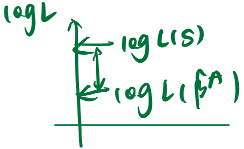
Generalised Linear Models (GLMs) II

Deviance

Deviance



From here on we will w.l.o.g. adopt the convention that $a(\phi) = \phi$.

Definition: the scaled deviance for model A is

$$\frac{D^A}{\phi} = -2 \log \frac{\mathcal{L}(\hat{\beta}^A)}{\mathcal{L}(S)}$$

where $\hat{\beta}^A$ is the MLE of β^A, the parameters in the model A, and $\mathcal{L}(S)$ is the maximum likelihood for the saturated model

The deviance is just D^A.

$\phi \cdot$ scaled deviance.

Example: normal

model A $\hat{\beta}^A \Rightarrow \hat{\theta}_i^A$ ϕ assume known

for all GLM $\mu_i = b'(\theta_i) = g^{-1}(x_i^T\beta)$. model saturated s. $\hat{\beta}^s = \hat{\theta}_i^s$ ϕ assume known

The saturated normal model uses y_i to estimate μ_i.

The deviance can be written as $D = \sum_i d_i$ where

$$d_i = (y_i - \hat{\mu}_i)^2$$

where $\hat{\mu}_i$ is the fitted mean using the MLE.

$$\frac{D^A}{\phi} = 2\left[\log^{L(S)} - \log^{L(A)}\right]$$

$$= 2\left[\sum_{i=1}^{n}\left[\frac{y_i\hat{\theta}_i^s - b(\hat{\theta}_i^s)}{\phi} + \cancel{c(y_i,\phi)}\right]\right.$$

$$\left. - \sum_{i=1}^{n}\left[\frac{y_i\hat{\theta}_i^A - b(\hat{\theta}_i^A)}{\phi} + \cancel{c(y_i,\phi)}\right]\right]$$

It is equivalent to SSE (sum of squared errors) in linear models.

$$= \frac{2}{\phi}\sum_{i=1}^{n}\left[y_i(\hat{\theta}_i^s - \hat{\theta}_i^A) - \left(b(\hat{\theta}_i^s) - b(\hat{\theta}_i^A)\right)\right]$$

or normal $\theta_i = \mu_i$ $b(\theta_i) = \frac{1}{2}\theta_i$

$\hat{\theta}_i^A = \hat{\mu_i} = x^T\beta$

or saturated $\hat{\theta}_i^s = y_i$

$$\sum_{i=1}^{n}[y_i(y_i - \hat{\mu}_i) - \tfrac{1}{2}(y_i^2 - \hat{\mu}_i^2)] = \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \rightarrow \text{scaled deviance}$$

$$D^A = \phi \cdot \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

Example: Poisson

exercise

The saturated Poisson model uses y_i to estimate $\mu_i = \lambda_i$.

The deviance can be written as $D = \sum_i d_i$ where

$$d_i = -2\left(y_i \log \frac{\hat{\mu}_i}{y_i} - (\hat{\mu}_i - y_i)\right)$$

where $\hat{\mu}_i$ is the fitted mean using the MLE.

exercise

Example: binomial

The saturated binomial model uses y_i to estimate μ_i, that is y_i/m_i to estimate p_i.

The deviance can be written as $D = \sum_i d_i$ where

$$d_i = -2\left(y_i \log \frac{\hat{\mu}_i}{y_i} + (m_i - y_i) \log \frac{m_i - \hat{\mu}_i}{m_i - y_i}\right)$$

where $\hat{\mu}_i$ is the fitted mean using the MLE.

Scaled deviance for testing model adequacy

The binomial and Poisson models:

- $\phi = 1$ and the scaled deviance is just the deviance.
- When the individual responses are somewhat normal, if the model is adequate then the scaled deviance will be $\approx \chi^2_{n-p}$, where the saturated model has n parameters (equal to the number of observations) and the fitted model has p parameters.
- As a rule of thumb we need $m_i p_i$ and $m_i(1 - p_i) \geq 5$ for the binomial model and $\lambda_i \geq 5$ for the Poisson model.

if scaled deviance
$> (1-\alpha)$th of χ^2_{n-p}
fitted
model is not adequate
loss a lot of information

The normal, gamma or inverse gaussian models:

- we have to estimate ϕ using $X^2/(n - p)$.
- we can't use the scaled deviance to test model adequacy.

Scaled deviance for model selection (nested models)

The difference between two scaled deviances can be used for model selection between two nested models (i.e., testing the significance of the extra parameters).

The binomial and Poisson models:

- For nested models, if the smaller model is correct then the difference between two scaled deviances is the log likelihood ratio and will be $\approx \chi^2_s$ where s is the difference in the number of parameters.
- Example : See "LR test using deviance" in Bliss.pdf

if the difference is bigger than what we expect, then the additional parameter in the full model is useful, pick the full model

The normal, gamma or inverse gaussian models:

- we have to estimate ϕ using $X^2/(n-p)$.
- see the next slide..

Scaled deviance for model selection (nested models)

For the normal model (linear model), if model A is nested within model B, under the null hypothesis that model A is correct we have

$$\frac{(D^A - D^B)/s}{\underbrace{X^2/(n-p)}_{\text{estimate } \phi.}} \sim F_{s,n-p}$$

where D^A and D^B denote deviances for models A and B, respectively, we have n observations, model A has $p - s$ parameters, and model B has p parameters.

- X^2 (Pearson's chi-squared) is calculated using model B. ϕ
- It is equivalent to the F-test in linear model.

For other models, this distributional result only holds approximately, but it can still be used for comparing two nested models. You will use it for comparing gamma models in the lab problem for the week 4.

(p-s parameters)
model A is nested in
model B.
(p)

LR: $\frac{D^A - D^B}{\phi}$

estimate ϕ with $\frac{X^2}{n-p}$

$\frac{D^A - D^B/s}{\frac{X^2}{n-p}} \sim F_{s,n-p}$

$\Rightarrow \frac{\frac{D^A - D^B}{\phi}/s}{\frac{X^2}{\phi}/n-p}$

ϕ is known

$\frac{D^A - D^B}{\phi} \sim \chi^2_s$ $\quad \frac{X^2}{\phi} \sim \chi_{n-p}$.

Scaled deviance for model selection (AIC)

The AIC is defined as

$$\text{AIC} = 2p - 2\log\mathcal{L}(\hat{\beta})$$

where $\hat{\beta}$ represent MLE of parameters in the model and p is the number of the parameters. Given a choice, we prefer the model with the smaller AIC.

If model B has s more parameters than model A (not necessarily nested within B), then

$$\begin{aligned}\text{AIC}^B - \text{AIC}^A &= 2s - 2\log\mathcal{L}(\hat{\beta}^B) + 2\log\mathcal{L}(\hat{\beta}^A)\\ &= 2s + \frac{D^B}{\phi} - \frac{D^A}{\phi}.\end{aligned}$$

Like the log likelihood ratio, the AIC needs an estimate of ϕ.

Diagnostics

Residuals

$r_i = y_i - \hat{\mu_i}$

$var(r_i) = var(y_i) = v(\mu_i)\, a(\phi)$.

for linear model $var(r_i)$ is constant across samples

$v(\mu_i) = 1$

Response residuals:

$$y_i - \hat{\mu}_i$$

Unless $v(\mu)$ is constant, as in the normal case, the response residuals are not homoskedastic and hence not very useful.

same variance across samples

```
> residuals(glmfit, type="response")
```

Pearson residuals:

$$r_P(i) = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}$$

$var\left(\frac{y_i - \mu_i}{\sqrt{v(\mu_i)}}\right) = a(\phi)$ → should be homoskedastic

since μ_i unknown, use estimate μ_i

Pearson residuals are (approximately) homoskedastic, and $\boxed{\sum_i r_P(i)^2 = X^2.}$

defind to estimate ϕ

```
> residuals(glmfit, type="pearson")
```

Residuals

Deviance residuals:

$$r_D(i) = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

where the deviance is $D = \sum_i d_i = \sum_i r_D(i)^2$.

```
> residuals(glmfit)
```

- As for linear models, patterns in the residuals indicate structure in the data that has not been captured by the model.
- We can plot the residuals against predictor variables, the responses, or the fitted means. Often a plot against $\hat{\eta}_i = x_i^T\hat{\beta}$ works well.
- With count data residual plots exhibit banding due to the discrete nature of the responses, and this can make it hard to see other patterns. In this case we can use a smoothed fit of the residuals to help spot trends/patterns.

banding pattern dominate ?

Leverage

"potential influence."

linear model

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^TX^{-1})X^T}_{} y$$

$$= \underset{n\times n}{H} y$$

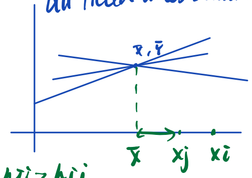
The leverage measures the potential influence of a point on the fitted model. We borrow the definition of leverage from the theory of linear models, and use the hat matrix from the IWLS algorithm.

In the IWLS algorithm

$$\hat{Z} = X\hat{\beta} = X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}Z,$$

h_{ii}: distance between $\bar{x}$ and x_i

all fitted lines should go through $(\bar{x}, \bar{y})$



where $\hat{\Sigma}_{ii} = (g'(\hat{\mu}_i))^2 v(\hat{\mu}_i)\phi$ (assuming $a(\phi) = \phi$). Then, the hat matrix for $\hat{Z}$ is

$$H = X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}.$$

The leverage for the i-th observation is H_{ii}, the i-th diagonal element of H.

if $h_{ii} > h_{jj}$

the point (x_i) far away from $\bar{x}$ has a potentially larger influence than the point close to $\bar{x}$ (x_j)

Note that H does not depend on ϕ (it cancels out).

A large leverage does not necessarily mean a point *has* influenced the fit.

```
> influence(glmfit)$hat
```

measures the potential influence → doesn't measure the direct influence

Studentised residuals

$\mathrm{var}(z) = H\,\mathrm{var}(z)\,H^T$

The variance of $\hat{Z}$ is

$$
\begin{aligned}
& H \mathrm{Var}\, Z H^T \\
&= [X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}]\Sigma[\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T] \\
&= X(X^T\Sigma^{-1}X)^{-1}X^T \quad = \quad H\Sigma.
\end{aligned}
$$

Whence $\mathrm{Var}\,(Z - \hat{Z}) = (I - H)\Sigma$.

$H\Sigma$ is symmetric

$(H\Sigma)^T = H\Sigma$

also Σ is diagonal

$(H\Sigma)^T = \Sigma^T H^T = \Sigma H^T$

$$\mathrm{var}(z-\hat{z}) = \mathrm{var}(z - Hz) = \mathrm{var}\big((I-H)z\big) = (I-H)\,\mathrm{var}(z)\,(I-H)^T$$

$$= (I-H)\Sigma(I-H)^T$$

$$= (\Sigma - H\Sigma)(I - H^T)$$

$$= \Sigma - \cancel{H\Sigma} - \Sigma H^T + \underbrace{H\Sigma H^T}_{\cancel{H\Sigma}}$$

$$= \Sigma - H\Sigma$$

$$= (I-H)\Sigma$$

Studentised residuals

ith diagonal element of I−H is 1−Hii

ith diagonal element of Σ is g'(μi)²v(μi) φ.

?.

From the above we have $\mathrm{Var}\,\frac{Z_i-\hat{Z}_i}{\sqrt{(1-H_{ii})g'(\mu_i)^2v(\mu_i)\phi}}=1$. Also

$$\begin{aligned} Z_i-\hat{Z}_i &= g(\mu_i)+(Y_i-\mu_i)g'(\mu_i)\underline{-[g(\mu_i)+(\hat{Y}_i-\mu_i)g'(\mu_i)]} \\ &= (Y_i-\hat{Y}_i)\underline{g'(\mu_i)} \end{aligned}$$

will be cancelled out.

so, noting that $\underline{\hat{Y}_i=\hat{\mu}_i}$,

$$\mathrm{Var}\,\frac{Y_i-\hat{\mu}_i}{\sqrt{(1-H_{ii})v(\hat{\mu}_i)\hat{\phi}}}\approx 1$$

→ Pearson residual

The LHS is just $r_P(i)/\sqrt{(1-H_{ii})\hat{\phi}} =: r_{SP}(i)$, which we call the i-th **studentised Pearson residual**.

Studentised residuals

Studentised Pearson residual:

$$r_{SP}(i) = \frac{r_P(i)}{\sqrt{(1 - H_{ii})\hat{\phi}}}$$

By analogy we define **Studentised deviance residual:**

$$r_{SD}(i) = \frac{r_D(i)}{\sqrt{(1 - H_{ii})\hat{\phi}}}$$

Jack-knife residuals

" direct influence "

A direct measure of the influence of a point is the **jack-knife residual**, which is the change in $\hat{\mu}_i$ when you remove y_i from the set of observations, then scaled to standardise the variance.

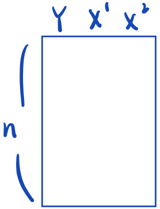
The jack-knife residual can be approximated by

$$\text{sign}(y_i - \hat{\mu}_i)\sqrt{(1 - H_{ii})r_{SD}^2(i) + H_{ii}r_{SP}^2(i)}.$$

$\Downarrow$

weighted sum

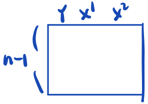
```
> rstudent(glmfit)
```



STEP 1 observation id

STEP 2. 1. estimate $\hat{\mu}_1$

eg. ⇒ remove 1st



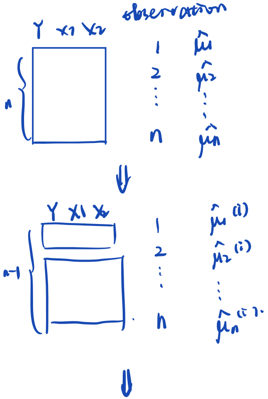
⇒ then estimate $\hat{\mu}_i^{(1)}$ → the estimation after removing 1st observation

⇒ ① calculate $\hat{\mu}_i - \hat{\mu}_i^{(1)}$

STEP 3. ② $\hat{\mu}_2 - \hat{\mu}_2^{(2)}$

} ⇒ generally $\dfrac{\hat{\mu}_i - \hat{\mu}_i^{(1)}}{\bigcirc}$ → standardised

influence of ith observation on "ith" fitted mean



Cook's distance

" influence of ith observation on ALL fitted mean "

Another measure of the influence of the i-th observation is **Cook's distance**:

$$\frac{(\hat{\beta}^{(i)} - \hat{\beta})^T X^T \hat{\Sigma}^{-1} X(\hat{\beta}^{(i)} - \hat{\beta})}{p}$$

where $\hat{\beta}^{(i)}$ is the estimate of β obtained when y_i is omitted.

```
> cooks.distance(glmfit)
```

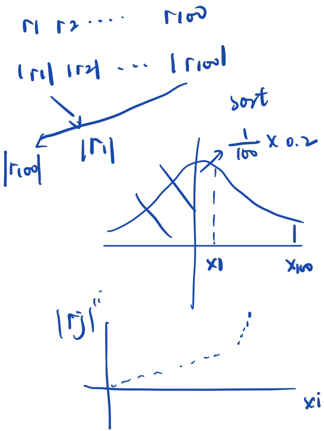
look at

$$\frac{\sum_{j=1}^{n}(\hat{\mu}_j - \hat{\mu}_j^{(i)})^2}{\bigcirc}$$

→ standardize.

the difference between $\hat{\mu}_j$ and $\hat{\mu}_j^{(i)}$ (after removing ith instance).

Plotting leverage/jack-knife residuals/cook's distance



- When plotting it is helpful to consider them ordered by absolute size.
- If we plot the ordered absolute values against the percentage points of a half-normal distribution, then it is easier to see if the largest values are in keeping with the others. That is, plot the i-th ordered absolute residual against $\Phi^{-1}((n+i)/(2n+1))$, for $i = 1, \ldots, n$.
- If all is well we expect to see a smooth plot, while a jump or kink in the tail indicates a potential problem.
- Note that leverage/jack-knife residuals/cook's distance will not be normal, so don't expect a straight line.

Checking linearity

- A non-linear link g (anything except the identity) makes it harder to check the assumption that $g(\mu_i) = \mathrm{x}_i^T \boldsymbol{\beta}$.

$g(\hat{\mu}_i) \to g(y_i)$.

- The easiest thing to do is to plot $g(y_i)$ against $\{x_{ij}\}_{i=1}^n$ for each j and look for linear relationships.

iterated weighted least square estimator.

- A more sophisticated approach is to plot $z_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$ against $\{x_{ij}\}_{i=1}^n$, where we use $\hat{\mu}_i$ for μ_i. From the IWLS algorithm we know that these plots should be linear.

- If there are non-linearities present then we can consider transforming $\{x_{ij}\}_{i=1}^n$. Transforming the responses y_i is often not a good idea for a glm, as this can break assumptions made about the distribution of Y_i.

Example

See Gala.pdf.

Learning goals

- Be able to define Generalised Linear Model (GLM).
- Be able to obtain the cannonical link function for GLM.
- (Challenging) Be able to explain ideas for the iterated weighted least squares (IWLS) algorithm.
- Be able to compute the variance of parameter estimates in GLM using IWLS algorithm.
- Understand (scaled) deviance
    - Be able to define (scaled) deviance.
    - Be able to compute (scaled) deviance.
    - Be able to use it to test model adequacy, perform model selection for nested models and non-nested models.
- Be able to perform diagnostics for GLM using R.
    - Understand what different 'diagnostics measurements' aim to measure.
    - Be able to obtain those diagnostic measurements from the *glm* output.
    - Be able to perform diagnostics for GLM using R.