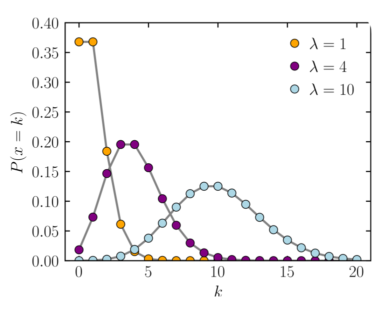
POISSON REGRESSION

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:
$Y \sim \text{Poisson}(\mu)$
- parameter space: $\mu > 0 \Rightarrow \Theta = (0, +\infty)$
- 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. $\eta_i = \tilde{x}_i^T \beta$
3. $g(\mu_i) = \eta_i$ with $g = \log$ LOGARITHMIC LINK FUNCTION "log-linear model"

Remarks:
- the log link allows mapping the linear predictor $\eta_i = \tilde{x}_i^T\beta \in \mathbb{R}$ to $\mathbb{R}^+$, the parameter space of μ_i
  indeed $\log(\mu_i) = \eta_i \Rightarrow \mu_i = e^{\eta_i} = e^{\tilde{x}_i^T\beta}$
  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 model is
$$\mathbb{P}(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \qquad \log(\mu_i) = \tilde{x}_i^T\underline{\beta} \Rightarrow \mu_i = e^{\tilde{x}_i^T\underline{\beta}}$$
$$= \frac{e^{-e^{\tilde{x}_i^T\beta}}\, e^{\tilde{x}_i^T\beta y_i}}{y_i!}$$

- INTERPRETATION of the regression parameters

Let's study the mean μ at two values x_j and $x_j + 1$ of the j-th covariate
at x_j we obtain $\mu_1 = \exp\{\beta_1 + \beta_2 x_2 + \dots + \beta_j x_j + \dots + \beta_p x_p\}$
at $(x_j + 1)$ $\mu_2 = \exp\{\beta_1 + \beta_2 x_2 + \dots + \beta_j (x_j + 1) + \dots + \beta_p x_p\}$

$$\frac{\mu_2}{\mu_1} = \frac{\exp\{\beta_1 + \beta_2 x_2 + \dots + \beta_j(x_j+1) + \dots + \beta_p x_p\}}{\exp\{\beta_1 + \beta_2 x_2 + \dots + \beta_j x_j + \dots + \beta_p x_p\}}$$
$$= \exp\{\cancel{\beta_1} + \cancel{\beta_2 x_2} + \dots + \beta_j(x_j+1) + \dots + \cancel{\beta_p x_p} - \cancel{\beta_1} - \cancel{\beta_2 x_2} - \dots - \beta_j x_j - \dots - \cancel{\beta_p x_p}\}$$
$$= \exp\{\cancel{\beta_j x_j} + \beta_j - \cancel{\beta_j x_j}\} = e^{\beta_j}$$
$$\Rightarrow \beta_j = \log\frac{\mu_2}{\mu_1} = \log\mu_2 - \log\mu_1 = \log\mathbb{E}(Y \mid x_j = x_j + 1) - \log\mathbb{E}(Y \mid x_j = x_j)$$

The parameter β_j represents the difference of the logs of the expected counts if we increase x_j of 1 unit, while keeping the other predictors fixed.

Or, if we write: $e^{\beta_j} = \dfrac{\mu_2}{\mu_1} \Rightarrow \mu_2 = \mu_1 \cdot e^{\beta_j}$
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.

- ESTIMATE

likelihood
$$L(\underline{\beta}) \propto \prod_{i=1}^n p(y_i \mid \underline{\beta}) = \prod_{i=1}^n \frac{e^{-\mu_i}\mu_i^{y_i}}{\cancel{y_i!}} \qquad \mu_i = e^{\tilde{x}_i^T\beta}$$
$$\propto e^{-\sum_{i=1}^n \mu_i}\prod_{i=1}^n \mu_i^{y_i}$$

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

$$\ell_*(\underline{\beta}) = \left\{\frac{\partial}{\partial\beta_r}\ell(\underline{\beta})\right\}_{r=1,\dots,p} \qquad \text{(score function)}$$

$$\frac{\partial}{\partial\beta_r}\ell(\underline{\beta}) = -\sum_{i=1}^n x_{ir} e^{\tilde{x}_i^T\beta} + \sum_{i=1}^n y_i x_{ir} = \sum_{i=1}^n x_{ir}(y_i - e^{\tilde{x}_i^T\beta})$$
$$= \sum_{i=1}^n x_{ir}(y_i - \mu_i)$$

for the vector $\underline{\beta}$
$$\frac{\partial}{\partial\underline{\beta}}\ell(\underline{\beta}) = -\sum_{i=1}^n \tilde{x}_i \cdot e^{\tilde{x}_i^T\beta} + \sum_{i=1}^n \tilde{x}_i y_i = -\sum_{i=1}^n \tilde{x}_i\mu_i + \sum_{i=1}^n \tilde{x}_i y_i = \sum_{i=1}^n \tilde{x}_i(y_i - \mu_i) = \underset{p\times n}{X^T}\underset{n\times 1}{(\underline{y} - \underline{\mu})}$$

The MLE $\hat{\underline{\beta}}$ is the solution of $X^T(\underline{y} - \underline{\mu}) = \underline{0}$ → it resembles the normal equations,
$X^T(\underline{y} - e^{X\underline{\beta}}) = 0$ but here $\underline{\mu}$ is a nonlinear 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 $\hat{\underline{\beta}}$.
However, notice that, similarly to the LM, since $\hat{\beta}$ is the solution of the equation, we obtain
$$\underset{p\times n}{X^T}\underset{n\times 1}{(\underline{y} - \hat{\underline{\mu}})} = 0 \Rightarrow \underset{p\times n}{\begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_p^T \end{bmatrix}} \cdot (\underline{y} - \hat{\underline{\mu}}) = \begin{bmatrix} \underline{x}_1^T(\underline{y} - \hat{\underline{\mu}}) \\ \vdots \\ \underline{x}_p^T(\underline{y} - \hat{\underline{\mu}}) \end{bmatrix} = \underline{0}$$

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
$$\ell_{**}(\underline{\beta}) = \left\{\frac{\partial^2\ell(\beta)}{\partial\beta_r\partial\beta_s}\right\}_{r,s=1,\dots,p} = -\sum_{i=1}^n x_{ir}x_{is}e^{\tilde{x}_i^T\underline{\beta}}$$
$$= -\sum_{i=1}^n x_{ir}x_{is}\mu_i < 0 \quad (\Rightarrow \hat{\underline{\beta}} \text{ is a max})$$

Hence the matrix is $\ell_{**}(\underline{\beta}) = -X^T U X$ with $U = \text{diag}\{\mu_1, \dots, \mu_n\} = \text{diag}\{e^{\tilde{x}_1^T\beta}, \dots, e^{\tilde{x}_n^T\beta}\} = U(\underline{\beta})$

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

- INFERENCE

inference here is based on approximate distributions
(we write "approximately distributed as" with $\dot\sim$; approximations get better for large n)

- 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 is $\hat{\beta}_j \,\dot\sim\, N(\beta, [j(\hat{\beta})^{-1}]_{(jj)})$

- Hence an approximate confidence interval with level $(1-\alpha)$ for β_j $(j = 1, \dots, p)$ can be obtained as
$$\hat{\beta}_j \pm \underbrace{z_{1-\frac{\alpha}{2}}}_{\text{quantile of level } 1-\frac{\alpha}{2} \text{ of a } N(0,1)}\sqrt{[j(\hat{\beta})^{-1}]_{(jj)}}$$

- A test $H_0: \beta_j = b_j$ vs $H_1: \beta_j \neq b_j$ is performed as
$$z_j = \frac{\hat{\beta}_j - b_j}{\sqrt{[j(\hat{\underline{\beta}})^{-1}]_{(jj)}}} \,\dot\sim\, N(0,1) \text{ under } H_0$$
the p-value is $\alpha^{obs} = \mathbb{P}_{H_0}(|z_j| \geq |z_j^{obs}|) = 2(1 - \Phi(z_j^{obs}))$

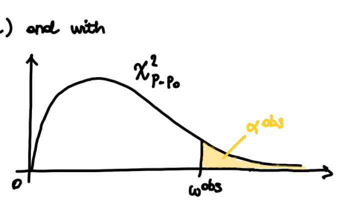
- TEST for comparing NESTED MODELS

we have a model $Y_i \sim \text{Pois}(\mu_i)$ $(i = 1, \dots, n)$
with $\log(\mu_i) = \tilde{x}_i^T\underline{\beta} = \beta_1 + \beta_2 x_{i2} + \dots + \beta_{p_0}x_{ip_0} + \underbrace{\boxed{\beta_{p_0+1}x_{i,p_0+1} + \dots + \beta_p x_{ip}}}_{\text{under } H_0 \text{ this is } = 0}$
we call it the "full" model (it is the proposed model).
we want to test
$$\begin{cases} H_0: \beta_{p_0+1} = \dots = \beta_p = 0 \\ H_1: \overline{H_0} \end{cases}$$
We can partition the vector $\underline{\beta} = \begin{bmatrix} \underline{\beta}^{(0)} \\ \underline{\beta}^{(1)} \end{bmatrix}$ $\quad \underline{\beta}^{(0)} \in \mathbb{R}^{p_0}$, $\underline{\beta}^{(1)} \in \mathbb{R}^{p-p_0}$
$$\begin{cases} H_0: \underline{\beta}^{(1)} = \underline{0} \\ H_1: \underline{\beta}^{(1)} \neq \underline{0} \end{cases}$$

under H_0 we have the "restricted model"
$Y_i \sim \text{Pois}(\mu_i)$ with $\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \dots + \beta_{p_0}x_{ip_0}$

To compare two nested models we use the LIKELIHOOD RATIO TEST: it compares the maximum of the likelihood of the full and of the restricted model:
$$W = 2\log\frac{\hat{L}(\text{model})}{\tilde{L}(\text{restricted})} = 2\{\hat{\ell}(\text{model}) - \tilde{\ell}(\text{restricted})\} \,\dot\sim\, \chi^2_{\underbrace{p-p_0}_{(\#\text{ covariates under } H_1) - (\#\text{ covariates under } H_0)}} \text{ under } H_0$$

We can denote with $\hat{\underline{\beta}} = (\hat{\underline{\beta}}^{(0)}, \hat{\underline{\beta}}^{(1)})$ the MLE under H_1 (full model) and with $\tilde{\underline{\beta}} = (\tilde{\underline{\beta}}^{(0)}, \underline{0})$ the MLE under H_0 (restricted). Then,
$$W = 2\{\ell(\hat{\underline{\beta}}^{(0)}, \hat{\underline{\beta}}^{(1)}) - \ell(\tilde{\underline{\beta}}^{(0)}, \underline{0})\} \,\dot\sim\, \chi^2_{p-p_0} \text{ under } H_0$$
with the data we compute w^{obs}
and the p-value is $\alpha^{obs} = \mathbb{P}_{H_0}(W \geq w^{obs}) = \mathbb{P}(\chi^2_{p-p_0} \geq w^{obs})$



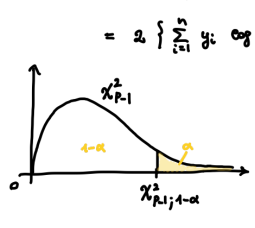
- TEST about the OVERALL SIGNIFICANCE

Similarly to the LM, we can test
$$\begin{cases} H_0: \beta_2 = \beta_3 = \dots = \beta_p = 0 \\ H_1: \overline{H_0} \end{cases}$$
We can use the test for nested models by setting $p_0 = 1$.
In this case we compare the full model with a model with only the intercept ("null model").

Under H_0: $Y_i \sim \text{Pois}(\mu_i)$ indep. $i = 1, \dots, n$
$\mu_i = e^{\beta_1} = \mu$
$$L(\mu) = \prod_{i=1}^n \frac{e^{-\mu}\mu^{y_i}}{y_i!} \propto e^{-n\mu}\mu^{\sum y_i}$$
$$\ell(\mu) = -n\mu + \sum_{i=1}^n y_i \cdot \log(\mu) = -n\mu + n\bar{y}\log\mu$$
$$\ell_*(\mu) = -n + \frac{n\bar{y}}{\mu}$$
$\ell_*(\mu) = 0 \Rightarrow -n\mu = -n\bar{y} \Rightarrow \tilde{\mu} = \bar{y}$ — MLE under the restricted ("null") model; we estimate the common mean using the sample mean
Moreover, since $\log(\mu_i) = \log\mu = \beta_1$ (one-to-one correspondence: bijective function)
we automatically obtain $\tilde{\beta}_1 = \log\tilde{\mu} = \log\bar{y}$
$$\ell_{**}(\mu) = -n\bar{y}\cdot\frac{1}{\mu^2}$$
$$\ell_{**}(\tilde{\mu}) = -\frac{n}{\bar{y}} < 0 \quad \text{it's a max}$$
Hence $\tilde{\ell}(\text{restricted}) = \ell(\tilde{\mu}) = -n\bar{y} + n\bar{y}\cdot\log\bar{y} = n(\bar{y}\log\bar{y} - \bar{y})$

Under H_1 we have the model with p covariates
$\mu_i = \exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_p x_{ip}\}$
we estimate $\hat{\underline{\beta}}$ numerically, and we compute
$$\hat{\ell}(\text{model}) = \ell(\hat{\underline{\beta}}) = -\sum_{i=1}^n \hat{\mu}_i + \sum_{i=1}^n y_i\log\hat{\mu}_i = -\sum_{i=1}^n e^{\tilde{x}_i^T\hat{\beta}} + \sum_{i=1}^n y_i\tilde{x}_i^T\hat{\beta}$$
The likelihood ratio test in this case is
$$W = 2\{\hat{\ell}(\text{model}) - \tilde{\ell}(\text{restricted})\} = 2\{\ell(\hat{\beta}) - \ell(\bar{Y})\} \,\dot\sim\, \chi^2_{p-1} \text{ under } H_0$$

With the data:
$$w^{obs} = 2\left\{-\sum_{i=1}^n \hat{\mu}_i + \sum_{i=1}^n y_i\log\hat{\mu}_i - n(\bar{y}\log\bar{y} - \bar{y})\right\}$$
$$= 2\left\{\sum_{i=1}^n y_i\log\hat{\mu}_i - \sum_{i=1}^n y_i\log\bar{y} - \sum_{i=1}^n \hat{\mu}_i + n\bar{y}\right\}$$
$$= 2\left\{\sum_{i=1}^n y_i\log\frac{\hat{\mu}_i}{\bar{y}} - \sum_{i=1}^n \hat{\mu}_i + n\bar{y}\right\}$$



reject H_0 if $w^{obs} \geq \chi^2_{p-1;1-\alpha}$ (quantile of level $1-\alpha$ of a χ^2_{p-1})

- TEST about the GOODNESS OF FIT of the model

we first need to introduce the concept of "SATURATED MODEL". This is the most elaborate model one can consider: i.e., a model with n parameters (one for each observation).
We want to estimate a model with n parameters using n observations $\Rightarrow$ we obtain $\tilde{\mu}_i^S = y_i$ $i = 1, \dots, n$
a model with a perfect fit (perfect but useless: interpolation, there is no simplification → the model is keeping all the erratic variability of the data and is not highlighting the underlying systematic behavior)
$$\left(\begin{aligned} &\ell(\mu_i) = -\mu_i + y_i\log\mu_i \\ &\Rightarrow \ell_*(\mu_i) = -1 + \frac{y_i}{\mu_i} \Rightarrow \frac{y_i}{\mu_i} - 1 = 0 \Rightarrow \tilde{\mu}_i^S = y_i \\ &\Rightarrow \ell(\tilde{\mu}_1^S, \dots, \tilde{\mu}_n^S) = \sum_{i=1}^n y_i\log y_i - \sum_{i=1}^n y_i \end{aligned}\right)$$

The ("full") model with p parameters can be compared with the saturated model using the likelihood ratio test.
For this particular case, this quantity is called DEVIANCE (or "residual deviance")
$$D = \text{deviance}(\text{model}) = 2\{\tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model})\}$$

Since $\tilde{\mu}_i^S = y_i$ for all i in the saturated model, the log-likelihood evaluated at $\tilde{\mu}_i^S$ is always
$$\tilde{\ell}(\text{saturated}) = \sum_{i=1}^n y_i\log y_i - \sum_{i=1}^n y_i$$

Hence, $D = 2\{\tilde{\ell}(\text{saturated}) - \ell(\underbrace{\hat{\underline{\mu}}}_{\text{estimate of } \underline{\mu} \text{ in the model; } \hat{\underline{\mu}} \text{ is a transformation of } \hat{\underline{\beta}}})\} = 2\Big\{\underbrace{\sum_{i=1}^n (y_i\log y_i - y_i)}_{\text{does not depend on } \beta} - \sum_{i=1}^n (y_i\log\hat{\mu}_i - \hat{\mu}_i)\Big\}$
$$= 2\Big\{\sum_{i=1}^n y_i\log\frac{y_i}{\hat{\mu}_i} - \underbrace{\sum_{i=1}^n(\cancel{y_i - \hat{\mu}_i})}_{=0 \text{ if the model includes the intercept}}\Big\} = 2\sum_{i=1}^n y_i\log\frac{y_i}{\hat{\mu}_i}$$

Since the saturated model has a perfect fit, for sure $\tilde{\ell}(\text{saturated}) > \hat{\ell}(\text{model}) \Rightarrow D > 0$.
Moreover, if the model with p covariates fits the data well, $\hat{\ell}(\text{model})$ will not be "too far" from $\tilde{\ell}(\text{saturated})$.
→ A good model will have a small deviance
→ WE DO NOT HAVE A DISTRIBUTION FOR THE DEVIANCE
When the LR test is used to compare the saturated model we lose the approximation to a χ^2 distribution
We can't do formal tests, a deviance $< n-p$ is generally ok. It indicates that the model fits the data well: it does not lose too much accuracy compared to the "perfect" saturated model.
Notice that deviance $= 0$ means that you fit the data perfectly, but with p parameters instead of n.
The residual deviance is more useful when used to compare different models (on the same data), with the same number of covariates p (but different covariates).

- RELATIONSHIP between SATURATED, PROPOSED and NULL MODEL

Notice that the saturated model and the null model (with only the intercept) are the two extreme cases:
- SATURATED model: n parameters
- proposed model: p parameters (nested)
- NULL model: 1 parameter (nested)

When you estimate a glm in R, the output returns 2 quantities:
"Residual deviance" and "Null deviance"

likelihood ratio test:
- between saturated and proposed (the proposed model can be seen as a "restricted" model w.r.t. the saturated)
  $2\{\tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model})\} = D(\text{model})$ "residual deviance"
- between saturated and null (also the null model can be seen as a restricted model w.r.t. the saturated)
  $2\{\tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{null})\} = D(\text{null})$ "null deviance"
- between model and null (test of overall significance)
$$2\{\hat{\ell}(\text{model}) - \tilde{\ell}(\text{null})\} = 2\{\hat{\ell}(\text{model}) + \tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{null})\}$$
$$= 2\{[\tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{null})] - [\tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model})]\}$$
$$= D(\text{null}) - D(\text{model})$$
we can write the LR test in terms of difference of the deviances

- MODEL CHECKING: RESIDUALS

In the linear model we had $\underline{y} = \hat{\underline{\mu}} + (\underline{y} - \hat{\underline{\mu}}) = \hat{\underline{\mu}} + \underline{e}$ and we studied the expected behavior of the residuals when the model assumptions are valid, to compare it with the observed behavior after fitting the model.
In the linear model the residuals were the "sample counterpart" of the errors: here we do not have them.
Since now it is not so clear how to define residuals, several versions have been proposed.
- Pearson's residuals: they are the analogous version of the standardised residuals in the LM.
$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \qquad i = 1, \dots, n$$
For the Poisson, we have $V(\mu) = \mu \Rightarrow e_i = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$ $\quad i = 1, \dots, n$

They have approximately zero mean and constant variance.