Tutorial 8

STAT3015/4030/7030 Generalised Linear Modelling

The Australian National University

Week 8, 2017

Overview

Deviance

In Tutorial 6 we have seen that

$$\textit{deviance} = \textit{Constant} - 2 \times \log(\textit{Maximum Likelihood}).$$

Now we formally introduce *deviance* or *residual deviance*, $D(\hat{Y}, Y)$, defined by

$$D(\hat{Y}, Y) = 2\phi\{\ell(Y, \phi) - \ell(\hat{Y}, \phi)\},$$

which measures the (scaled) difference between the log-likelihood for the **observed data** and the log-likelihood of the the **fitted values**, and thus small values of the deviance indicate that a model fits the observed data well.

Deviance

In the previous silde, the log-likelihood function

$$\ell(\mu, \phi) = \sum_{i=1}^{n} \left\{ \frac{Y_i b(\mu_i) - c(\mu_i)}{\phi} + d(Y_i, \phi) \right\}$$

is calculated based on the types of GLM fitted.

In **RStudio**, we can use “model$deviance” to extract deviance.

Scaled deviance

For independent observations Y_i and exponential family errors, we have

$$D(\hat{Y}, Y) = 2\sum_{i=1}^{n}\{Y_i(\hat{\theta}_{saturated} - \hat{\theta}) - b(\hat{\theta}_{saturated}) + b(\hat{\theta})\}.$$

(exponential family and $b(\cdot)$ functions on Page 33 of the brick)

Then we can write a likelihood ratio statistic of comparing a saturated model and the model of interest as

$$\text{Likelihood ratio} = D^* = \frac{D(\hat{Y}, Y)}{\phi}$$

Dispersion

The dispersion parameter ϕ indicates if we have more or less than the expected variance. We have already seen that $\phi = 1$ for Binomial and Poisson distributions. In the **summary** output we have **dispersion** parameter defined as

$$\phi_{assumed} = \begin{cases} MSE = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, & \text{Normal} \\ 1, & \text{Binomial and Poisson} \\ CV = \frac{1}{n-p}\sum_{i=1}^{n}(\frac{Y_i - \hat{Y}_i}{\hat{Y}_i})^2, & \text{Gamma} \end{cases}$$

where CV is the estimated coefficient of variation (relative standard deviation) for the gamma distribution.

Alternative estimates of dispersion

An alternative estimate of ϕ for all GLMs is

$$\phi_{alt} = \frac{D(\hat{Y}, Y)}{n - p}.$$

If $\phi_{alt} = \phi_{assumed} \longrightarrow$ model is "good".
If $\phi_{alt} < \phi_{assumed} \longrightarrow$ model is **under-dispersed**.
If $\phi_{alt} > \phi_{assumed} \longrightarrow$ model is **over-dispersed**.

Goodness of fit test

We can also use deviance to assess model fit.

If dispersion ϕ is known (Poisson and Binomial GLM),

$$\frac{D(\hat{Y}, Y)}{\phi} \sim \chi^2_{n-p} \quad \text{under } H_0$$

$$D(\hat{Y}_S, Y) - D(\hat{Y}_L, Y) \sim \chi^2_{df_S - df_L} \quad \text{under } H_0$$

The difference in deviance between two “nested” models is a measure of how much better the “larger” model is at fitting the data.

Q1

(a) Don't forget to include "**weights**" in fitting the model as the number of months in service are different.

(b) It is good to know how to use matplot function. `help(matplot)`

(c) Use "*" to create the interaction term.

(d) A good explanation of the difference between Wald's test and Student's *t* test can be found here.

Q2

(a) The dispersion for the exponential distribution is $1/\alpha$ with $\alpha = 1$. Use this fact to test the assumption of an exponential distribution.
α is a parameter in Gamma distribution,
e.g., $f(x) = \beta^\alpha / \Gamma(\alpha) x^{\alpha-1} \exp(-\beta x)$.

(b) Time needed for the concentration to be halved. Solve

$$0.5\exp(\alpha) = \exp(\alpha + \beta_T t),$$

get

$$t = \log(0.5)/\beta_T.$$

(c) Fit indicators of temperature (as a combined term) last in the model and check summary output.

$$\frac{D(\hat{Y}, Y)}{\phi} \sim \chi^2_{n-p} \quad \text{under } H_0.$$