STATS305B – Lecture 7

Jonathon Taylor
Scribed by Michael Howes

01/26/22

Contents

1 Binary GLMs

1.1 Fitting

Let F be a CDF with density f. The deviance for a binary GLM with link function $g = F^{-1}$ is,

$$\text{DEV}(\beta|Y) = 2\sum_{i=1}^{n} -Y_i \frac{F(X_i^T\beta)}{F(X_i^T\beta)(1-F(X_i^T\beta))} - \log\left(1 - F(X_i^T\beta)\right).$$

Thus,

$$\nabla \text{DEV}(\beta|Y) = 2\sum_{i=1}^{n} X_i \frac{f(X_i^T\beta)^2}{F(X_i^T\beta)(1-F(X_i^T\beta))}\left[\frac{Y_i - F(X_i^T\beta)}{f(X_i^T\beta)}\right].$$

Since $X_i^T\beta = F^{-1}(\mathbb{E}_\beta[Y_i])$, we have that $\mathbb{E}_\beta[F(X_i^T\beta) - Y_i] = 0$. Thus, we have

$$\mathbb{E}[\nabla^2 \text{DEV}(\beta|Y)] = 2\sum_{i=1}^{n} X_i X_i^T \frac{f(X_i^T\beta)^2}{F(X_i^T\beta)(F(X_i^T\beta)-1)} = 2X^T W_\beta X,$$

where $W_\beta = \text{diag}\left(\frac{f(X_i)^2}{F(X_i^T\beta)(1-F(X_i^T\beta))}\right)$. We can also rewrite $\nabla \text{DEV}(\beta|Y)$ in terms of the W_β,

$$\nabla \text{DEV}(\beta|Y) = 2X^T W_\beta \left(\frac{Y - F(X\beta)}{f(X\beta)}\right).$$

To fit a binary glm, we can use Fisher scoring. Fisher scoring is an iterative quasi-Newton method given by

$$\begin{aligned}\widehat{\beta}^{(k+1)} &= \widehat{\beta}^{(k)} - \mathbb{E}_{\beta^{(k)}}[\nabla^2 \text{DEV}(\beta^{(k)}|Y)]^{-1}\nabla \text{DEV}(\widehat{\beta}^{(k)}|Y) \\ &= (X^T W_{\widehat{\beta}^{(k)}} X)^{-1} X^T W_{\widehat{\beta}^{(k)}} \left(X\widehat{\beta}^{(k)} + \frac{Y - F(X\widehat{\beta}^{(k)})}{f(X\widehat{\beta}^{(k)})}\right).\end{aligned}$$

Note that the above method is a form of iterative re-weighted least squares. This was important for historical reasons since least squares was one of the main algorithms implemented on early computers. When we view Fisher scoring as iterative re-weighted least squares, the response at step $k+1$ is,

$$Z^{(k+1)} = X\widehat{\beta}^{(k)} + \frac{Y - F(X\widehat{\beta}^{(k)})}{f(X\widehat{\beta}^{(k)})} = g(\mathbb{E}_{\beta^{(k)}}(Y)) + g'(\mathbb{E}_{\widehat{\beta}^{(k)}}(Y))(Y - \mathbb{E}_{\widehat{\beta}^{(k)}}[Y]).$$

This is because if $\mu = F(X^T\beta)$, then $g(\mu) = F^{-1}(\mu) = X^T\beta$ and $g'(\mu) = \frac{1}{F'(F^{-1}(\mu))} = \frac{1}{f(X\beta)}$. The weight matrix at time $k+1$ is,

$$\begin{aligned} W^{(k+1)} &= \text{diag}\left(\frac{f(X_i^T\widehat{\beta}^{(k)})^2}{F(X_i^T\widehat{\beta}^{(k)})(1 - F(X_i^T\widehat{\beta}^{(k)}))}\right) \\ &= \text{diag}\left(\frac{F(X_i^T\widehat{\beta}^{(k)})(1 - F(X_i^T\widehat{\beta}^{(k)}))}{f(X_i^T\widehat{\beta}^{(k)})^2}\right)^{-1} \\ &= \frac{1}{f(X\widehat{\beta}^{(k)})^2}\,\text{Var}_{\widehat{\beta}^{(k)}}(Y)^{-1} \\ &= g'(\widehat{\mu}^{(k)})^2 V(\widehat{\mu}^{(k)})^{-1}, \end{aligned}$$

where $\widehat{\mu}^{(k)} = \mathbb{E}_{\widehat{\beta}}[Y]$. This shows how Fisher scoring can be generalized to other glms. Instead of minimizing the deviance, we simply run iterative re-weighted least squares with features X, iterative response

$$Z^{(k)} = g(\widehat{\mu}^{(k)}) + g'(\widehat{\mu}^{(k)})(Y - \widehat{\mu}^{(k)})$$

and iterative weights $W^{(k)} = g'(\widehat{\mu}^{(k)})^2 V(\widehat{\mu}^{(k)})^{-1}$. This is important because in general for glms we do not have a full model from which we can calculate and optimize a likelihood. We just have the functions g and V. Note that, in the binary case, if the model is true, then

$$\widehat{\beta} \approx \mathsf{N}(\beta^*, (X^T W_{\widehat{\beta}} X)^{-1}).$$

If the model is not true, then we have the sandwich form

$$\widehat{\beta} - \beta^* \approx \mathsf{N}(0, Q_{\beta^*}^{-1}\Sigma_{\beta^*}Q_{\beta^*}^{-1}),$$

where $Q_{\beta^*} = \mathbb{E}_{\beta^*}[X^T W_{\beta^*} X]$ and $\Sigma_{\beta^*} = \text{Var}(X^T(Y - \pi_{\beta^*}(X)))$. In practice, we can estimate Q_{β^*} with $X^T W_{\widehat{\beta}} X$ and bootstrap for Σ_{β^*} or we can bootstrap directly for $\text{Var}(\widehat{\beta})$.

1.2 Inference

The difference of deviance can be used as a likelihood ratio test. If $M_R \subseteq M_F$ are two models, then if M_R contains the true model

$$\text{DEV}(M_R) - \text{DEV}(M_F) \overset{n\to\infty}{\sim} \chi^2_{df_R - df_F}.$$

Unlike in linear regression, this test is different to the Wald test for a single predictor (i.e. when M_F is M_R plus one additional predictor).

2 Poisson data

Suppose $Y \sim \mathsf{Poisson}(\lambda)$, then for $y = 0, 1, 2, \ldots$, we have

$$\mathbb{P}_\lambda(Y = y) = \frac{\lambda^y}{y!}\exp(-\lambda) = \exp(y\log(\lambda) - \lambda)\frac{1}{y!}.$$

The canonical link for this family is $g = \log$, giving the model $\log(\lambda_i) = X_i^T\beta$ where $\lambda_i = \mathbb{E}[Y_i|X_i]$. This model is called the *log-linear model*. The variance function for this model is $\text{Var}(Y_i|X_i) = \lambda_i$. The deviance is

$$\text{DEV}(\beta|Y) = 2\sum_{i=1}^{n} -Y_i X_i^T\beta + e^{X_i^T\beta} + Y_i\log(Y_i) - Y_i,$$

the terms $Y_i\log(Y_i) - Y_i$ come from the saturated model where we have $\lambda_i = Y_i$. The gradient and Hessian of the deviance are thus,

$$\nabla\,\text{DEV}(\beta|Y) = 2X^T(\exp(X\beta) - Y) = 2X^T(\mathbb{E}_\beta[Y] - Y),$$

and

$$\nabla^2\,\text{DEV}(\beta|Y) = 2X^T W_\beta X,$$

where $W_\beta = \text{Var}_\beta(Y) = \exp(X\beta)$. Thus, we can fit this model by using Newton–Raphson. This gives us the iterative rule,

$$\begin{aligned}
\widehat{\beta}^{(k+1)} &= \widehat{\beta}^{(k)} - \nabla^2\,\text{DEV}(\widehat{\beta}^{(k)}|Y)^{-1}\left(\nabla\,\text{DEV}(\widehat{\beta^{(k)}}|Y)\right) \\
&= \widehat{\beta}^{(k)} - \left(X^T W_\beta X\right)^{-1} X^T(\mathbb{E}_{\widehat{\beta}^{(k)}}[Y] - Y) \\
&= \widehat{\beta}^{(k)} + \left(X^T\exp(X\beta)X\right)^{-1} X^T(Y - \exp(X\beta)).
\end{aligned}$$

This iterative algorithm once again corresponds to a form of iterative re-weighted least squares, with response,

$$\begin{aligned}
Z^{(k+1)} &= X\widehat{\beta}^{(k)} + (Y - \exp(X\widehat{\beta}^{(k)}))/\exp(X\widehat{\beta}^{(k)}) \\
&= g(\widehat{\lambda}^{(k)}) + g'(\widehat{\lambda}^{(k)})(Y - \widehat{\lambda}^{(k)})
\end{aligned}$$

where $\widehat{\lambda}^k = \exp(X\widehat{\beta}^{(k)}) = \mathbb{E}_{\widehat{\beta}^{(k)}}[Y]$. and weight matrix,

$$\begin{aligned}
W^{(k+1)} &= \exp(X\widehat{\beta}^{(k)}) \\
&= \text{Var}_{\widehat{\beta}^{(k)}}(Y).
\end{aligned}$$

2.1 Residuals

Like the binary models, there are two types of residuals for the log-linear model. We have the Pearson residuals,

$$e_i = \frac{Y_i - \widehat{\lambda}_i}{\sqrt{\widehat{\lambda}_i}} = \frac{Y_i - \mathbb{E}_{\widehat{\lambda}_i}[Y]}{\sqrt{\text{Var}_{\widehat{\lambda}_i}(Y)}}.$$

We also have the deviance residuals. Note that

$$\text{DEV}(\widehat{\lambda}|Y) = \sum_{i=1}^{n}\text{DEV}(\widehat{\lambda}_i|Y_i).$$

Thus, we can define the deviance residuals as

$$d_i = \text{sign}(Y_i - \widehat{\lambda}_i)\sqrt{\text{DEV}(\widehat{\lambda}_i|Y_i)},$$

recall that $\text{DEV}(\widehat{\lambda}_i|Y_i) = 2\left(\widehat{\lambda}_i - Y_i\log(\widehat{\lambda}_i) - Y_i + Y_i\log(Y_i)\right)$. We also have a hat matrix for the log-linear model. It is given by

$$H_{\widehat{\beta}} = W_{\widehat{\beta}}^{1/2}X(X^T W_{\widehat{\beta}}X)^{-1}X^T W_{\widehat{\beta}}^{1/2},$$

where $W_{\widehat{\beta}} = \exp(X\widehat{\beta})$.

2.2 "Unnatural" models for Poisson data

We can get other models for Poisson data by changing the link function g. The link function satisfies $g(\lambda_i) = X_i^T\beta$ and hence $\lambda_i = g^{-1}(X_i^T\beta)$. The natural choice if $g = \log$ but two other choices are `identity`: $g(\lambda) = \lambda$ and `inverse`: $g(\lambda) = 1/\lambda$. These can be used in `glm()` in R by specifying the link function. For a link function g, the deviance is

$$\mathrm{DEV}(\beta|Y) = 2\sum_{i=1}^{n}\left[g^{-1}(X_i^T\beta) - Y_i\log\left(g^{-1}(X_i^T\beta)\right) - Y_i + Y_i\log(Y_i)\right].$$

We can fit a Poisson glm with Fisher scoring which we can present as an IRLS algorithm with

$$Z^{(k+1)} = X\widehat{\beta}^{(k)} + g'(\widehat{\lambda}^{(k)})(Y - \widehat{\lambda}^{(k)}),$$

and

$$W^{(k+1)} = g'(\widehat{\lambda}^{(k)})^2 V(\widehat{\lambda}^{(k)})^{-1},$$

where $\widehat{\lambda}^{(k)} = g^{-1}(X\widehat{\beta}^{(k)})$ and $V(\lambda) = \lambda$. Note that

$$\nabla\,\mathrm{DEV}(\beta|Y) = 2X^T\left(\frac{\mathbb{E}_\beta[Y] - Y}{g'(\mathbb{E}_\beta[Y])\mathbb{E}_\beta[Y]}\right),$$

and

$$\mathbb{E}_\beta[\nabla^2\,\mathrm{DEV}(\beta|Y)] = 2X^T\left(\frac{1}{g'(\mathbb{E}_\beta[Y])^2\mathbb{E}_\beta[Y]}\right)2X^T W_\beta X,$$

where $W_\beta = \mathrm{diag}\left(\frac{1}{g'(\lambda_i)^2\,\mathrm{Var}_{\lambda_i}(Y)}\right) = \mathrm{diag}\left(\frac{1}{g'(\lambda_i)^2\lambda_i}\right)$, where $\lambda_i = g^{-1}(X_i^T\beta)$. We again have a sandwich estimator for the variance of $\widehat{\beta}^{(k)}$,

$$\widehat{\beta} - \beta^* \approx \mathsf{N}(0, Q^{-1}\Sigma Q^{-1}),$$

where $Q = X^T W_{\beta^*} X$ and $\Sigma = \mathrm{Var}(X^T W^{(\infty)} Z^{(\infty)})$. When the model is correct, $X^T W_{\beta^*} X \approx \Sigma$.

3 Over-dispersion

The Poisson model requires that $\mathbb{E}[Y_i] = \mathrm{Var}(Y_i)$ but this may be far from true. In simple clustering models we in fact have $\mathrm{Var}(Y_i) = \phi\mathbb{E}[Y_i]$ where ϕ has to be estimated from the data. In a Poisson glm we can estimate ϕ with

$$\widehat{\phi} = \frac{1}{n-p}\sum_i e_i^2,$$

where p is the number of parameters and e_i are the Pearson residuals. Another way to incorporate over-dispersion is to work with a negative binomial distribution. Consider the following non-standard parametrization of the negative binomial distribution, $Y \sim \mathsf{Negativebinomial}(\mu, k)$

$$\mathbb{P}(Y = y) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1 - \frac{k}{\mu+k}\right)^y.$$

For a fixed k, this is a one-dimensional exponential family with $\mathbb{E}[Y] = \mu$ and $\mathrm{Var}(Y) = \mu + \frac{\mu^2}{k}$. Thus, by varying k, we get different amounts of over-dispersion in our model. The parameter k can be estimated from the data.