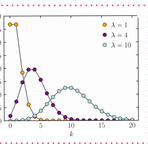
Poisson regression

If Y_i is a count variable, with values in $R_{Y_i} = \{0, 1, 2, \dots\}$, assuming a Gaussian distribution is not adequate.

The most common distribution for a count variable is the Poisson.

Recall that:

- $Y_i \sim \text{Poisson}(\lambda_i) \quad \lambda_i > 0$
- parameter space: $\Theta = (0, +\infty)$
- support: $Y_i = R_{Y_i} = \{0, 1, 2, \dots\}$
- probability mass function: $p(y_i|\lambda_i) = P(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$
- moments: $\mathbb{E}[Y_i] = \lambda_i$, $\text{var}(Y_i) = \lambda_i$



Poisson regression: assumptions

1. $Y_i \sim \text{Poisson}(\lambda_i)$ independent for $i = 1, \dots, n$
2. $\eta_i = \vec{x}_i^T \vec{\beta}$
3. $\log(\lambda_i) = \eta_i$ LOGARITHIC LINK FUNCTION "log-linear model"

Remarks:

- the log link allows mapping the linear predictor $\eta_i = \vec{x}_i^T\vec{\beta} \in \mathbb{R}$ to $(0, +\infty)$, the parameter space of λ_i:
  indeed $\log(\lambda_i) = \eta_i \Rightarrow \lambda_i = e^{\eta_i} = e^{\vec{x}_i^T\vec{\beta}} > 0$
  We could also use other link functions, however, the log link leads to better theoretical properties (it is the "canonical" link).
- non-constant variance: the Poisson distribution assumes that $\text{var}(Y_i) = \mathbb{E}[Y_i]$
  Hence $\text{var}(Y_i) = \lambda_i$ (different between units, by construction).

The distribution of Y_i hence is

$$P(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \qquad \log(\lambda_i) = \vec{x}_i^T\vec{\beta} \Rightarrow \lambda_i = e^{\vec{x}_i^T\vec{\beta}}$$

$$= \frac{e^{-e^{\vec{x}_i^T\vec{\beta}}} e^{\vec{x}_i^T\vec{\beta} y_i}}{y_i!}$$

Interpretation of the model parameters

Let's study the mean $\mathbb{E}[Y_i]$ for two individuals i and k with all the covariates equal except the j-th one, for which we assume $x_{ij} = x_{kj} + \varepsilon$

i.e., $x_{ih} = x_{kh}$ for $h = 1, \dots, p$ $h \neq j$; $x_{ij} = x_{kj} + \varepsilon$

For individual i we get

$$\mathbb{E}[Y_i] = \lambda_i = e^{\vec{x}_i^T\vec{\beta}} = \exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{ij-1} + \beta_j x_{ij} + \beta_{j+1} x_{ij+1} + \dots + \beta_p x_{ip}\}$$

For individual k we get

$$\mathbb{E}[Y_k] = \lambda_k = e^{\vec{x}_k^T\vec{\beta}} = \exp\{\beta_1 + \beta_2 x_{k2} + \dots + \beta_{j-1} x_{kj-1} + \beta_j (x_{ij} + \varepsilon) + \beta_{j+1} x_{kj+1} + \dots + \beta_p x_{kp}\}$$

$$= \exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{ij-1} + \beta_j x_{ij} + \beta_j \varepsilon + \beta_{j+1} x_{ij+1} + \dots + \beta_p x_{ip}\}$$

If we study the ratio

$$\frac{\mathbb{E}[Y_i]}{\mathbb{E}[Y_k]} = \frac{\lambda_i}{\lambda_k} = \frac{\exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{ij-1} + \beta_j (x_{ij} + \varepsilon) + \beta_{j+1} x_{ij+1} + \dots + \beta_p x_{ip}\}}{\exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{ij-1} + \beta_j x_{ij} + \beta_j \varepsilon + \beta_{j+1} x_{ij+1} + \dots + \beta_p x_{ip}\}}$$

$$= \exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{ij-1} + \beta_j (x_{ij} + \varepsilon) + \beta_{j+1} x_{ij+1} + \dots + \beta_p x_{ip} - \beta_1 - \beta_2 x_{i2} - \dots - \beta_{j-1} x_{ij-1} - \beta_j x_{ij} - \beta_{j+1} x_{ij+1} - \dots - \beta_p x_{ip}\}$$

$$= \exp\{\beta_j (x_{ij} + \varepsilon) - \beta_j x_{ij}\} = \exp\{\beta_j \varepsilon\}$$

(all terms except the j-th simplify since we assumed $x_{ih} = x_{kh}$ for $h \neq j$)

$$\Rightarrow \frac{\lambda_i}{\lambda_k} = e^{\beta_j}$$

$$\Rightarrow \beta_j = \log\frac{\lambda_i}{\lambda_k} = \log\lambda_i - \log\lambda_k = \log\mathbb{E}[Y_i \mid x_j = x_{ij} + \varepsilon] - \log\mathbb{E}[Y_i \mid x_j = x_{ij}]$$

The parameter β_j represents the difference in the log of the expected counts if we increase x_j of ε unit, while keeping the other covariates fixed.

or, if we write $e^{\beta_j} = \frac{\lambda_i}{\lambda_k}$ $\Rightarrow$ $\lambda_i = \lambda_k \cdot e^{\beta_j}$ $\Rightarrow$ $\mathbb{E}[Y_i \mid x_j = x_{ij} + \varepsilon] = \mathbb{E}[Y_i \mid x_j = x_{ij}] \cdot e^{\beta_j}$

The expected counts change of a multiplicative factor e^{β_j} if we increase the j-th covariate of ε unit, while keeping the other covariates fixed.

- ESTIMATION

data $(y_1, \dots, y_n)$ from $Y_i \sim \text{Pois}(\lambda_i) = \text{Pois}(e^{\vec{x}_i^T\vec{\beta}})$ indep.

joint density

$$p(y_1, \dots, y_n) = \prod_{i=1}^n p(y_i) = \prod_{i=1}^n \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \prod_{i=1}^n \frac{e^{-e^{\vec{x}_i^T\vec{\beta}}} e^{\vec{x}_i^T\vec{\beta} y_i}}{y_i!} = \frac{e^{-\sum_i e^{\vec{x}_i^T\vec{\beta}}} e^{\sum_i \vec{x}_i^T\vec{\beta} y_i}}{\prod_i y_i!}$$

likelihood

$$L(\vec{\beta}) \propto \prod_{i=1}^n p(y_i; \vec{\beta}) = e^{-\sum_i e^{\vec{x}_i^T\vec{\beta}}} e^{\sum_i \vec{x}_i^T\vec{\beta} y_i}$$

log-likelihood

$$\ell(\vec{\beta}) = -\sum_i e^{\vec{x}_i^T\vec{\beta}} + \sum_i y_i \vec{x}_i^T\vec{\beta}$$

score function

$$g_n(\vec{\beta}) = \left\{\frac{\partial \ell(\vec{\beta})}{\partial \beta_r}\right\}_{r=1,\dots,p}$$

$$\frac{\partial}{\partial \beta_r}\ell(\vec{\beta}) = -\sum_i x_{ir} e^{\vec{x}_i^T\vec{\beta}} + \sum_i y_i x_{ir} = \sum_i x_{ir}(y_i - e^{\vec{x}_i^T\vec{\beta}})$$

Hence the score function can be written as a function of the entire vector $\vec{\beta}$ as:

$$\frac{\partial}{\partial \vec{\beta}}\ell(\vec{\beta}) = -\sum_i \vec{x}_i e^{\vec{x}_i^T\vec{\beta}} + \sum_i \vec{x}_i y_i = \sum_i \vec{x}_i (y_i - e^{\vec{x}_i^T\vec{\beta}}) = X^T(\vec{y} - \vec{\lambda})$$

The MLE $\hat{\vec{\beta}}$ is the solution of the equation $\ell_n(\vec{\beta}) = 0$

$\Rightarrow$ solution of $X^T(\vec{y} - \vec{\lambda}) = \vec{0}$, i.e. $X^T(\vec{y} - e^{X\vec{\beta}}) = 0$

(it resembles the normal equations in the Gaussian LM, however, here $\vec{\lambda}$ is a non-linear function of $\vec{\beta}$)

This equation does not have an analytical solution: the maximum is found numerically using iterative optimization methods.

Hence we do not have a closed-form expression for the MLE $\hat{\vec{\beta}}$.

Remark:

notice that, similarly to the LM, since $\hat{\vec{\beta}}$ is the solution of the equation, we obtain

$$X^T(\vec{y} - e^{X\hat{\vec{\beta}}}) = \vec{0}$$

$$\Rightarrow X^T(\vec{y} - \hat{\vec{\lambda}}) = \vec{0}$$

$$\begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_n^T \end{bmatrix}^T \cdot (\vec{y} - \hat{\vec{\lambda}}) = \begin{bmatrix} \sum_i x_{i1}(y_i - \hat{\lambda}_i) \\ \vdots \\ \sum_i x_{ip}(y_i - \hat{\lambda}_i) \end{bmatrix} = \vec{0}$$

$p \times n$

If the model includes the intercept $\Rightarrow$ $x_{i1} = 1 \; \forall i$

$$\Rightarrow \sum_i (y_i - \hat{\lambda}_i) = \sum_i (y_i - \hat{\lambda}_i) \cdot 1 = 0$$

second derivative

$$g_{nn}(\vec{\beta}) = \left\{\frac{\partial^2 \ell(\vec{\beta})}{\partial \beta_r \partial \beta_s}\right\}_{r,s=1,\dots,p} = -\sum_i x_{ir} x_{is} e^{\vec{x}_i^T\vec{\beta}} = -\sum_i x_{ir} x_{is} \lambda_i$$

In matrix form we get $g_{nn}(\vec{\beta}) = -X^T U X$ with U an $n \times n$ diagonal matrix

$$U = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = \text{diag}\{\lambda_1, \dots, \lambda_n\} = \text{diag}\{e^{\vec{x}_1^T\vec{\beta}}, \dots, e^{\vec{x}_n^T\vec{\beta}}\}$$

$\Rightarrow$ it is a function of $\vec{\beta}$ $\Rightarrow$ $U = U(\vec{\beta})$

The observed information evaluated at the MLE $\hat{\vec{\beta}}$ is

$$j(\hat{\vec{\beta}}) = -g_{nn}(\vec{\beta})\big|_{\hat{\vec{\beta}}} = X^T U(\hat{\vec{\beta}}) X$$

where $U(\hat{\vec{\beta}}) = \text{diag}\{e^{\vec{x}_1^T\hat{\vec{\beta}}}, \dots, e^{\vec{x}_n^T\hat{\vec{\beta}}}\}$

- INFERENCE

Inference here is based on approximate distributions

Remarks:

- notation: $Y \sim p(\cdot)$ means "Y is approximately distributed as $p(\cdot)$"
- approximation improves as n increases (good approximation if the sample size is large)

- DISTRIBUTION of the MAXIMUM LIKELIHOOD ESTIMATOR of the REGRESSION PARAMETERS

$$\hat{\vec{\beta}} \sim N_p(\vec{\beta}, j(\hat{\vec{\beta}})^{-1})$$

the marginal distribution for the j-th element is $\hat{\beta}_j \sim N(\beta_j, \mathbb{E}[j(\hat{\vec{\beta}})^{-1}]_{jj})$ $\quad j = 1, \dots, p$

(element of the matrix $j(\hat{\vec{\beta}})^{-1}$ in position (j,j))

- CONFIDENCE INTERVAL FOR β_j

A pivotal quantity is

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{[j(\hat{\vec{\beta}})^{-1}]_{jj}}} \sim N(0,1)$$

a confidence interval with level $(1-\alpha)$ for β_j $(j = 1, \dots, p)$ can be obtained as

$$P\left(z_{\frac{\alpha}{2}} < \frac{\hat{\beta}_j - \beta_j}{\sqrt{[j(\hat{\vec{\beta}})^{-1}]_{jj}}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

(Gaussian is symmetric: $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$; quantile of level $1-\frac{\alpha}{2}$ of a $N(0,1)$)

with the steps:

$$\hat{\beta}_j - \sqrt{[j(\hat{\vec{\beta}})^{-1}]_{jj}} \cdot z_{1-\frac{\alpha}{2}} < \beta_j < \hat{\beta}_j + \sqrt{[j(\hat{\vec{\beta}})^{-1}]_{jj}} \cdot z_{1-\frac{\alpha}{2}}$$

$$\Rightarrow \beta_j \in \hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \sqrt{[j(\hat{\vec{\beta}})^{-1}]_{jj}}$$

- TEST ABOUT β_j

consider the test $\begin{cases} H_0: \beta_j = b_j \\ H_1: \beta_j \neq b_j \end{cases}$

we can use the Wald statistic

$$Z_j = \frac{\hat{\beta}_j - b_j}{\sqrt{[j(\hat{\vec{\beta}})^{-1}]_{jj}}} \sim N(0,1) \text{ under } H_0$$

the observed value of the test is z_j^{obs}

- if we use a fixed significance level α:
  $\alpha = P_{H_0}(|Z_j| > z_{1-\frac{\alpha}{2}})$
  $\Rightarrow$ reject region is $R = (-\infty, -z_{1-\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, +\infty)$



- if we use the observed significance level:
  the p-value is $\alpha^{obs} = P_{H_0}(|Z_j| \geq |z_j^{obs}|) = 2\left(1 - \Phi(|z_j^{obs}|)\right)$