Discussion # 12

Zifeng Zhao

Dec 4th, 2015
Maximum Likelihood function for GLM’s

Solve score equations, for \( j = 1, \ldots, p \), \( S_j(\beta) = \frac{\partial l}{\partial \beta_j} = 0 \), where \( l \) is the log-likelihood function and

\[
l = \sum_{i=1}^{m} \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} = \sum_i l_i.
\]

Let \( S(\beta) = \frac{\partial l}{\partial \beta} \), MLE is the solution of \( S(\beta) = 0 \).

To solve \( S(\beta) = 0 \), iterative method is required for most GLMs. The Newton-Raphson algorithm uses the observed derivative of the score (gradient) and Fisher scoring method uses the expected derivative of the score (i.e., Fisher’s information matrix, \(-I_m\)).
Maximum Likelihood function for GLM's

The algorithm for New-Raphson method:

- 1. Find an initial value $\hat{\beta}^{(0)}$.

- 2. For step $j$ to step $j + 1$, update $\hat{\beta}^{(j)}$ via
  
  $$\hat{\beta}^{(j+1)} = \hat{\beta}^{(j)} - V(\hat{\beta}^{(j)})^{-1} S(\hat{\beta}^{(j)})$$.

- 3. Evaluate convergence using changes in log L or $\|\hat{\beta}^{(j+1)} - \hat{\beta}^{(j)}\|$.

- 4. Iterate until convergence criterion is satisfied.

where $V(\beta) = \frac{\partial S(\beta)}{\partial \beta}$.
The algorithm for Fisher-Scoring:

1. Find an initial value $\hat{\beta}^{(0)}$.

2. For step $j$ to step $j + 1$, update $\hat{\beta}^{(j)}$ via
   $$\hat{\beta}^{(j+1)} = \hat{\beta}^{(j)} + I(\hat{\beta}^{(j)})^{-1} S(\hat{\beta}^{(j)}).$$

3. Evaluate convergence using changes in log L or $\|\hat{\beta}^{(j+1)} - \hat{\beta}^{(j)}\|$.

4. Iterate until convergence criterion is satisfied.

where $I(\beta) = -E\left(\frac{\partial S(\beta)}{\partial \beta}\right)$.

Remark: For canonical link, these two algorithms are the same.
**Iterative Weighted Least Squares (IWLS)**

Consider the GLM model $g(\mu) = \eta$, let $\hat{\eta}_0$ be the current estimate of the linear predictor, with corresponding fitted value $\hat{\mu}_0$ derived from the link function $g(\mu) = \eta$.

Step 1: Given the initial estimator $\hat{\eta}_0$, define the adjusted dependent variate with

$$z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\left(\frac{d\eta}{d\mu}\right)_0$$

where the derivative of the link is evaluated at $\hat{\mu}_0$.

Note that $z$ is just a linearized form of the link function applied to the data

$$g(y) \approx g(\mu) + (y - \mu)g'(\mu)$$
Iterative Weighted Least Squares (IWLS)

Step 2: Given \( \hat{\eta}_0 \), \( z_0 \) is an approximation of \( g(\mu) + (y - \mu)g'(\mu) \), which has mean \( g(\mu) = x'\beta \) and variance \( \text{Var}(y)g'(\mu)^2 = a(\phi)V_0(\mu)(\frac{dn}{d\mu})_0^2 \).

So \( z_0 \approx (x'\beta, a(\phi)V_0(\frac{dn}{d\mu})_0) \), then we can use weighted least square to solve \( \hat{\beta}_1 \). The weight matrix \( W \) is a diagonal matrix with element \( W_0^{-1} = V_0(\frac{dn}{d\mu})_0^2 \).

Step 3: After we get \( \hat{\beta}_1 \), form a new estimation of \( \eta \), go back to Step 1. Repeat until convergence.
IWLS is equivalent to Fisher-Scoring method. The log likelihood for a SINGLE observation, in canonical form, is given by

\[ l = \left\{ y\theta - b(\theta) \right\}/a(\phi) + c(y, \phi). \]

Then

\[ \frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta} \frac{\partial \theta}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \beta_j}, \]

where we know \( b'(\theta) = \mu \) and \( b''(\theta) = V \), so we have \( \frac{\partial \mu}{\partial \theta} = V \), also \( \frac{\partial \eta}{\partial \beta_j} = x_j \).
Justification of IWLS (Nelder P.40-41)

Together, we have

\[ \frac{\partial l}{\partial \beta_j} = \frac{(y - \mu)}{a(\phi)} \frac{1}{V} \frac{d\mu}{d\eta} x_j = \frac{W}{a(\phi)} (y - \mu) \frac{d\eta}{d\mu} x_j \]

where \( W^{-1} = V (\frac{d\eta}{d\mu})^2 \).

The maximum-likelihood equations for \( \beta_j \) are given by

\[ \sum W (y - \mu) \frac{d\eta}{d\mu} x_j = 0 \]

where \( \sum \) without a suffix denotes summation over the units,
Fisher’s scoring method uses the gradient vector $\frac{\partial l}{\partial \beta} = u$ and minus the expected value of the Hessian matrix,

$$-E(\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}) = A.$$  

Given the current estimate $b$ of $\beta$, derive an adjustment $\delta b$ defined as the solution of

$$A\delta b = u.$$  

The components of $u$ (omitting the dispersion factor) are

$$u_r = \frac{\partial l}{\partial \beta_r} = \sum W(y - \mu) \frac{d\eta}{d\mu} x_r.$$
So we have

\[ A_{rs} = -E\left( \frac{\partial u_r}{\partial \beta_s} \right) = -E\left( \sum [(y-\mu) \frac{\partial}{\partial \beta_s} \{ W \frac{d\eta}{d\mu} x_r \} + W \frac{d\eta}{d\mu} x_r \frac{\partial}{\partial \beta_s} (y-\mu)] \right) \]

\[ = \sum W \frac{d\eta}{d\mu} x_r \frac{\partial \mu}{\partial \beta_s} = \sum Wx_r x_s. \]

If we denote \( W \) be the \( n \times n \) diagonal weight matrix and \( X = (x_1|x_2 \ldots |x_n)^T \) be the model matrix, then the Fisher information matrix \( A \) can be expressed as,

\[ A = X'WX \]
The updated estimate of $\beta$ using Fisher Scoring will be $b^* = b + \delta b$ and it satisfies,

$$Ab^* = Ab + A\delta b = Ab + u.$$ 

where $(Ab)_r = \sum_s A_{rs} b_s = \sum Wx_r \eta$, so together we have

$$(Ab^*)_r = \sum Wx_r \{\eta + (y - \mu) \frac{d\eta}{d\mu}\}$$

Let $z = \eta + (y - \mu) \frac{d\eta}{d\mu}$, we have

$$(Ab^*)_r = \sum Wx_r z$$
Together, we have

\[ Ab^* = X'WZ, \]

where \( Z = (z_1, |, \ldots, |z_n)^T \), i.e.

\[ X'WXb^* = X'WZ, \]

which is exactly the IWLS.
Besides deviance, the other important measure of discrepancy/goodness of fit is the Generalized Pearson $X^2$ statistic,

$$X^2 = \sum (y - \hat{\mu})^2 / V(\hat{\mu}),$$

where $V(\hat{\mu})$ is the estimated variance function for the distribution concerned.

For the Normal distribution, $X^2$ is the residual sum of squares.
In Binary, \( E(Y_i) = p_i, \ Var(Y_i) = p_i(1 - p_i)/m \);
In Poisson, \( E(Y_i) = \mu_i, \ Var(Y_i) = \mu_i \)

When the variance assumption of the GLM is broken, particularly, actual variance of \( Y_i \) is bigger than that implied by the GLM, it’s called over-dispersion.

Intuition: Then we are just fitting a MIS-SPECIFIED model.

Generally speaking, when the variance assumption of the GLM is broken but the link function and choice of predictors is correct, the estimates of \( \beta \) are consistent, but the standard errors will be wrong.
Poisson Regression in GLM (Faraway 3.1)

\[ Y_i \sim \text{Poisson}(\mu_i), \text{ where } g(\mu_i) = x_i' \beta, \text{ the canonical link function is } g(\cdot) = \log(\cdot). \]

Now we have a data set consisting of 30 Islands, and response is the number of species of turtles found on each island.

```r
> library(faraway)
> data(gala)
> str(gala)
'data.frame': 30 obs. of 7 variables:
$ Species : num 58 31 3 25 2 18 24 10 8 2 ...
$ Endemics : num 23 21 3 9 1 11 0 7 4 2 ...
$ Area : num 25.09 1.24 0.21 0.1 0.05 ...
$ Elevation: num 346 109 114 46 77 119 93 168 71 112 ...
$ Nearest : num 0.6 0.6 2.8 1.9 1.9 8 6 34.1 0.4 2.6 ...
$ Scruz : num 0.6 26.3 58.7 47.4 1.9 ...
$ Adjacent : num 1.84 572.33 0.78 0.18 903.82 ...

> gala=gala[,,-2]
> modp <- glm(Species ~ .,family=poisson, gala)
```
Poisson Regression in GLM (Faraway 3.1)

> summary(modp)

Call:
glm(formula = Species ~ ., family = poisson, data = gala)

Deviance Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-8.2752</td>
<td>-4.4966</td>
<td>-0.9443</td>
<td>1.9168</td>
<td>10.1849</td>
</tr>
</tbody>
</table>

Coefficients:

| (Intercept) | Estimate | Std. Error | z value | Pr(>|z|) |
|-------------|----------|------------|---------|---------|
| Area        | -5.799e-04 | 2.627e-05  | -22.074 | < 2e-16 *** |
| Elevation   | 3.541e-03  | 8.741e-05  | 40.507  | < 2e-16 *** |
| Nearest     | 8.826e-03  | 1.821e-03  | 4.846   | 1.26e-06 *** |
| Scruz       | -5.709e-03 | 6.256e-04  | -9.126  | < 2e-16 *** |
| Adjacent    | -6.630e-04 | 2.933e-05  | -22.608 | < 2e-16 *** |

Signif. codes:  0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
OVER-DISPERSION

For a Poisson distribution, the mean is equal to the variance. We plot this estimated variance against the mean,

```r
> mu=as.numeric(modp$fitted)
> var=(gala$Species-mu)^2;plot(log(mu),log(var))
> abline(0,1)
```
Over-dispersion

We see that the variance is proportional to, but larger than, the mean. It may have an over-dispersion.

Assume now we have a dispersion parameter $\phi$, then $\text{Var}(y_i) = \phi \mu_i$. It can be estimated by

$$\hat{\phi} = \frac{X^2}{n - p} = \frac{\sum(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

```r
> dp <- sum(residuals(modp, type="pearson")^2)/modp$df.res
> dp
[1] 31.74914
```
OVER-DISPERSION

> summary(modp, dispersion=dp)

Call:
glm(formula = Species ~ ., family = poisson, data = gala)

Deviance Residuals:

  Min 1Q Median 3Q Max
-8.2752 -4.4966 -0.9443 1.9168 10.1849

Coefficients:

                  Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.1548079  0.2915897 10.819  < 2e-16 ***
Area          -0.0005799  0.0001480  -3.918  8.95e-05 ***
Elevation     0.0035406  0.0004925   7.189  6.53e-13 ***
Nearest       0.0088256  0.0102621   0.860    0.390
Scruz         -0.0057094  0.0035251  -1.620    0.105
Adjacent     -0.0006630  0.0001653  -4.012  6.01e-05 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

(Dispersion parameter for poisson family taken to be 31.74914)