

# **Part II: Generalized Linear Models**

## **Chapter II.2**

### **Theory of Generalized Linear Models**

Model Fit Evaluation - Model Selection

# Evaluating Model Fit for GLMs

## II.2.32 Definition

Consider a random sample  $\mathbf{Y} = (Y_1, \dots, Y_n)'$  of  $n$  independent responses with  $Y_i \sim \text{EDF}(\vartheta_i, \phi)$ , modeled by a GLM  $\boldsymbol{\eta} = g(\mathbf{E}(\mathbf{Y})) = \mathbf{X}\boldsymbol{\beta}$ , denoted as  $\mathcal{M}$ , with (known) design matrix  $\mathbf{X} \in \mathbb{R}^{n \times p}$  and (unknown) parameter vector  $\boldsymbol{\beta} \in \Theta \subseteq \mathbb{R}^p$ .

Then

- $\hat{\boldsymbol{\mu}}$  denotes the MLE of  $\boldsymbol{\mu} = \mathbf{E}(\mathbf{Y})$  under  $\mathcal{M}$  and the vector  $\hat{\mathbf{Y}} = \hat{\boldsymbol{\mu}} = g^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}})$  is called the vector of predicted values,
- $\mathbf{e} = \hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \hat{\boldsymbol{\mu}}$  is the vector of *raw residuals*, and
- $\ell(\hat{\boldsymbol{\beta}}) = \ell(\hat{\boldsymbol{\mu}}; \mathbf{y})$  is the maximum log-likelihood corresponding to model  $\mathcal{M}$  evaluated at the MLE  $\hat{\boldsymbol{\mu}}$  (i.e.  $\hat{\boldsymbol{\beta}}$ ).

## II.2.33 Definition (saturated model)

A GLM that describes  $\mathbf{Y} = (Y_1, \dots, Y_n)'$  perfectly is called **saturated model** and is denoted by  $\mathcal{M}_{\text{sat}}$ . If  $\boldsymbol{\beta}_{\text{sat}}$  is the parameter vector of  $\mathcal{M}_{\text{sat}}$  and  $\hat{\boldsymbol{\mu}}_{\text{sat}}$  the corresponding vector of predicted values, then it holds  $\boldsymbol{\beta}_{\text{sat}} \in \Theta \subseteq \mathbb{R}^n$  (i.e.  $p = n$ ),  $\hat{\boldsymbol{\mu}}_{\text{sat}} = \mathbf{Y}$  and consequently  $\mathbf{e} = \mathbf{0}$ .

### ➤ II.2.34 Remark

- The quality of the fit of a model  $\mathcal{M}$  is assessed by comparing the maximum log-likelihood  $\ell(\hat{\boldsymbol{\mu}}; \mathbf{y})$  for  $\mathcal{M}$  to  $\ell_{sat} = \ell(\mathbf{y}; \mathbf{y})$ , which is the *saturated* maximum log-likelihood (corresponding to model  $\mathcal{M}_{sat}$ ).
- It is obvious that  $\ell(\hat{\boldsymbol{\mu}}; \mathbf{y}) < \ell_{sat}$  always, with the fit of model  $\mathcal{M}$  being as better as its log-likelihood approaches the saturated log-likelihood.

### II.2.35 Definition (Deviance)

The **goodness of fit** of a GLM  $\mathcal{M}$  is expressed, in terms of the log-likelihood difference, by the statistic  $-2[\ell(\hat{\boldsymbol{\mu}}; \mathbf{y}) - \ell_{sat}]$ . For the exponential dispersion family (s. slide 38) it equals

$$-2[\ell(\hat{\boldsymbol{\mu}}; \mathbf{y}) - \ell_{sat}] = 2 \left( \sum_{i=1}^n \frac{y_i \tilde{\theta}_i - b(\tilde{\theta}_i)}{a(\phi; i)} - \sum_{i=1}^n \frac{y_i \hat{\theta}_i - b(\hat{\theta}_i)}{a(\phi; i)} \right),$$

with  $\hat{\theta}_i$  and  $\tilde{\theta}_i$  the MLE of  $\theta_i$  under models  $\mathcal{M}$  and  $\mathcal{M}_{sat}$ , respectively.

Furthermore, for  $a(\phi; i) = \phi/w_i$  (s. Remark II.2.8), the statistic becomes

$$D_s(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \frac{D(\mathbf{y}; \hat{\boldsymbol{\mu}})}{\phi} = \frac{2}{\phi} \sum_{i=1}^n w_i \left( y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)] \right)$$

and is called the *scaled deviance*. The statistic  $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$  is known as **deviance**.

### II.2.36

Since  $\ell(\hat{\boldsymbol{\mu}}; \mathbf{y}) < \ell_{sat}$ , it always holds  $D(\mathbf{y}; \hat{\boldsymbol{\mu}}) \geq 0$ , with the fit of model  $\mathcal{M}$  becoming poorer as the deviance becomes greater.

## II.2.37 Remark (deviance of GLMs with canonical link)

- ① For normal GLMs, the deviance equals the residual sum of squares (SSE):

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2.$$

- ② For Poisson GLMs, the deviance equals the **likelihood ratio statistic (LRS)**  $G^2$  for testing model  $\mathcal{M}$  ( $H_0$ ) against the saturated model  $\mathcal{M}_{sat}$  ( $H_1$ ):

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = G^2(\mathcal{M}) = 2 \sum_{i=1}^n y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right).$$

- ③ For GLMs with response distributions satisfying  $\text{Var}(Y_i) = v(\mu_i)$ ,  $i = 1, \dots, n$  (s. Remark II.2.18), and  $\phi = 1$ , the corresponding *score statistic*<sup>a</sup> is of the form

$$W_S(\mathcal{M}) = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}.$$

For Poisson distributed  $Y_i$ s ( $v(\mu_i) = \mu_i$ ), it becomes the *Pearson's*  $X^2$  statistic.

---

<sup>a</sup>For the LRS, the score statistic and the corresponding asymptotic statistical tests see Theorems S1.59, S1.61 and S1.63 of the course 'Mathematics of Data Science'.

## > II.2.38 Remark

For normal GLMs and under the assumption that  $\mathcal{M}$  is true, the deviance and the score statistics are both exact  $\chi^2$  distributed.

For the other response distributions, these statistics can be used for testing model's  $\mathcal{M}$  goodness of fit, if their asymptotic distribution can be specified, which is not always the case. This is possible for example, for Poisson GLMs with large means  $\mu_i$ s or for binomial GLMs with  $Y_i \sim \mathcal{B}(m_i, \mu_i)/m_i$  having sufficiently large  $m_i$ s.

In this case,  $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$  and  $G^2(\mathcal{M})$ , for example, are both approximately  $\chi^2_{df}$  distributed, with  $df = n - p$  (e.g. sample size – number of parameters in the model).

# Model Complexity

These goodness-of-fit tests do not account for **model complexity** while they are increasing in sample size  $n$ , giving thus significant values for good models if the sample size is large.

Alternatively, the fit of a GLM can be evaluated by the criteria introduced below.

## II.2.39 Definition (AIC and BIC)

The fit of a model  $\mathcal{M}$  can be evaluated by

- **Akaike's Information Criterion** (Akaike, 1974):  $AIC = -2\ell(\hat{\mu}; \mathbf{y}) + 2p$ .  
It is based on the maximum likelihood under  $\mathcal{M}$  but penalizes its value for model complexity, where  $p$  is the number of parameters in the model. The better is a model, the smaller is the corresponding  $AIC$ .
- **Bayesian Information Criterion** (Schwarz, 1978):  $BIC = -2\ell(\hat{\mu}; \mathbf{y}) + (\log n)p$ , which is another maximum likelihood based measure, incorporating Bayesian thinking, that beyond complexity takes into account also the sample size.

# Model Selection in GLMs

Deviance plays a predominant role in comparing GLMs. Model selection is based on comparing *nested models*.

## II.2.40 Definition (nested models)

Let  $\mathcal{M}_1$  be a GLM having  $p_1$  parameters. Let also  $\mathcal{M}_0$  be a simpler GLM, produced from  $\mathcal{M}_1$  by eliminating  $r$  of its  $p_1$  parameters. Then,  $\mathcal{M}_0$  is said to be *nested* in  $\mathcal{M}_1$  and denoted by  $\mathcal{M}_0 \subset \mathcal{M}_1$ . Model  $\mathcal{M}_0$  is more parsimonious than  $\mathcal{M}_1$  having  $p_0 = p_1 - r$  parameters.

## II.2.41 Remark

If the response distributions are in the EDF with  $\phi = 1$  (like Poisson and binomial responses), then the deviance of a model is equal to the corresponding LRS for testing its fit.

If  $\hat{\mu}_0$  and  $\hat{\mu}_1$  are the MLEs of  $\mu$  under  $\mathcal{M}_0$  and  $\mathcal{M}_1$ , respectively, then, for  $\phi = 1$ , the deviances of models  $\mathcal{M}_0$  and  $\mathcal{M}_1$  are

$$D(\mathbf{y}; \hat{\mu}_j) = -2 [\ell(\hat{\mu}_j; \mathbf{y}) - \ell(\mathbf{y}; \mathbf{y})], \quad j = 0, 1.$$

If  $(\mathcal{M}_0 \subset \mathcal{M}_1)$ , since reducing the number of model's parameters implies increase of model's distance from the perfect fit of the saturated model, it will always be  $D(\mathbf{y}; \hat{\mu}_0) > D(\mathbf{y}; \hat{\mu}_1)$ .



### ► II.2.42 Comparison of two nested models

Consider models  $\mathcal{M}_0$  and  $\mathcal{M}_1$  with  $\mathcal{M}_0 \subset \mathcal{M}_1$ , for response distributions in EDF with  $\phi = 1$ , both applied on the same observations  $\mathbf{y}$ . Then the difference in their deviances is

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1) = -2 [\ell(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}_1; \mathbf{y})] = G^2(\mathcal{M}_0, \mathcal{M}_1) ,$$

where  $G^2(\mathcal{M}_0, \mathcal{M}_1)$  is the LRS for testing the null hypothesis that  $\mathcal{M}_0$  holds against the alternative that  $\mathcal{M}_1$  holds.

### ► II.2.43 Remark

In particular, if the appropriate assumptions hold (see Remark II.2.38), then the difference in deviances

$$D(\hat{\boldsymbol{\mu}}_0; \hat{\boldsymbol{\mu}}_1) = D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1),$$

is under  $\mathcal{M}_0$  approximately  $\chi_r^2$  distributed, where  $r = p_1 - p_0$  is the difference between the number of parameters of the two compared models. This result is the key for asymptotic comparison of models.

► For Poisson log-linear models,  $D(\hat{\boldsymbol{\mu}}_0; \hat{\boldsymbol{\mu}}_1)$  simplifies to

$$G^2(\mathcal{M}_0 | \mathcal{M}_1) = 2 \sum_{i=1}^n y_i \log \left( \frac{\hat{\mu}_{i1}}{\hat{\mu}_{i0}} \right) = G^2(\mathcal{M}_0) - G^2(\mathcal{M}_1) .$$

Normal GLM:  
(canonical link)  
↪ identity

$$\hat{\theta}_i = \hat{\mu}_i, \quad b(\hat{\theta}_i) = \hat{\theta}_i^2/2, \quad \alpha(\phi) = \sigma^2$$

$$\tilde{\theta}_i = y_i, \quad b(\tilde{\theta}_i) = y_i^2/2$$

$$D(y; \hat{\mu}) = 2 \sum_{i=1}^n [y_i (\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]$$

$$= 2 \sum_{i=1}^n \left[ y_i (y_i - \hat{\mu}_i) - \frac{y_i^2}{2} + \frac{\hat{\mu}_i^2}{2} \right] = \dots = \underbrace{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2}_{= SSE}$$

Poisson GLM:  $\hat{\theta}_i = \log(\hat{\mu}_i)$ ,  $b(\hat{\theta}_i) = e^{\hat{\theta}_i} = \hat{\mu}_i$

↪ log-link

$$\tilde{\theta}_i = \log(y_i), \quad b(\tilde{\theta}_i) = y_i$$

$$\alpha(\phi) = \frac{\phi}{w_i}, \quad \phi = w_i = 1$$

$$D(y; \hat{\mu}) = 2 \sum_{i=1}^n \left[ y_i \log(y_i / \hat{\mu}_i) - y_i + \hat{\mu}_i \right] = 2 \sum_{i=1}^n y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + 2 \sum_{i=1}^n (\hat{\mu}_i - y_i)$$

II. 2.42:

$$G^2(\mathcal{M}_0, \mathcal{M}_1) = -2 [\ell(\hat{\mu}_0; y) - \ell(\hat{\mu}_1; y)] \\ + 2 [\ell(y; y) - \ell(y; y)]$$

$$= -2 [\ell(\hat{\mu}_0; y) - \ell_{\text{sat}}] - 2 [\ell(\hat{\mu}_1; y) - \ell_{\text{sat}}]$$

Rem. II 2.41

$$= D(y; \hat{\mu}_0) - D(y; \hat{\mu}_1)$$

II 2.43:

$$G^2(\mathcal{M}_0 | \mathcal{M}_1) = -2 \sum_{i=1}^n \hat{y}_i \log \left( \frac{\hat{\mu}_{i0}}{\hat{\mu}_{i1}} \right) = 2 \hat{\mu}_1' (\log \hat{\mu}_1 - \log \hat{\mu}_0)$$



$$G^2(\mathcal{M}_0) - G^2(\mathcal{M}_1) = -2 y' (\log \hat{\mu}_0 - \log y) + 2 y' (\log \hat{\mu}_1 - \log y) \\ = 2 y' (\log \hat{\mu}_1 - \log \hat{\mu}_0)$$

$$\Rightarrow G^2(\mathcal{M}_0) - G^2(\mathcal{M}_1) = 2 y' (X_1 \hat{\beta}_1 - \underbrace{X_0 \hat{\beta}_0}_{= \log \hat{\mu}_1})^{\otimes} \quad \text{with } \beta_1' = (\beta_0', \beta_1')$$

↳ Expand  $\beta_0$  to  $\beta_1^* = \begin{pmatrix} \beta_0 \\ 0 \end{pmatrix}$  with  $0$  a 0-vector of length  $p_1 - p_0$

$$\hookrightarrow X_0 \hat{\beta}_0 = X_1 \hat{\beta}_1^*$$

$$\otimes \Rightarrow G^2(\mathcal{M}_0) - G^2(\mathcal{M}_1) \stackrel{**}{=} 2 \hat{\mu}_1' (X_1 \hat{\beta}_1 - X_1 \hat{\beta}_1^*)$$

$**$  Likelihood Eq.:

$$y' X_1 = \hat{\mu}_1' X_1$$

# 'Best Model' Selection

## ▶ II.2.44 Remark


Upon considering a sequence of nested models from a very simple  $\mathcal{M}_0$  up to the saturated  $\mathcal{M}_{\text{sat}}$ ,

$$\mathcal{M}_0 \subset \mathcal{M}_1 \subset \mathcal{M}_2 \subset \dots \subset \mathcal{M}_{\text{sat}} ,$$

the importance of the parameters added gradually can be evaluated by successive comparisons of neighbor models, in a **stepwise** manner, moving “*forwards*” or “*backwards*”.

Hence, comparisons of nested models serve for developing procedures of ‘best model’ selection. Furthermore, once the ‘best model’ is selected, models comparison can serve as a tool for evaluating the individual importance of each parameter or group of parameters.

Model selection can also be based on *AIC* or *BIC*.

 These issues will be discussed and illustrated in the context of log-linear models for multi-way tables.