STATS305B – Lecture 6

Jonathon Taylor
Scribed by Michael Howes

01/24/22

Contents

1 Model diagnostics in logistic regression

1.1 Grouped goodness-of-fit-tests

If our covariates X are grouped (for instance X is categorical), then we can use a G^2 or X^2 statistic to measure our model's goodness-of-fit. If we do not have groups, then we can create groups by partitioning the feature space. These goodness-of-fit tests are measuring variation in the counts that is unexplained by our model.

1.2 Pearson's residuals

Define the Pearson's residuals to be,

$$e_i = \frac{Y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}}.$$

If the model is true and if we wrote π_i instead of $\widehat{\pi}_i$, then the above residuals would be independent and with mean 0 and variance 1. But $\widehat{\pi}_i$ is fit to Y, and so e_i may be dependent and have variance less than 1. The residuals are used in *Pearson's χ^2 statistic*,

$$X^2 = \sum_{i=1}^{n} e_i^2.$$

The statistic X^2 can be used as an alternative to the deviance.

1.3 Deviance residuals

An alternative to the Pearson's residuals is to use the decomposition,

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

where $\text{DEV}(\widehat{\pi}_i|Y_i) = -2\left(Y_i \log(\widehat{\pi}_i) + (1-Y_i)\log(1-\widehat{\pi}_i)\right)$ in the binary case. The *deviance residuals* are defined to be,

$$\text{sign}\left(Y_i - \widehat{\pi}_i\right)\sqrt{\text{DEV}(\widehat{\pi}_i|Y_i)}.$$

1.4 Standardized residuals and hat matrices

Is there a way to adjust Pearson's residuals so that they have variance 1? In ordinary least squares regression, we know that

$$r_i = \frac{Y_i - \widehat{Y}_i}{\sqrt{R_{ii}}},$$

have variance σ^2. Where R is the orthogonal projection onto the orthogonal complement of the range of X. That is, $R = I - X(X^TX)^{-1}X^T = I - H$. Can we find something analogous to the hat matrix in logistic regression? Recall that,

$$\widehat{\beta} - \beta^* \approx (X^T W_{\beta^*} X)^{-1} X\left(Y - \pi_{\beta^*}(X)\right) = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}R_{\beta^*}(X,Y),$$

where $\widetilde{X} = W_{\beta^*}(X)^{1/2}X$ and $R_\beta(X,Y) = W_{\beta^*}(X)^{-1/2}(Y - \pi_{\beta^*}(X))$. Thus, logistic regression looks like weighted least squares. Recall the in weighted least squares we use the estimator,

$$\widehat{\beta}_W = \underset{\beta}{\text{argmin}}\left\{\sum_{i=1}^{n} W_i(Y_i - X_i^T\beta)^2\right\}.$$

If $\widetilde{X} = W^{1/2}X$ and $\widetilde{Y} = W^{1/2}Y$, then

$$\widehat{\beta}_W = \underset{\beta}{\text{argmin}}\left\{\sum_{i=1}^{n}(\widetilde{Y}_i - \widetilde{X}_i^T\beta)^2\right\} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\widetilde{Y}.$$

The vector $R_\beta(X,Y) = W_{\beta^*}(X)^{-1/2}(Y - \pi_{\beta^*}(X))$ has independent entries with mean 0 and variance 1. Thus, when viewing logistic regression as weighted least squares, the appropriate hat matrix is,

$$H_{\beta^*} = \widetilde{X}(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T = W_{\beta^*}(X)^{1/2}X(X^TW_{\beta^*}(X)X)^{-1}X^TW_{\beta^*}(X)^{1/2}.$$

Which we have to estimate with,

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

Thus, the standardized residuals are,

$$r_i = \frac{e_i}{\sqrt{1 - H_{\widehat{\beta},ii}}}.$$

These residuals are the leverage scores $H_{\widehat{\beta},ii}$ can be used similarly to how they are used in OLS.

1.5 Analogies of R^2

In OLS, we have $R^2 = \frac{SST - SSE}{SST}$. The analog for logistic regression is thus,

$$R^2 = \frac{\text{DEV}(M_0) - \text{DEV}(M)}{\text{DEV}(M_0)},$$

where M is our model and M_0 is the model with just an intercept.

1.6 Confusion matrices and AUC

So far we have been modelling $\pi(x) = \mathbb{P}(Y|X = x)$ via $\widehat{\pi}(x)$. To make a prediction $\widehat{y}(x) \in \{0, 1\}$, we can pick a threshold c and define,

$$\widehat{y}(x) = \begin{cases} 1 & \text{if } \widehat{\pi}(x) \geq c, \\ 0 & \text{if } \widehat{\pi}(x) < c. \end{cases}$$

For each fixed c, we can create a *confusion matrix* that records the number of correct and incorrect predictions and the predicted values. The confusion matrix looks like this,

		Actual	
		0	1
Fitted	0	True negative (TN)	False negative (FN)
	1	False positive (TP)	True positive (TP)

From this we can calculate the true positive rate (TPR) and the false positive rate (FPR). These are,

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}.$$

Ideally we would have $TPR \approx 1$ and $FPR \approx 0$, but it can be hard to achieve both simultaneously. Both TPR and FPR are functions of our chosen threshold c. By plotting the pair (FPR, TPR) as c varies from 0 to 1, we produce an ROC curve. We can measure our classifier by how far the ROC curve is above the line $y = x$. This can be summarized by the AUC which is the area under the ROC curve. A random assignment of 0 and 1 to $\widehat{y}$ would give a curve with $AUC = 0.5$.

1.7 Model selection

We can use the AIC to select models. The AIC of a model M with $p(M)$ parameters is defined to be

$$\begin{aligned} AIC(M) &= -2 \log L(M) + 2p(M) \\ &= \text{DEV}(M) + 2p(M) + 2g(Y), \end{aligned}$$

where $L(M)$ is the maximized likelihood of the data under the model M. We also have the BIC which is,

$$BIC(M) = -2 \log L(M) + \log(n)p(M).$$

We can then choose a model by picking the one which minimizes $AIC(M)$ or $BIC(M)$. The AIC is a generalization of Mallow's C_p statistic and the B in BIC stands for Bayesian. A warning: if AIC or BIC or something else is used to pick a model M, then all the inference results we derived no longer hold. This is because we are using the data twice. Once to pick the model but then again to do inference. For the inference results, the model was chosen separately to the data.

2 Generalized linear models

Suppose that,

$$f(y|\theta) = \alpha(\theta)b(y) \exp(yQ(\theta)),$$

where,

$$\alpha(\theta) = \int_{\mathcal{Y}} \exp(yQ(\theta))b(y)m(dy).$$

This means that our data y comes from an exponential family with sufficient statistic y, natural parameters $\eta = Q(\theta)$ and reference measure m with density $b(y)$. We can write,

$$\alpha(\theta)^{-1} = \exp(\Lambda(\eta)) = \int_{\mathcal{Y}} \exp(y\eta) b(y) m(dy).$$

Standard calculations give,

$$\nabla\Lambda(\eta) = \mathbb{E}_\eta[Y],$$

and

$$\nabla^2\Lambda(\eta) = \mathrm{Var}_\eta(Y).$$

Thus, the function $\eta \mapsto \mathbb{E}_\eta[Y]$ is increasing and invertible on its range. It follows that $\mathrm{Var}_\eta(Y)$ can be written as a function of $\eta = (\nabla\Lambda)^{-1}(\mathbb{E}_\eta[Y])$. Thus, $\mathrm{Var}(Y) = V(\mu)$ where V is a function and $\mu = \mathbb{E}_\eta[Y]$. Here are some examples,

1. Poisson: If $Y \sim \mathsf{Poisson}(\mu)$, then

$$P_\mu(Y = y) = \exp(y\log(\mu))\frac{e^{-\mu}}{y!}.$$

   The natural parameter if $\log(\mu)$ and here $\mathrm{Var}_\mu(Y) = \mu = V(\mu) = V(\mathbb{E}_\mu[Y])$. The natural regression model is $\log(\mu_i) = X_i^T\beta$.

2. Bernoulli: If $Y \sim \mathsf{Bernoulli}(\pi)$, then

$$P_\pi(Y = y) = \exp(y\log(\pi) + (1-y)\log(1-\pi)) = \exp(y\log(\pi/(1-\pi)) - \log(1-\pi)).$$

   Thus, the natural parameter is $\eta = \log(\pi/(1-\pi))$, and we have our familiar logistic regression model. In this case, $\mathrm{Var}_\pi(Y) = \pi(1-\pi) = V(\pi) = V(\mathbb{E}_\pi(Y))$.

3. Other binary models. Let F be a fixed CDF and let $\pi_i = F(X_i^T\beta)$ so that $F^{-1}(\pi_i) = X_i^T\beta$. Different choices of F give us different GLMS for binary data. Some examples are

   - Logistic: $F^{-1}(\pi) = \mathsf{logit}(\pi)$.
   - Probit: $F^{-1}(\pi) = \Phi^{-1}(\pi)$.
   - cloglog: $F^{-1}(\pi) = -\log(-\log(\pi))$.

   Each of these can be used in R's `glm(family = binomial())` by specifying the link. The parameters in other models may be less interpretable than in the logistic model.

A generalized linear model is model for $Y|X$ where we have i.i.d. data $(X_i, Y_i)_{i=1}^n$. We specify the following,

1. We model $\eta_i = g(\mathbb{E}(Y_i|X_i)) = g(\mu_i) = X_i^T\beta$ where g is the *link function* for the model.
2. If $\mathrm{Var}(Y_i|X_i) = \phi V(\mathbb{E}(Y_i|X_i)) = \phi V(\mu_i)$ for some *dispersion parameter* $\phi > 0$ and *variance function* V.
3. The glm is specified by the pair (g, V). If $Y_i|X_i$ comes from a one-dimensional exponential family, then V is determined and there is "canonical" choice for g.

We will now briefly talk about fitting a glm for binary data with $\pi_i = F^{-1}(X_i^T\beta)$ where F is a CDF with density f. The deviance for such a model is,

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

Like logistic regression, we could try to minimize the deviance by using Newton–Raphson. The gradient of the deviance is,

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

Clearly the Hessian is going to be quite complicated. However, $\mathbb{E}[Y_i|X_i] = F(X_i^T\beta)$. Thus, if we calculate the Hessian, and then compute the expected value of the Hessian, many terms will cancel. Indeed,

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

where $W_\beta(X) = \operatorname{diag}\left(\frac{f(X_i^T\beta)}{F(X_i^T\beta)(1-F(X_i^T\beta))}\right)$, which looks a lot like our Hessian from logistic regression. But is it valid to replace the Hessian with the expected value of the Hessian when doing Newton–Raphson? It turns out this it is okay and is an example of an optimization technique called a *quasi-Newton* method. This is how we will optimize the deviance in glms.

3 Quasi-Newton methods

In Newton–Raphson, we approximate the objective function l with its quadratic Taylor's approximation. We then minimize the quadratic Taylor's approximation and iterate. More precisely, suppose we want to minimize l, and we have a current guess β_c. We then have,

$$l(\beta) \approx l(\beta_c) + \nabla l(\beta_c)^T(\beta-\beta_c) + \frac{1}{2}(\beta-\beta_c)^T\nabla^2 l(\beta_c)(\beta-\beta_c).$$

We can easily minimize the quadratic approximation which gives us a new value β_n. Standard calculus gives,

$$\beta_n = \beta_c - \nabla^2 l(\beta_c)^{-1}\nabla l(\beta_c).$$

Iterating this gives us Newton–Raphson. *Quasi-Newton* methods are a class of methods where we replace $\nabla^2 l(\beta_c)$ with another positive semi-definite matrix $H(\beta_c)$ such that $H(\beta_c)$ is larger than $\nabla^2 l(\beta_c)$ in the positive semi-definite ordering. More specifically suppose that we have a matrix $H(\beta_c)$ such that for all β,

$$l(\beta) \le l(\beta_c) + \nabla l(\beta_c)^T(\beta-\beta_c) + \frac{1}{2}(\beta-\beta_c)^T H(\beta_c)(\beta-\beta_c).$$

If we set $\beta_n = \beta_c - H^{-1}\nabla l(\beta_c)$, then β_n minimizes the above quadratic and hence,

$$l(\beta_n) \le l(\beta_c) + \nabla l(\beta_c)^T(\beta_n-\beta_c) + \frac{1}{2}(\beta_n-\beta_c)^T H(\beta_c)(\beta_n-\beta_c) \le l(\beta_c).$$

Thus, β_n reduces the objective l. A quasi-Newton method is an iterative version of this procedure. Some examples are,

- If $l(\beta) = \operatorname{DEV}(\beta|Y)$ and $H(\beta) = \mathbb{E}_\beta[\nabla^2 l(\beta)]$, then we get a method called *Fisher scoring*. This is a common way to fit glms.
- If l is any objective and $H(\beta) = \lambda_\beta I$, then the resulting method is gradient descent with step-size λ_β^{-1}. The step-size has to be sufficiently small to guarantee that $H(\beta)$ is bigger than $\nabla^2 l(\beta)$ in the positive semi-definite ordering.