1.7 Deviance

Deviance

One way of assessing the adequacy of a model is to compare it with a more general model with the maximum number of parameters that can be estimated. This is called the **saturated** model.

If there are N observations $Y_i, i = 1, \ldots, N$, then a saturated model can be specified with N parameters. Also called the **maximal** or **full model**. (In general, however, the maximum number m of parameters that can be estimated can be smaller than N, e.g., if observations are repeated.)

We write $\boldsymbol{\theta}_{\max}$ *for the parameter vector of the saturated model* and $\hat{\boldsymbol{\theta}}_{\max}$ for the maximum likelihood estimator of $\boldsymbol{\theta}_{\max}$.

The likelihood for the saturated model $L(\hat{\boldsymbol{\theta}}_{\max}; \mathbf{y})$ will be *larger* than any other likelihood function for these observations, with the same assumed distribution and link function, because it provides the most complete description of the data.

Let $L(\hat{\boldsymbol{\theta}}; \mathbf{y})$ denote the maximum value of the likelihood function for the model of interest. Then the **likelihood ratio**

$$\lambda = \frac{L(\hat{\boldsymbol{\theta}}_{max}; \mathbf{y})}{L(\hat{\boldsymbol{\theta}}; \mathbf{y})}$$

is a way of assessing the goodness of fit for the model. In practice

$$\log \lambda = \ell(\hat{\boldsymbol{\theta}}_{max}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}; \mathbf{y})$$

is used. *Large values of* $\log \lambda$ *suggest that the model of interest is a poor description of the data relative to the saturated model.*

The **Deviance** or **log likelihood ratio statistic** is defined as

$$D = 2[\ell(\hat{\boldsymbol{\theta}}_{max}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}; \mathbf{y})].$$

We know that

$$\ell(\boldsymbol{\theta}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) = -\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

or, equivalently,

$$2[\ell(\hat{\boldsymbol{\theta}};\mathbf{y})-\ell(\boldsymbol{\theta};\mathbf{y})]=(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top\mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})$$

and consequently the asymptotic distribution of this statistic is

$$2[\ell(\hat{\boldsymbol{\theta}};\mathbf{y})-\ell(\boldsymbol{\theta};\mathbf{y})]\dot{\sim}\chi^2_p$$

We now have:

$$\begin{aligned}
D &= 2[\ell(\hat{\boldsymbol{\theta}}_{max};\mathbf{y})-\ell(\hat{\boldsymbol{\theta}};\mathbf{y})] \\
&= 2[\ell(\hat{\boldsymbol{\theta}}_{max};\mathbf{y})\pm\ell(\boldsymbol{\theta}_{max};\mathbf{y})\pm\ell(\boldsymbol{\theta};\mathbf{y})-\ell(\hat{\boldsymbol{\theta}};\mathbf{y})] \\
&= 2[\ell(\hat{\boldsymbol{\theta}}_{max};\mathbf{y})-\ell(\boldsymbol{\theta}_{max};\mathbf{y})]-2[\ell(\hat{\boldsymbol{\theta}};\mathbf{y})-\ell(\boldsymbol{\theta};\mathbf{y})] \\
&\quad +2[\ell(\boldsymbol{\theta}_{max};\mathbf{y})-\ell(\boldsymbol{\theta};\mathbf{y})]
\end{aligned}$$

- The **first term** $2[\ell(\hat{\boldsymbol{\theta}}_{max};\mathbf{y})-\ell(\boldsymbol{\theta}_{max};\mathbf{y})]$ has distribution χ^2_m, where m is the number of parameters in the saturated model
- The **second term** $2[\ell(\hat{\boldsymbol{\theta}};\mathbf{y})-\ell(\boldsymbol{\theta};\mathbf{y})]$ has distribution χ^2_p, where p is the number of parameters in the model of interest
- The **third term** $2[\ell(\boldsymbol{\theta}_{max};\mathbf{y})-\ell(\boldsymbol{\theta};\mathbf{y})]$ is a positive constant, which is zero if the model of interest has a fit which is as good as the saturated model; it can be considered a (usually negligible) non-centrality parameter

Therefore, the sampling distribution of the deviance is

$$D\dot{\sim}\chi^2_{m-p}(\nu) \tag{1.7.1}$$

where $\nu=2[\ell(\boldsymbol{\theta}_{\max};\mathbf{y})-\ell(\boldsymbol{\theta};\mathbf{y})]$ is a non-centrality parameter.

Remarks:

- The distribution is exact if the response variable is normally distributed
- For some other distributions, D can be calculated and used directly as a goodness of fit statistic

Example: Binomial distribution

If the response variables $\mathrm{Y}_1, \ldots, \mathrm{Y}_N$ are independent and $\mathrm{Y}_i \sim \mathrm{Bin}(n_i, p_i)$, then the log-likelihood is

$$\ell(\mathbf{p}; \mathbf{y}) = \sum_{i=1}^{N} \left[Y_i \log p_i - Y_i \log(1 - p_i) + n_i \log(1 - p_i) + \log \binom{n_i}{Y_i} \right]$$

For a **saturated** model, the p_i's are all different.

The MLE are $\hat{p}_i = \frac{\mathrm{Y}_i}{n_i}$ and

$$\ell(\hat{\boldsymbol{p}}_{max}; \mathbf{y}) = \sum_{i=1}^{N} \left[\mathrm{Y}_i \log\left(\frac{\mathrm{Y}_i}{n_i}\right) - \mathrm{Y}_i \log\left(\frac{n_i - \mathrm{Y}_i}{n_i}\right) + n_i \log\left(\frac{n_i - \mathrm{Y}_i}{n_i}\right) + \log \binom{n_i}{\mathrm{Y}_i} \right]$$

For any other model, the dimension of the parameter is $p < N$; let's call $\hat{p}_i^*$ *the MLE for a non-saturated model* and $\hat{Y}_i = n_i \hat{p}_i^*$ *the fitted values*; then

$$\ell(\hat{\boldsymbol{p}}^*; \mathbf{y}) = \sum \left[\mathrm{Y}_i \log\left(\frac{\hat{\mathrm{Y}}_i}{n_i}\right) - \mathrm{Y}_i \log\left(\frac{n_i - \hat{\mathrm{Y}}_i}{n_i}\right) + n_i \log\left(\frac{n_i - \hat{\mathrm{Y}}_i}{n_i}\right) + \log \binom{n_i}{\mathrm{Y}_i} \right]$$

And the **deviance** is

$$D = 2 \sum_{i=1}^{N} \left[\mathrm{Y}_i \log\left(\frac{\mathrm{Y}_i}{\hat{\mathrm{Y}}_i}\right) + (n_i - \mathrm{Y}_i) \log\left(\frac{n_i - \mathrm{Y}_i}{n_i - \hat{\mathrm{Y}}_i}\right) \right].$$

Nested model

We say that *model M_0 is nested in model M_1 if M_0 results as a special case of M_1*.

For instance, if we partition $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta}^\top = \left(\boldsymbol{\theta}^{(1)^\top}, \boldsymbol{\theta}^{(2)^\top}\right)$$

where $\boldsymbol{\theta}$ has length p and $\boldsymbol{\theta}^{(1)}$ has length q,

Then model M_1 could assume unrestricted $\boldsymbol{\theta}$, whereas M_0 restricts, e.g, $\boldsymbol{\theta}^{(2)} = \mathbf{0}$.

The **scaled deviance** can be used for model comparison.

For two nested linear models, the difference ΔD between the two deviance statistics generally follows a χ^2 distribution.
The degrees of freedom equal the difference in the dimensions of the two models, that is:

$$\Delta D = D_0 - D_1 = 2[\ell(\hat{\boldsymbol{\theta}}_{\max}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}_0; \mathbf{y})] - 2[\ell(\hat{\boldsymbol{\theta}}_{\max}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}_1; \mathbf{y})] = 2[\ell(\hat{\boldsymbol{\theta}}_1; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}_0; \mathbf{y}$$

Then $D_0 \sim \chi^2(N-q)$ and $D_1 \sim \chi^2(N-p)$, then

$$\Delta D \sim \chi^2(p-q)$$

when N is large.

If the values of ΔD is in the critical region, then we would reject H_0 in favour of H_1 on the grounds that model M_1 provides a significantly better description of the data.

Check your understanding

This is a non-assessed self-practice. Attempt the question below and press submit to be able to see the solution.

Question

[Dobson and Barnett (2018, Exercise 5.1)]

Consider the single response variable Y with $\mathrm{Y} \sim \mathrm{Bin}(n, \pi)$.

1. Find the Wald statistic $(\hat{\pi} - \pi)^\top \mathcal{I}(\hat{\pi} - \pi)$, where $\hat{\pi}$ is the maximum likelihood estimator of π and $\mathcal{I}$ the information.
2. Verify that the Wald statistic is the same as the score statistics $\mathrm{U}^\top \mathcal{I}^{-1} \mathrm{U}$.
3. Find the deviance $2\left(\ell(\hat{\pi}; y) - \ell(\pi; y)\right)$. Note that this is an adaptation of the deviance for the case where there is only one predictor and therefore no saturated/non-saturated models.
4. For large sample. both the Wald/score statistic and the deviance approximately have the χ^2_1 distribution. For $n = 10$ and $y = 3$, use both statistics to assess the adequacy of the models:
    - $\pi = 0.1$;
    - $\pi = 0.3$;
    - $\pi = 0.5$.

    Do the two statistics lead to the same conclusions?

No response