Lecture 11: Model Adequacy, Deviance

(Text Sections 5.5-5.7)

Deviance is an important idea associated with a fitted GLM. It can be used to test the fit of the link function and linear predictor to the data, or to test the significance of a particular predictor variable (or variables) in the model. The following discussion applies to GLMs where the random component is in the 1-parameter exponential family (e.g., Poisson and binomial).

To understand deviance, we first need to return to the concept of a *saturated model*. To review, let Y_{ik} be the response observed on the k^{th} replicate of the i^{th} distinct combination of covariate values $(x_{i1}, \ldots, x_{ip})$, $k = 1, \ldots, n_i$ and $i = 1, \ldots, K$. Here $n \equiv \sum_{i=1}^{K} n_i$ and $K \leq n$. K is the maximum number of parameters that we can estimate from these data. Let $\mu_i \equiv \mathrm{E}[Y_{ik}]$. The model of interest specifies a distribution for Y_{ik} and $g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j \equiv \eta_i$ where $p \leq K$ is the number of parameters to be estimated. In other words, this model *constrains* the means μ_i to lie on the surface given by η_i.

In the GLM setting we define the saturated model as the GLM with the same distribution and link function as the model of interest, but with $g(\mu_i) = \psi_i$ for $i = 1, \ldots, K$. In other words, the saturated model allows a different mean response ($\mu_i = g^{-1}(\psi_i)$) for each group of replicates, and hence has K parameters to be estimated. We will denote this vector of parameters by $\boldsymbol{\psi}$. We can think of the saturated model as having the most general possible mean structure for the data since the means μ_i are *unconstrained*[1]. The saturated model is also referred to as the *full model* or *maximal model*.

With respect to testing the fit of the link function and linear predictor, suppose we have fit a GLM to some data. One way to assess the adequacy of the model is to compare this fitted model with the fitted saturated model. Let $\mathcal{L}_S(\boldsymbol{\psi}; \mathbf{y})$ and $\mathcal{L}(\boldsymbol{\beta}; \mathbf{y})$ be the likelihoods corresponding to the saturated and proposed model, respectively. We know that $\mathcal{L}_S(\boldsymbol{\psi}; \mathbf{y}) \geq \mathcal{L}(\boldsymbol{\beta}; \mathbf{y})$ since the model of interest is a special case of the saturated model. Comparing $\mathcal{L}_S(\boldsymbol{\psi}; \mathbf{y})$ and $\mathcal{L}(\boldsymbol{\beta}; \mathbf{y})$ – or equivalently $l_S(\boldsymbol{\psi}) \equiv \log \mathcal{L}_S(\boldsymbol{\psi}; \mathbf{y})$ and $l(\boldsymbol{\beta}) \equiv \log \mathcal{L}(\boldsymbol{\beta}; \mathbf{y})$ – is one means of assessing how well our assumed link function and form of the linear predictor fit the data. In particular, we would expect that $l(\boldsymbol{\beta}) \approx l_S(\boldsymbol{\psi})$ if the model is suitable in these respects.

We define the *deviance* or *likelihood ratio statistic*, D, as

$$D = 2[l_S(\hat{\boldsymbol{\psi}}) - l(\hat{\boldsymbol{\beta}})],$$

where $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\beta}}$ are the MLEs of the saturated and proposed model, respectively.

Under some conditions (which we will discuss for each type of response data individually),

[1] Some authors – and S-PLUS – form the saturated model*as if there were no replicates*, i.e. allowing n different parameters, one for each observation. These authors might argue that there are never any true replicates, e.g., experimental conditions always change slightly from one run to the next.

if the proposed model describes the data nearly as well as the saturated model, then asymptotically

$$D \sim \chi^2_{K-p},$$

where K and p are the number of parameters in the saturated and proposed models, respectively. If the proposed model is poor, D will be larger than predicted by the χ^2_{K-p} distribution.

Example: Poisson GLM

GOAL: Derive D for samples $Y_1, \ldots, Y_n$ from the Poisson(λ_i) distribution with

$$\log(\lambda_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

where the vectors $(x_{i1}, \ldots, x_{ip})$ are all distinct.

The log-likelihood for the model of interest is

$$l(\beta) = \sum_{i=1}^{n} y_i \log \lambda_i - \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \log y_i!,$$

where $\lambda_i = \exp\left\{\sum_{j=1}^{p} x_{ij}\beta_j\right\}$. The maximum value of this function occurs at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

For the maximal model, we have parameters $\mu_1, \ldots, \mu_n$, where $\mu_i = \mathrm{E}[Y_i]$. The log-likelihood for this model is

$$\begin{aligned} l_S(\boldsymbol{\mu}) &= \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \log y_i! \\ &= \left(\sum_{i\in C} y_i \log \mu_i - \sum_{i\in C} \mu_i - \sum_{i\in C} \log y_i!\right) - \sum_{i\in C^c} \mu_i \end{aligned}$$

where $C = \{i : Y_i > 0\}$. Since we require $\hat{\mu}_i \geq 0$ for all i, we can see by inspection that $\hat{\mu}_i = 0$ for all $i \notin C$. For $i \in C$, we compute $\hat{\mu}_i$ in the usual way:

$$\begin{aligned} \frac{\partial}{\partial \mu_i} l_S(\boldsymbol{\mu}) &= \frac{y_i}{\mu_i} - 1 \\ 0 &= \frac{y_i}{\hat{\mu}_i} - 1 \\ \hat{\mu}_i &= y_i. \end{aligned}$$

Therefore,

$$l_S(\hat{\boldsymbol{\mu}}) = \sum_{i\in C} y_i \log y_i - \sum_{i\in C} y_i - \sum_{i\in C} \log y_i!.$$

The deviance is then

$$\begin{aligned} D &= 2[l_S(\hat{\boldsymbol{\mu}}) - l(\hat{\boldsymbol{\beta}})] \\ &= 2\left[\sum_{i\in C} y_i \log y_i - \sum_{i=1}^{n} y_i \log \hat{\lambda}_i - \sum_{i=1}^{n} \left(y_i - \hat{\lambda}_i\right)\right] \end{aligned}$$

Here, D does not contain unknown parameters. So, it can be calculated completely from the data. If D is large relative to the χ^2_{n-p} distribution, then we have evidence against the null hypothesis that our model fits the data well. In other words, the maximal model explains significantly more variation in the data than our (smaller) model of interest. □

Example: Normal (linear) model

GOAL: Derive D for samples $Y_1, \ldots, Y_n$ from the $\mathrm{N}(\mu_i, \sigma^2)$ distribution with

$$\mu_i = \sum_{j=1}^{p} x_{ij}\beta_j,$$

where σ^2 is known and the vectors $(x_{i1}, \ldots, x_{ip})$'s are distinct.

The log-likelihood for the model of interest is

$$l(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2 - \frac{n}{2}\log(2\pi\sigma^2)$$

The maximum value of this function occurs at $\hat{\mu}_i = \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$.

For the maximal model, we have parameters $\theta_1, \ldots, \theta_n$, where $\theta_i = \mathrm{E}[Y_i]$. The log-likelihood for this model is

$$l_S(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta_i)^2 - \frac{n}{2}\log(2\pi\sigma^2)$$

So,

$$\begin{aligned}
\frac{\partial}{\partial\theta_i} l_S(\boldsymbol{\theta}) &= \frac{1}{\sigma^2}(y_i - \theta_i) \\
0 &= y_i - \hat{\theta}_i \\
\hat{\theta}_i &= y_i.
\end{aligned}$$

Therefore,

$$l_S(\hat{\boldsymbol{\theta}}) = -\frac{n}{2}\log(2\pi\sigma^2)$$

The deviance is then

$$\begin{aligned}
D &= 2[l_S(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\beta}})] \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2
\end{aligned}$$

Note the connection with the RSS! Under the null hypothesis that the model fits the data well, $D \sim \chi^2_{n-p}$ *exactly* (in this case). □

The above deviance test compares the proposed model to the most general (saturated) model, and hence asks the question, "Can we use a more parsimonious model to describe the data as well as the most general model does?"

Another test of interest concerns the *null model*. In this context, the null model is defined as the GLM with the same distribution and link function as the proposed model, but with $g(\mu_i) = \phi$ for some parameter ϕ. In other words, the null model assumes that all observations have the *same distribution* with common parameter ϕ. Like the saturated model, the null model does not depend on predictor variables. While the saturated most is the most general model, the null model is the most restricted model.

We can use the deviance to compare the proposed model to the null model. This test asks the question, "Do we gain a significant improvement in fit by including predictor variables in the model?" Specifically, if $l_N(\hat{\theta})$ is the maximized log-likelihood of the null model, then asymptotically,

$$2[l(\hat{\boldsymbol{\beta}}) - l_N(\hat{\theta})] \sim \chi^2_{p-1},$$

where p and 1 are the number of parameters in the proposed and null models, respectively.

Similarly, we can use the *change in deviance*, much like the change in residual sums of squares, to test the significance of one or more predictor variables. In particular, let the reduced model be the same as the proposed model, but excluding the predictor variables of interest. Let $\boldsymbol{\beta}^*$ be the vector of parameters associated with the reduced model, and let $l_R(\hat{\boldsymbol{\beta}}^*)$ be the maximized log-likelihood. Then asymptotically,

$$2[l(\hat{\boldsymbol{\beta}}) - l_R(\hat{\boldsymbol{\beta}}^*)] \sim \chi^2_{p-q},$$

where p and q are the number of parameters in the proposed and reduced models, respectively.

S-PLUS gives the deviance (labelled "residual deviance") in the summary output . It also gives the "null deviance", which is

$$ND = 2[l_S(\hat{\boldsymbol{\phi}})] - l_N(\hat{\phi})].$$

This quantity is not useful in itself. However,

$$ND - D = 2[l(\hat{\boldsymbol{\beta}}) - l_N(\hat{\phi})],$$

which is the quantity we need for comparing the proposed model to the null model. The `anova` command provides the changes in deviance associated with dropping parameters sequentially from the model in the column labelled `Deviance` (similar to the usual ANOVA table, which provides changes in sums of squares).

NOTE: If there are replicates in the data set, you should **ignore** the "residual deviance" given by S-PLUS and compute it yourself. Specifically, you should fit the saturated model by defining a factor with one level for each distinct combination of covariate values. You should then compare (2 times) the maximized log-likelihood of this model to that of the proposed model.