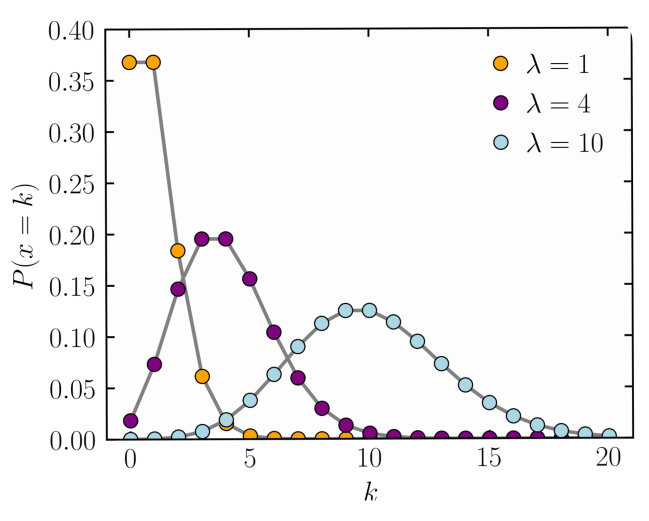
21 Nov - LEC 10

POISSON REGRESSION

If Y_i is a count variable, with values in $\mathbb{N}_0 = \{0, 1, 2, \dots\}$, assuming a Gaussian distribution is not adequate.
The most common distribution for a count variable is the Poisson.

Recall that:
$Y \sim \text{Poisson}(\mu) \qquad \mu > 0$
- parameter space: $\Theta = (0, +\infty)$
- support: $\mathcal{Y} = \mathbb{N}_0 = \{0, 1, 2, \dots\}$
- probability mass function $p(y; \mu) = \mathbb{P}(Y = y) = \dfrac{e^{-\mu} \mu^y}{y!}$
- moments: $\mathbb{E}[Y] = \mu$
  $\text{var}(Y) = \mu$



POISSON REGRESSION : ASSUMPTIONS
1. $Y_i \sim \text{Poisson}(\mu_i)$ independent for $i = 1, \dots, n$
2. $\eta_i = \tilde{\underline{x}}_i^T \underline{\beta}$
3. $\log(\mu_i) = \eta_i$ LOGARITHMIC LINK FUNCTION "log-linear model"

Remarks:
- the log link allows mapping the linear predictor $\eta_i = \tilde{\underline{x}}_i^T \underline{\beta} \in \mathbb{R}$ to $(0, +\infty)$, the parameter space of μ_i
  indeed $\log(\mu_i) = \eta_i \Rightarrow \mu_i = e^{\eta_i} = e^{\tilde{\underline{x}}_i^T \underline{\beta}} > 0$
  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 $\text{var}(Y_i) = \mu_i$ (different between units, by construction).

The distribution of Y_i hence is

$$\mathbb{P}(Y_i = y_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \qquad \log(\mu_i) = \tilde{\underline{x}}_i^T \underline{\beta} \Rightarrow \mu_i = e^{\tilde{\underline{x}}_i^T \underline{\beta}}$$
$$= \frac{e^{-e^{\tilde{x}_i^T \beta}} \, e^{\tilde{x}_i^T \beta y_i}}{y_i!}$$

INTERPRETATION OF THE MODEL PARAMETERS

Let's study the mean $\mathbb{E}[Y]$ for two individuals i and k with all the covariates equal except the j-th one, for which we assume $x_{kj} = x_{ij} + 1$
i.e, $x_{ih} = x_{kh}$ for $h = 1, \dots, p$ $h \neq j$, $x_{kj} = x_{ij} + 1$

For individual i we get
$$\mathbb{E}[Y_i] = \mu_i = e^{\tilde{\underline{x}}_i^T \underline{\beta}} = \exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{i,j-1} + \beta_j x_{ij} + \beta_{j+1} x_{i,j+1} + \dots + \beta_p x_{ip}\}$$

For individual k we get
$$\mathbb{E}[Y_k] = \mu_k = e^{\tilde{\underline{x}}_k^T \underline{\beta}} = \exp\{\beta_1 + \beta_2 x_{k2} + \dots + \beta_{j-1} x_{k,j-1} + \beta_j x_{kj} + \beta_{j+1} x_{k,j+1} + \dots + \beta_p x_{kp}\}$$
$$= \exp\{\beta_1 + \beta_2 x_{k2} + \dots + \beta_{j-1} x_{k,j-1} + \beta_j (x_{ij} + 1) + \beta_{j+1} x_{k,j+1} + \dots + \beta_p x_{kp}\}$$

If we study the RATIO
$$\frac{\mathbb{E}[Y_k]}{\mathbb{E}[Y_i]} = \frac{\mu_k}{\mu_i} = \frac{\exp\{\beta_1 + \beta_2 x_{k2} + \dots + \beta_{j-1} x_{k,j-1} + \beta_j (x_{ij} + 1) + \beta_{j+1} x_{k,j+1} + \dots + \beta_p x_{kp}\}}{\exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{i,j-1} + \beta_j x_{ij} + \beta_{j+1} x_{i,j+1} + \dots + \beta_p x_{ip}\}}$$
$$= \exp\{\cancel{\beta_1} + \cancel{\beta_2 x_{k2}} + \dots + \cancel{\beta_{j-1} x_{k,j-1}} + \beta_j (x_{ij} + 1) + \cancel{\beta_{j+1} x_{k,j+1}} + \dots + \cancel{\beta_p x_{kp}} -$$
$$- \cancel{\beta_1} - \cancel{\beta_2 x_{i2}} - \dots - \cancel{\beta_{j-1} x_{i,j-1}} - \beta_j x_{ij} - \cancel{\beta_{j+1} x_{i,j+1}} - \dots - \cancel{\beta_p x_{ip}}\}$$
$$= \exp\{\beta_j (x_{ij} + 1) - \beta_j x_{ij}\}$$
$$= \exp\{\beta_j x_{ij} + \beta_j - \beta_j x_{ij}\} = \exp\{\beta_j\}$$

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

$$\Rightarrow \frac{\mu_k}{\mu_i} = e^{\beta_j}$$

$$\Rightarrow \beta_j = \log \frac{\mu_k}{\mu_i} = \log \mu_k - \log \mu_i = \log \mathbb{E}[Y \mid x_j = x_{ij} + 1] - \log \mathbb{E}[Y \mid x_j = x_{ij}]$$

The parameter β_j represents the DIFFERENCE IN THE LOG OF THE EXPECTED COUNTS IF WE INCREASE x_j OF 1 UNIT, WHILE KEEPING THE OTHER COVARIATES FIXED.

or, if we write $e^{\beta_j} = \dfrac{\mu_k}{\mu_i} \Rightarrow \mu_k = \mu_i \cdot e^{\beta_j} \Rightarrow \mathbb{E}[Y \mid x_j = x_{ij} + 1] = \mathbb{E}[Y \mid x_j = x_{ij}] \cdot e^{\beta_j}$

The expected counts change of a MULTIPLICATIVE FACTOR e^{β_j} if we increase the j-th covariate of 1 unit, while keeping the other covariates fixed.

- ESTIMATION

data $(y_1, \dots, y_n)$ from $Y_i \sim \text{Pois}(\mu_i) = \text{Pois}(e^{\tilde{\underline{x}}_i^T \underline{\beta}})$ indep.
joint density
$$p(y_1, \dots, y_n) = \prod_{i=1}^n p(y_i) = \prod_{i=1}^n \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} =$$
$$= \prod_{i=1}^n \frac{e^{-e^{\tilde{x}_i^T \beta}} \, e^{\tilde{x}_i^T \beta y_i}}{y_i!} = \frac{e^{-\sum_{i=1}^n e^{\tilde{x}_i^T \beta}} \, e^{\sum_{i=1}^n \tilde{x}_i^T \beta y_i}}{\prod_{i=1}^n y_i!}$$

likelihood
$$L(\underline{\beta}) \propto \prod_{i=1}^n p(y_i; \underline{\beta}) \propto e^{-\sum_{i=1}^n e^{\tilde{x}_i^T \beta}} \, e^{\sum_{i=1}^n \tilde{x}_i^T \beta y_i}$$

log-likelihood
$$\ell(\underline{\beta}) = -\sum_{i=1}^n e^{\tilde{x}_i^T \beta} + \sum_{i=1}^n y_i \tilde{x}_i^T \beta$$

score function
$$\ell_*(\underline{\beta}) = \left\{ \frac{\partial}{\partial \beta_r} \ell(\beta) \right\}_{r = 1, \dots, p}$$
$$\frac{\partial}{\partial \beta_r} \ell(\underline{\beta}) = -\sum_{i=1}^n x_{ir} e^{\tilde{x}_i^T \beta} + \sum_{i=1}^n y_i x_{ir} = \sum_{i=1}^n x_{ir} (y_i - e^{\tilde{x}_i^T \beta})$$

Hence the score function can be written as a function of the entire vector $\underline{\beta}$ as:
$$\frac{\partial}{\partial \underline{\beta}} \ell(\beta) = -\sum_{i=1}^n \tilde{x}_i e^{\tilde{x}_i^T \beta} + \sum_{i=1}^n \tilde{x}_i y_i = \sum_{i=1}^n \tilde{x}_i (y_i - \underbrace{e^{\tilde{x}_i^T \beta}}_{\mu_i}) = X^T (\underline{y} - \underline{\mu})$$

The MLE $\hat{\underline{\beta}}$ is the solution of the equation $\ell_*(\underline{\beta}) = 0$
$\Rightarrow$ solution of $X^T (\underline{y} - \underline{\mu}) = \underline{0}$ → it resembles the normal equations in the Gaussian LM
$X^T (\underline{y} - e^{X\underline{\beta}}) = 0$ however, here $\underline{\mu}$ is a non-linear function of $\underline{\beta}$

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 $\hat{\underline{\beta}}$.

Remark:
notice that, similarly to the LM, since $\hat{\underline{\beta}}$ is the solution of the equation, we obtain
$$X^T (\underline{y} - e^{X\hat{\underline{\beta}}}) = \underline{0}$$
$$\Rightarrow X^T (\underline{y} - \hat{\underline{\mu}}) = \underline{0}$$
$$\underset{p \times n}{\begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_p^T \end{bmatrix}} \cdot (\underline{y} - \hat{\underline{\mu}}) = \begin{bmatrix} \underline{x}_1^T (\underline{y} - \hat{\underline{\mu}}) \\ \vdots \\ \underline{x}_p^T (\underline{y} - \hat{\underline{\mu}}) \end{bmatrix} = \underline{0}$$

If the model includes the intercept $\Rightarrow \underline{x}_1 = \underline{1}_n$
$$\Rightarrow \underline{x}_1^T (\underline{y} - \hat{\underline{\mu}}) = \underline{1}^T (\underline{y} - \hat{\underline{\mu}}) = \sum_{i=1}^n (y_i - \hat{\mu}_i) = 0$$

second derivative $\ell_{**}(\underline{\beta}) = \left\{ \dfrac{\partial^2 \ell(\beta)}{\partial \beta_r \partial \beta_s} \right\}_{r, s = 1, \dots, p} = -\sum_{i=1}^n x_{ir} x_{is} e^{\tilde{x}_i^T \underline{\beta}}$
$$= -\sum_{i=1}^n x_{ir} x_{is} \mu_i$$

In matrix form we get $\ell_{**}(\underline{\beta}) = -X^T U X$ with U an $n \times n$ diagonal matrix
$$U = \begin{bmatrix} \mu_1 & 0 & 0 & \cdots & 0 \\ 0 & \mu_2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \mu_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & \mu_n \end{bmatrix} = \text{diag}\{\mu_1, \dots, \mu_n\} = \text{diag}\{e^{\tilde{x}_1^T \beta}, \dots, e^{\tilde{x}_n^T \beta}\}$$

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

The OBSERVED INFORMATION evaluated at the MLE $\hat{\underline{\beta}}$ is
$$j(\hat{\underline{\beta}}) = -\ell_{**}(\underline{\beta}) \big|_{\underline{\beta} = \hat{\underline{\beta}}} = X^T U(\hat{\underline{\beta}}) X$$
where $U(\hat{\underline{\beta}}) = \text{diag}\{e^{\tilde{x}_1^T \hat{\beta}}, \dots, e^{\tilde{x}_n^T \hat{\beta}}\}$

- INFERENCE

inference here is based on APPROXIMATE distributions
Remarks:
- notation: we write "Y approximately distributed as (some distribution p(y))" as "$Y \,\dot\sim\, p(y)$"
- approximations get better with n (large samples ⇝ better approximation)

- DISTRIBUTION of the MAXIMUM LIKELIHOOD ESTIMATOR of the REGRESSION PARAMETERS
$$\hat{\underline{\beta}} \,\dot\sim\, N_p(\underline{\beta}, j(\hat{\underline{\beta}})^{-1})$$
the marginal distribution for the j-th element is $\hat{\beta}_j \,\dot\sim\, N(\beta_j, [j(\hat{\beta})^{-1}]_{jj})$ $\qquad j = 1, \dots, p$
element of the matrix $j(\hat{\beta})^{-1}$ in position (j, j)

- CONFIDENCE INTERVAL FOR β_j

A pivotal quantity is
$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}}} \,\dot\sim\, N(0, 1)$$
a confidence interval with level $(1-\alpha)$ for β_j $(j = 1, \dots, p)$ can be obtained as
$$\mathbb{P}\left(z_{\frac{\alpha}{2}} < \frac{\hat{\beta}_j - \beta_j}{\sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}}} < z_{1-\frac{\alpha}{2}} \right) = 1 - \alpha$$
Gaussian is symmetric $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$
quantile of level $1-\frac{\alpha}{2}$ of a $N(0,1)$

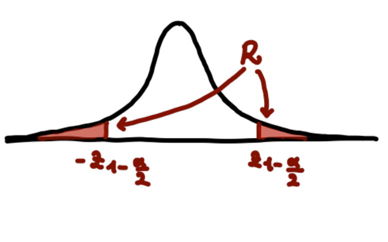
with the data:
$$\hat{\beta}_j - \sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}} \cdot z_{1-\frac{\alpha}{2}} < \beta_j < \hat{\beta}_j + \sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}} \cdot z_{1-\frac{\alpha}{2}}$$
$$\Rightarrow \beta_j \in \hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}}$$

- 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 test statistic
$$z_j = \frac{\hat{\beta}_j - b_j}{\sqrt{[j(\hat{\underline{\beta}})^{-1}]_{jj}}} \,\dot\sim\, N(0, 1) \text{ under } H_0$$
the observed value of the test is z_j^{obs}

- if we use a fixed significance level α
  $\alpha = P_{H_0}(|z_j| > z_{1-\frac{\alpha}{2}})$
  → 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 $\alpha^{obs} = \mathbb{P}_{H_0}(|z_j| \geq |z_j^{obs}|) = 2(1 - \Phi(|z_j^{obs}|))$