STAT3015/4030/7030:
Generalised Linear Modelling
GLMS - Analysis of Deviance

Semester 2 2017

Originally prepared by Bronwyn Loong

Goodness of Fit and Deviance Statistic

How do we decide on the final model for our data? Consider the range of possibilities

- Null model -

Goodness of Fit and Deviance Statistic

How do we decide on the final model for our data? Consider the range of possibilities

- Null model - no relationship between predictors and response. That is, fit a common mean μ for all y, and the data is represented entirely as random variation.
- Full/Saturated model -

Goodness of Fit and Deviance Statistic

How do we decide on the final model for our data? Consider the range of possibilities

- Null model - no relationship between predictors and response. That is, fit a common mean μ for all y, and the data is represented entirely as random variation.
- Full/Saturated model - data is explained exactly, estimate n parameters with exactly n data points. Drawbacks?

Goodness of Fit and Deviance Statistic

How do we decide on the final model for our data? Consider the range of possibilities

- Null model - no relationship between predictors and response. That is, fit a common mean μ for all y, and the data is represented entirely as random variation.
- Full/Saturated model - data is explained exactly, estimate n parameters with exactly n data points. Drawbacks? the model tells us no more than the data itself, uninformative.

Goodness of Fit and Deviance Statistic

Likelihood ratio statistics for comparison between full (saturated) model and the model under consideration:

$$\mathrm{2logLRT} = 2\mathrm{log}\Lambda = 2(\log L(\hat{\theta}_{sat}, \phi|y) - \log L(\hat{\theta}, \phi|y)) \sim \chi^2_\nu$$

ν: difference in number of parameters between two models

Provided observations are independent and for EF distributions this simplifies to

$$\mathrm{Scaled\ \ Deviance} = 2\mathrm{log}\Lambda = 2\sum_i^n \frac{(y_i(\hat{\theta}_{sat} - \hat{\theta}) - b(\hat{\theta}_{sat}) + b(\hat{\theta}))}{\phi} \tag{1}$$

$$= \frac{D(Y, \hat{Y})}{\phi} \tag{2}$$

$D(Y, \hat{Y})$ is called the *deviance*. $D^* = \frac{D(Y,\hat{Y})}{\phi}$ is called the *scaled deviance*.

Goodness of Fit and Deviance Statistic

Exercise: What is the deviance for the linear regression model with identity link and normal error structure, and fitted values $\hat{Y}_i$.

Goodness of Fit and Deviance Statistic

We showed previously that for the normal distribution, $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \theta^2/2$.

$$
\begin{aligned}
D(Y, \hat{Y}) &= \sum_i 2(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)) \\
&= \sum_i 2\left(y_i(y_i - \hat{y}_i) - \frac{y_i^2}{2} + \frac{\hat{y}_i^2}{2}\right) \\
&= \sum_i 2y_i^2 - 2y_i\hat{y}_i - y_i^2 + \hat{y}_i^2 \\
&= \sum_i y_i^2 - 2y_i\hat{y}_i + \hat{y}_i^2 \\
&= \sum_i (y_i - \hat{y}_i)^2 = SSE
\end{aligned}
$$

Goodness of Fit and Deviance Statistic

Alternatively, working from first principles, and noting that the log-likelihood for $Y \sim N(\mu, \sigma^2)$ is

$$\log L(Y|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 - \frac{n}{2}\log(2\pi\sigma^2)$$

$$\begin{aligned} D(Y, \hat{Y}) = 2\log\Lambda \times \sigma^2 &= 2\sigma^2\left\{-\frac{1}{2\sigma^2}\sum_i (y_i - y_i)^2 - n\phi\log(2\pi\sigma^2)\right\} \\ &+ 2\sigma^2\left\{-\frac{1}{2\sigma^2}\sum_i (y_i - \hat{y}_i)^2 + n\phi\log(2\pi\sigma^2)\right\} \\ &= \sum_i (y_i - \hat{y}_i)^2 = SSE \end{aligned}$$

Goodness of Fit and Deviance Statistic

A larger value of $D(Y, \hat{Y})$ suggests a poorer fit. How can we use the deviance statistic to assess model fit?

If ϕ is known,

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

So if $\phi = 1$, then

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

Goodness of Fit and Deviance Statistic

In R, $D(Y, \hat{Y})$ is reported as the residual deviance. Because ϕ must be known, we can only apply the goodness of fit test to Poisson and Binomial GLMS where ϕ=1. For GLMS such as the Gaussian and Gamma, we usually do not know the value of the dispersion parameter ϕ, and so this test cannot be used. Though, as we will see later, we can often estimate the disperison and use a suitably scaled deviance test.

IMPORTANT: the goodness-of-fit test is not appropriate for binary data (that is, observations are reported as Y=0,1). Why? Because the chi-squared approximation requires expected frequencies of success for each observation to exceed one, this is not possible with the definition of binary data.

Hypothesis Testing

Like the sums of squared errors under the least-squares approach to parameter estimation, the deviance measures the deviation of the fitted values from the observed data.

Similarly, we will use a *difference in deviances* measure to test if a larger model provides a significantly better fit to the data.

The difference in deviance is a measure of how important the subset of predictors which have been left out of the smaller model are to explain relationships in the observed data.

Hypothesis Testing

To compare a large model "L" to a small model "S", we calculate the difference in deviances
Suppose the extra parameters in the larger model are denoted by $\beta_{\bar{S}}$.

$$H_0 : \beta_{\bar{S}} = 0 \text{ vs } H_A : \beta_{\bar{S}} \neq 0$$

1. Dispersion ϕ known (Poisson and Binomial GLM)
For α significance level, reject H_0 if

$$D(Y, \hat{Y}_S) - D(Y, \hat{Y}_L) \geq \chi^2_{(df_S - df_L)}(1 - \alpha)$$

That is, the difference in deviances $D(Y, \hat{Y}_S) - D(Y, \hat{Y}_L)$ is asymptotically χ^2 with degrees of freedom equal to the difference in the number of identifiable parameters in the two models.

Hypothesis Testing

2. Dispersion ϕ unknown (Normal and Gamma GLM)
The χ^2 test for nested models cannot be used if ϕ must be estimated. Where ϕ is usually not known, we can insert an estimate for ϕ and compute an F-statistic of the form

$$F = \frac{(D(Y, \hat{Y}_S) - D(Y, \hat{Y}_L))/(df_S - df_L)}{\hat{\phi}_L}$$

Compare the F-stat to the quantiles of a $F_{df_S - df_L, df_L}$ distribution.

For the normal linear model (Gaussian GLM) the F-statistic has an exact F-distribution. For other GLMs with a free dispersion parameter that must be estimated, the statistic is only approximately F distributed.

Hypothesis Testing

Exercise: Specify the drop in deviance test statistic for the overall significance of a normal linear multiple regression with (p-1 predictors). That is, testing the null hypothesis
($H_0 : \beta_1 = ... = \beta_{p-1} = 0$)

Hypothesis Testing

In this case the smaller model is the null model with just an intercept term $Y_i = \beta_0 + \epsilon_i$. The fitted values are $\hat{Y}_i = \bar{Y}$ for all data points. So $D(Y, \hat{Y}_S) = \sum_i (Y_i - \bar{Y})^2 = SST$.

The larger model has fitted values $\hat{Y}_i = X_i^T \hat{\beta}$. So $D(Y, \hat{Y}_L) = \sum_i (Y_i - \hat{Y}_i)^2 = SSE$.

Hence, we have

$$\frac{(SST - SSE)/(p-1)}{\hat{\phi}_L} = \frac{SSR/(p-1)}{\hat{\sigma}_L^2} = \frac{MSR}{MSE} = F$$

So the deviance statistic is equivalent to the standard F-statistic in multiple linear regression to test for significance of subsets of predictors.

Hypothesis Testing

(R code example - Weighted regression on pages 46 to 48 of the brick)

(R code example - Car insurance data on pages 48 to 54 of the brick)

(Additional R code example - Bliss insect data)