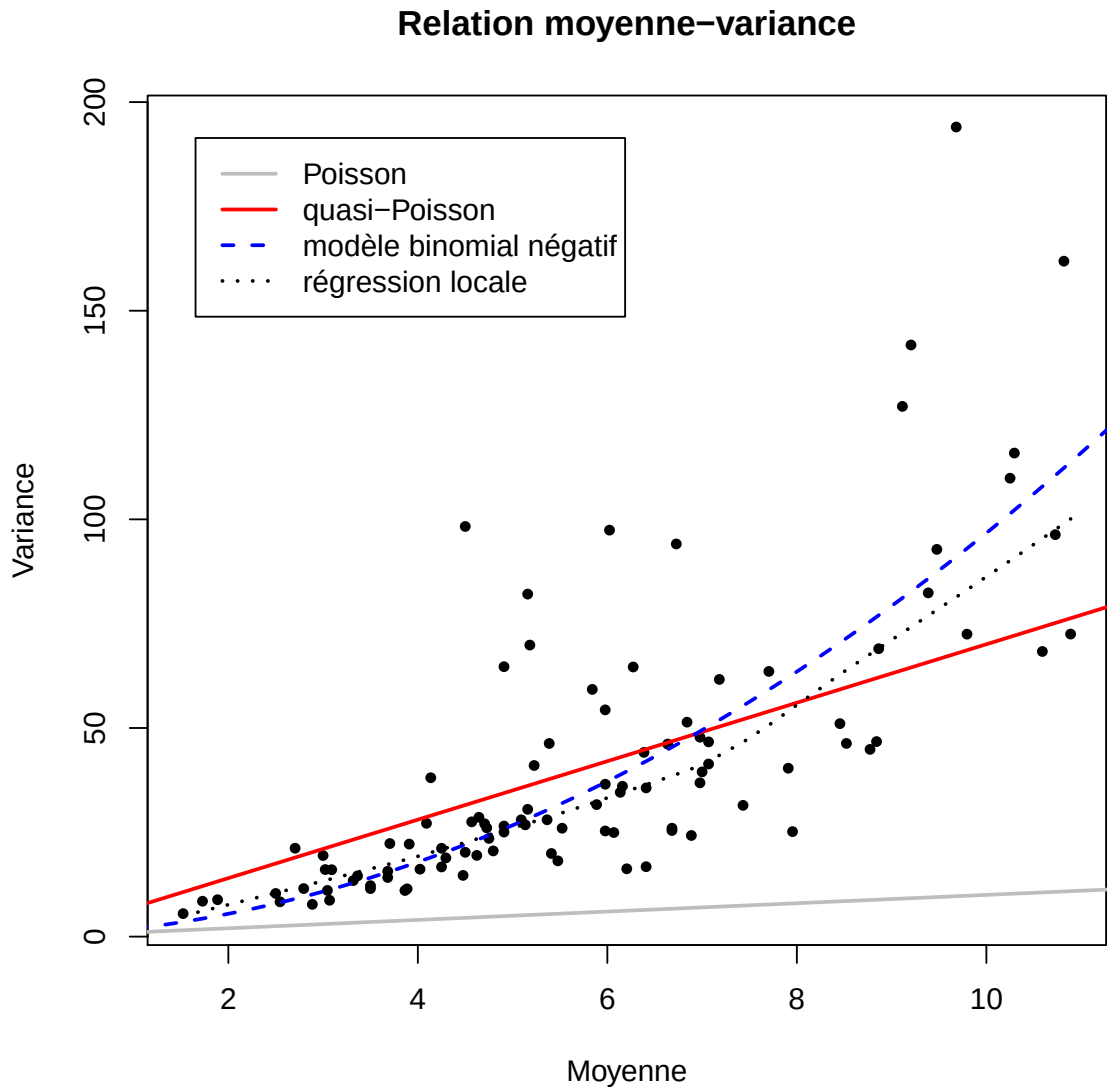
Choosing between quasi-Poisson, NB,...

Friendly et al. (2015) propose a graphical display based on the estimated mean-variance relation :

1. fit a NB regression model and obtain fitted response values,
2. divide the range of fitted values [8] in sub-intervals (according to centiles, for example),
3. obtain the empirical mean and variance of the count response within each sub-interval,
4. plot the (mean, variance) pairs,
5. plot the estimated mean-variance relation from Poisson, quasi-Poisson, NB... and eventually, a local regression estimate,
6. choose the model with best fit.

8. or equivalently, of linear predictors

Choosing between quasi-Poisson, NB,...



A non-exhaustive bibliography

- Cameron A.C., Trivedi P.K. *Regression analysis of count data*. Cambridge University Press, 1998.
- Dupuy J.-F. *Méthodes statistiques pour l'analyse de données de comptage sur-dispersées*. ISTE Ltd, 2018.
- Friendly M., Meyer D. *Discrete Data Analysis with R : Visualization and Modeling Techniques for Categorical and Count Data*, Chapman and Hall/CRC, 2015.
- Hilbe J.M. *Negative Binomial Regression*. Cambridge University Press, 2011.
- Kleiber C., Zeileis A. *Applied Econometrics with R*. Springer, 2008.