# STAT347: Generalized Linear Models
# Lecture 4

Today's topics: Agresi Chapters 4.4.6, 4.5, 4.7

- Model diagnosis with residuals
- Computation of the ML estimate
- Example: building a GLM

## 1 Model checking with the residuals

As in the linear models, we can examine the residuals to help us check whether a model fits poor or not, and whether there are any outliers in the observations.

Three types of residuals:

- Pearson residual:
$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}$$
where $v(\hat{\mu}_i) = \widehat{\mathrm{Var}}(y_i)$. For instance, if $y_i \sim \text{Poisson}(\mu_i)$ then $v(\hat{\mu}_i) = \hat{\mu}_i$. As we have shown in Lecture 2, in general $v(\hat{\mu}_i) = b''(\hat{\theta}_i)a(\hat{\phi})$.

- Deviance residual:
$$d_i = \sqrt{D(y_i, \hat{\mu}_i)} \times \text{sign}(y_i - \hat{\mu}_i)$$
For instance, for the Gaussian linear model, $D(y_i, \hat{\mu}_i) = (y_i - \hat{\mu}_i)^2/\sigma^2$, and the deviance residual is the same as the Pearson residual. As a rule of thumb, an observation is fitted poorly by the GLM model if $|d_i| > 2$.

- As in the linear models, the mean of $e_i$ is typically smaller than 1 as $\hat{\mu}_i$ is estimated. After some calculations (see Chapter 4.4.5), one can compute a more accurate variance of $y_i - \hat{\mu}_i$.

  Standardized residual:
$$r_i = \frac{e_i}{\sqrt{1 - \hat{h}_{ii}}}$$
where $h_{ii}$ is the $i$th diagonal element of the $H_W$ defined equation (4.19) of the Agresti chapter 4.4.5.

## 2 Computation

Score equation:
$$\dot{L}(\beta) = X^T D V^{-1}(y - \mu) = 0$$

where

$$L(\beta) = \sum_i [y_i \theta_i - b(\theta_i)]$$

(This is the log-likelihood ignoring the quantities involving $\phi$ that does not affect the estimation of $\beta$)

## 2.1 Newton's method

Second-order approximation of $L(\beta)$

$$L(\beta) \approx L(\beta^{(t)}) + \dot{L}(\beta^{(t)})^T(\beta - \beta^{(t)}) + \frac{1}{2}(\beta - \beta^{(t)})^T \ddot{L}(\beta^{(t)})(\beta - \beta^{(t)})$$

at $t$th iteration. If $\ddot{L}(\beta^{(t)}) \preceq 0$, then maximizing the second-order approximation is equivalent to solving

$$\dot{L}(\beta) \approx \dot{L}(\beta^{(t)}) + \ddot{L}(\beta^{(t)})(\beta - \beta^{(t)}) = 0$$

We have

$$\beta^{(t+1)} = \beta^{(t)} - \ddot{L}(\beta^{(t)})^{-1} \dot{L}(\beta^{(t)})$$

- Newton's method is a general algorithm for optimizing twice-differentiable functions.
- Converge to the global maximum if $L(\beta)$ is strongly concave
  - If $g(\cdot)$ is the canonical link, then $L(\beta)$ is concave in $\beta$

    $$-\ddot{L}(\beta^{(t)}) = X^T W^{(t)} X = X^T V^{(t)} X = -\mathbb{E}\left(\ddot{L}(\beta^{(t)})\right) \succeq 0$$

  - If $g(\cdot)$ is a general link, then $L(\beta)$ is NOT guaranteed to be concave in $\beta$
  - If $-\ddot{L}(\beta^{(t)})$ is not non-negative, than step $i$ does not maximize the quadratic approximation and Newton's method may not converge.
  - We can use another quadratic approximation that works better in practice: Fisher scoring method

## 2.2 Fisher scoring method

In lecture 2, we showed that $-\mathbb{E}\left(\ddot{L}(\beta)\right) \succeq 0$ for any $\beta$.

Instead of using the Hessian $\ddot{L}(\beta^{(t)})$, use its expectation

$$J^{(t)} = \mathbb{E}\left(\ddot{L}(\beta^{(t)})\right) = -X^T W^{(t)} X$$

instead of $\ddot{L}(\beta^{(t)})$ itself in the second-order approximation. Each iteration becomes:

$$\beta^{(t+1)} = \beta^{(t)} - \left(J^{(t)}\right)^{-1} \dot{L}(\beta^{(t)})$$

### 2.3 Iteratively reweighted least squares (IRLS)

We can make a connection between the optimization for GLM and weighted least squares estimation.

Recall the score equation:

$$\dot{L}(\beta) = X^T D V^{-1}(y-\mu) = 0$$

where $V = \mathrm{diag}(\mathrm{Var}(y_1), \cdots, \mathrm{Var}(y_n))$ and $D = \mathrm{diag}\left(g'(\mu_1), \cdots, g'(\mu_n)\right)^{-1}$, $y = (y_1, \cdots, y_n)$ and $\mu = (\mu_1, \cdots, \mu_n)$.

Also in lecture 2, we used the notation $\eta_i = X_i^T\beta = g(\mu_i)$. Thus, $D = \mathrm{diag}\left(\frac{\partial\mu_1}{\partial\eta_1}, \cdots, \frac{\partial\mu_n}{\partial\eta_n}\right)$. We also defined the diagnoal matrix $W = D^2V^{-1}$. Thus,

$$\dot{L}(\beta) = X^T D V^{-1}(y-\mu) = X^T W D^{-1}(y-\mu)$$

We can make a first order approximation of $\mu$

$$\mu = \mu^{(t)} + D^{(t)}(\eta - \eta^{(t)})$$

then

$$\dot{L}(\beta) \approx X^T W^{(t)}(z^{(t)} - X\beta)$$

where

$$z^{(t)} = X\beta^{(t)} + \left(D^{(t)}\right)^{-1}(y - \mu^{(t)})$$

is a linear approximation of $\eta$ at the $t$th iteration.

Thus, at the $t+1$th iteration, we solve

$$X^T W^{(t)}(z^{(t)} - X\beta) = 0$$

which can be considered as a weighted linear regression with observations $z_i^{(t)}$ and weight $w_i$ for each sample $i$.

- IRLS is equivalent to Fisher scoring, see Section 4.5.4
- weight matrix $W^{(t)} \approx \mathrm{Var}\left(z^{(t)}\right)^{-1}$

## 3 Data examples

Please check the R notebook 2.

Next time: Chapter 5.1 - 5.2, binary data model, application scenarios