21 Nov - LEC 10

• TEST for comparing NESTED MODELS
(test about a SUBSET OF COEFFICIENTS)

We have a model $Y_i \sim \text{Pois}(\mu_i)$ $i=1,\dots,n$ indep.

with $\log(\mu_i) = \underset{\sim}{x}_i^T \underline{\beta}$

$$= \beta_1 + \beta_2 x_{i2} + \dots + \beta_{p_0} x_{ip_0} + \beta_{p_0+1} x_{i,p_0+1} + \dots + \beta_p x_{ip}$$

we call it the "full" model (it is the proposed model).

we want to test

$$\begin{cases} H_0: \beta_{p_0+1} = \dots = \beta_p = 0 \\ H_1: \text{at least one } \beta_r \neq 0 \quad r \in \{p_0+1, \dots, p\} \end{cases}$$

We can partition the vector $\underline{\beta} = \begin{bmatrix} \underline{\beta}^{(0)} \\ \underline{\beta}^{(1)} \end{bmatrix}$ $\quad \underline{\beta}^{(0)} \in \mathbb{R}^{p_0}$, $\quad \underline{\beta}^{(1)} \in \mathbb{R}^{p-p_0}$

with $\underline{\beta}^{(0)} = [\beta_1, \beta_2, \dots, \beta_{p_0}]^T$

and $\underline{\beta}^{(1)} = [\beta_{p_0+1}, \dots, \beta_p]^T$

$$\begin{cases} H_0: \underline{\beta}^{(1)} = \underline{0} \\ H_1: \underline{\beta}^{(1)} \neq \underline{0} \end{cases}$$

under H_0 we have the "restricted" model

$Y_i \sim \text{Pois}(\mu_i^0)$ indep $i=1,\dots,n$

with $\log(\mu_i^0) = \beta_1 + \beta_2 x_{i2} + \dots + \beta_{p_0} x_{ip_0}$

To compare two nested models we use the LIKELIHOOD RATIO TEST: it compares the MAXIMUM of the LIKELIHOOD under the full and under the restricted models.

- Remark: the likelihood is $L(\beta)$. The maximum is obtained at the MLE.
  under H_1, $\underline{\beta} = (\underline{\beta}^{(0)}, \underline{\beta}^{(1)})^T$, the MLE is $\hat{\underline{\beta}} = (\hat{\underline{\beta}}^{(0)}, \hat{\underline{\beta}}^{(1)})^T$
  under H_0, $\underline{\beta} = (\underline{\beta}^{(0)}, \underline{0})^T$, the MLE is $\tilde{\underline{\beta}} = (\tilde{\underline{\beta}}^{(0)}, \underline{0})^T$
  - likelihood under the full model: $L(\text{model})$, with maximum $\hat{L}(\text{model}) = L(\hat{\underline{\beta}}^{(0)}, \hat{\underline{\beta}}^{(1)})$
  - likelihood under the restricted model: $L(\text{restricted})$, with maximum $\tilde{L}(\text{restricted}) = L(\tilde{\underline{\beta}}^{(0)}, \underline{0})$

  When studied as functions of the ML estimator $\hat{\beta}$, they are random quantities (with a distribution).

LIKELIHOOD RATIO TEST

$$W = 2 \log \frac{\hat{L}(\text{model})}{\tilde{L}(\text{restricted})} = 2 \{ \hat{\ell}(\text{model}) - \tilde{\ell}(\text{restricted}) \}$$

$$= 2 (\ell(\hat{\underline{\beta}}^{(0)}, \hat{\underline{\beta}}^{(1)}) - \ell(\tilde{\underline{\beta}}^{(0)}, \underline{0})) \dot{\sim} \chi^2_{p-p_0} \quad \text{under } H_0$$

(# covariates under H_1) - (# covariates under H_0)

with the data we compute $w^{obs} = 2(\ell(\hat{\underline{\beta}}) - \ell(\tilde{\underline{\beta}}))$.

What values do we expect under H_0 and H_1?

First, notice that $\tilde{\underline{\beta}}$ is a CONSTRAINED estimate $\Rightarrow \hat{\ell}(\text{model}) \geq \tilde{\ell}(\text{restricted})$

$\Rightarrow w^{obs} \geq 0$

If H_0 is true, the estimates under H_0 and H_1 will be similar

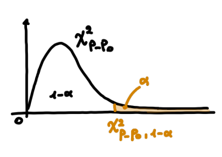
$\Rightarrow \hat{\ell}(\text{model}) \approx \tilde{\ell}(\text{restricted}) \Rightarrow w^{obs} \approx 0$

If H_0 is not true, removing $\underline{\beta}^{(1)}$ will lead to a worse fit, hence

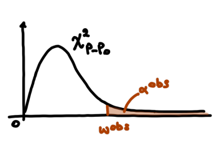
$\Rightarrow \hat{\ell}(\text{model}) \gg \tilde{\ell}(\text{restricted}) \Rightarrow w^{obs} \gg 0$

The reject region will comprise large values of the test

- fixed significance level α
  $\alpha = \mathbb{P}_{H_0}(W > \chi^2_{p-p_0; 1-\alpha})$
  $R = (\chi^2_{p-p_0, 1-\alpha} ; +\infty)$
  ↳ $(1-\alpha)$-quantile of a χ^2 distribution with $p-p_0$ d.o.f



- pvalue
  $\alpha^{obs} = \mathbb{P}_{H_0}(W > w^{obs})$
  with $W \overset{H_0}{\sim} \chi^2_{p-p_0}$



• TEST ABOUT THE OVERALL SIGNIFICANCE

Similarly to the LM, we can test

$$\begin{cases} H_0: \beta_2 = \beta_3 = \dots = \beta_p = 0 \\ H_1: \exists r \in \{2, \dots, p\} : \beta_r \neq 0 \end{cases}$$

We can use the test for nested models with $p_0 = 1$.

In this case we compare the full model with a model with only the intercept (null model).

Under H_0:

$Y_i \sim \text{Pois}(\mu_i)$ independent for $i=1,\dots,n$

with $\mu_i = e^{\beta_1} = \mu$ equal for all i

Remark: $\mu = e^{\beta_1} \Leftrightarrow \beta_1 = \log \mu$ this is a reparameterization (one-to-one correspondence)

likelihood

$$L(\mu) = \prod_{i=1}^n \frac{e^{-\mu} \mu^{y_i}}{y_i!} \propto e^{-n\mu} \mu^{\sum y_i}$$

log-likelihood

$$\ell(\mu) = -n\mu + \sum_{i=1}^n y_i \cdot \log(\mu) = -n\mu + n\bar{y} \log \mu$$

score function

$$\ell_*(\mu) = -n + \frac{n\bar{y}}{\mu}$$

likelihood equation

$$\ell_*(\mu) = 0 \Rightarrow -n\mu = -n\bar{y} \Rightarrow \tilde{\mu} = \bar{y}$$

MLE of $\mu = \mathbb{E}[Y_i]$ under the null model

we estimate the common mean using the sample mean

Moreover, since $\log \mu_i = \log \mu = \beta_1$

we automatically obtain $\tilde{\beta}_1 = \log \tilde{\mu} = \log \bar{y}$

second derivative

$$\ell_{**}(\mu) = -n\bar{y} \cdot \frac{1}{\mu^2} \Rightarrow \ell_{**}(\tilde{\mu}) = -\frac{n}{\bar{y}} < 0 \quad \text{it's a max}$$

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

$$\hat{\ell}(\text{model}) = \ell(\hat{\underline{\beta}}) = -\sum_{i=1}^n \hat{\mu}_i + \sum_{i=1}^n y_i \log \hat{\mu}_i = -\sum_{i=1}^n e^{\underset{\sim}{x}_i^T \hat{\underline{\beta}}} + \sum_{i=1}^n y_i \underset{\sim}{x}_i^T \hat{\underline{\beta}}$$

The likelihood ratio test in this case is

$$W = 2 \{ \hat{\ell}(\text{model}) - \tilde{\ell}(\text{restricted}) \} = 2 \{ \ell(\hat{\beta}) - \ell(\bar{Y}) \} \dot{\sim} \chi^2_{p-1} \quad \text{under } H_0$$

With the data:

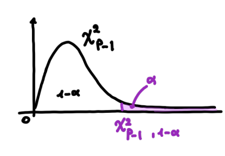
$$w^{obs} = 2 \left\{ -\sum_{i=1}^n \hat{\mu}_i + \sum_{i=1}^n y_i \log \hat{\mu}_i - n(\bar{y} \log \bar{y} - \bar{y}) \right\}$$

$$= 2 \left\{ \sum_{i=1}^n y_i \log \hat{\mu}_i - \sum_{i=1}^n y_i \log \bar{y} - \sum_{i=1}^n \hat{\mu}_i + n\bar{y} \right\}$$

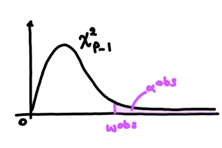
$$= 2 \left\{ \sum_{i=1}^n y_i \log \frac{\hat{\mu}_i}{\bar{y}} - \sum_{i=1}^n \hat{\mu}_i + n\bar{y} \right\}$$

The reject region comprises large values of the test

- fixed significance level α
  $\alpha = \mathbb{P}_{H_0}(W > \chi^2_{p-1; 1-\alpha})$
  $R = (\chi^2_{p-1, 1-\alpha} ; +\infty)$
  ↳ $(1-\alpha)$-quantile of a χ^2 distribution with $p-1$ d.o.f



- pvalue
  $\alpha^{obs} = \mathbb{P}_{H_0}(W > w^{obs})$
  with $W \overset{H_0}{\sim} \chi^2_{p-1}$



• GOODNESS OF FIT OF THE MODEL

we first need to introduce the concept of "saturated model".

SATURATED MODEL: it is the most elaborate model one can estimate: i.e., a model with n parameters (one for each observation).

$\Rightarrow$ model for n observations based on n parameters

$Y_i \sim \text{Pois}(\mu_i)$ independent for $i=1,\dots,n$

parameters $(\mu_1, \dots, \mu_n)$

$$L(\mu_1, \dots, \mu_n) = \prod_{i=1}^n [e^{-\mu_i} \mu_i^{y_i}] = e^{-\sum_{i=1}^n \mu_i} \prod_{i=1}^n \mu_i^{y_i}$$

$$\ell(\mu_1, \dots, \mu_n) = -\sum_{i=1}^n \mu_i + \sum_{i=1}^n \log \mu_i^{y_i} = -\sum_{i=1}^n \mu_i + \sum_{i=1}^n y_i \log \mu_i$$

$$\ell_*(\mu_1, \dots, \mu_n) = \left[\frac{\partial}{\partial \mu_1} \ell(\underline{\mu}) \dots \frac{\partial}{\partial \mu_n} \ell(\underline{\mu}) \right]^T \quad \text{with} \quad \frac{\partial}{\partial \mu_i} \ell(\underline{\mu}) = -1 + \frac{y_i}{\mu_i}$$

$$\frac{\partial}{\partial \mu_i} \ell(\underline{\mu}) = 0 \Rightarrow \frac{y_i}{\mu_i} = 1 \Rightarrow \tilde{\mu}_i^S = y_i$$

MLE of μ_i under the saturated model

The log-likelihood computed at $(\tilde{\mu}_1^S, \dots, \tilde{\mu}_n^S)$ is

$$\ell(\tilde{\mu}_1^S, \dots, \tilde{\mu}_n^S) = -\sum_{i=1}^n \tilde{\mu}_i^S + \sum_{i=1}^n y_i \log \tilde{\mu}_i^S$$

$$= -\sum_{i=1}^n y_i + \sum_{i=1}^n y_i \log y_i \qquad (1)$$

Remark: the saturated model has a perfect fit since it models each observation with itself: it is INTERPOLATING THE DATA.

However, it is useless since it is not providing any simplification of the phenomenon, and it is not modeling its variability.

Since this model has error $= 0$ when fitting y_i ($\tilde{y}_i^S = y_i \Rightarrow \tilde{y}_i^S - y_i = 0$ for all i) we can use it as a benchmark to evaluate the proposed ("full") model.

The proposed model with p parameters can be compared with the saturated model using the likelihood ratio test.

For this particular case, the LR test is called DEVIANCE (or "residual deviance")

$$2 \{ \tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model}) \} = \text{deviance}(\text{model}) = D$$

We have seen that $\tilde{\ell}(\text{saturated}) = \sum_{i=1}^n y_i \log y_i - \sum_{i=1}^n y_i$

Hence the deviance of a model is

$$D = 2 \{ \ell(\tilde{\underline{\mu}}^S) - \ell(\hat{\underline{\beta}}) \} = 2 \left\{ \sum_{i=1}^n y_i \log y_i - \sum_{i=1}^n y_i - \sum_{i=1}^n y_i \underbrace{\underset{\sim}{x}_i^T \hat{\underline{\beta}}}_{\log \hat{\mu}_i} + \sum_{i=1}^n \underbrace{e^{\underset{\sim}{x}_i^T \hat{\underline{\beta}}}}_{\hat{\mu}_i} \right\}$$

$$= 2 \left\{ \sum_{i=1}^n y_i \log y_i - \sum_{i=1}^n y_i - \sum_{i=1}^n y_i \log \hat{\mu}_i + \sum_{i=1}^n \hat{\mu}_i \right\}$$

$$= 2 \left\{ \sum_{i=1}^n y_i \log \frac{y_i}{\hat{\mu}_i} - \cancel{\sum_{i=1}^n (y_i - \hat{\mu}_i)} \right\}$$

$= 0$ if the model includes the intercept

$$= 2 \sum_{i=1}^n y_i \log \frac{y_i}{\hat{\mu}_i}$$

Since the saturated model has a perfect fit,

$$\tilde{\ell}(\text{saturated}) \geq \hat{\ell}(\text{model}) \Rightarrow D \geq 0.$$

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 model that fits the data well will have a small deviance

→ WE DO NOT HAVE A DISTRIBUTION FOR THE DEVIANCE

When the likelihood ratio test is applied to the saturated model, we lose the approximation to the χ^2 distribution. Hence, we can't do formal tests.

However, just as an indication, 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 ($p<n$) 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
- NULL model: 1 parameter

(nested) (nested)

Let's analyze the different likelihood ratio tests that we can define with these models:

I. between the SATURATED AND PROPOSED

the proposed model can be seen as a restricted model w.r.t. the saturated model

LR test: $2 \{ \tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model}) \} = D(\text{model})$

this is the DEVIANCE OF THE MODEL (RESIDUAL DEVIANCE)

II. between the SATURATED AND THE NULL

Also the null model can be seen as a restricted model w.r.t. the saturated model

LR test: $2 \{ \tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{null}) \} = D(\text{null})$

this is the NULL DEVIANCE

III. between the MODEL AND THE NULL (test about the overall significance)

LR test: $2 \{ \hat{\ell}(\text{model}) - \tilde{\ell}(\text{null}) \}$

$$= 2 \{ \hat{\ell}(\text{model}) + \tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{null}) \}$$

$$= 2 \{ [\tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{null})] - [\tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model})] \}$$

$$= D(\text{null}) - D(\text{model})$$

↳ we can write the LR test for testing the overall significance of the model as a difference of the two deviances.

Remark: estimating a GLM in R, the output returns both the residual deviance and the null deviance.

• 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 standardized 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 $V(\mu) = \mu \Rightarrow e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \qquad i = 1, \dots, n$

They have approximately zero mean and constant variance.