STATS305B – Lecture 5

Jonathon Taylor
Scribed by Michael Howes

01/19/22

Contents

1 Logistic regression

Suppose we have data $(X_i, Y_i)_{i=1}^n$ where $X_i \in \mathbb{R}^p$ and $Y_i \in \{0, 1\}$. The likelihood of $y = (Y_i)_{i=1}^n$ is determined by $\pi_i = \mathbb{P}(Y_i = 1 | X_i = 1)$. To work with natural parameters we use

$$\eta_i = \mathsf{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right).$$

So that $\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$. The likelihood is thus,

$$-\log L(\eta | Y) = \sum_{i=1}^n -Y_i \eta_i + \log(1 + e^{\eta_i}).$$

In the logistic regression model we parametrize η as a linear function of X, that is $\eta_i = X_i^T \beta$. The likelihood for β is thus,

$$-\log L(\beta | Y) = -(X\beta)^T Y + \sum_{i=1}^n \log(1 + e^{X_i^T \beta}).$$

We then fix β by minimizing the above negative log likelihood.

1.1 An example in R

In R, a logistic regression model can be fitted by using the command `glm()` and specifying the input `family = binomial`. The syntax for fitting a model to a data set is analogous to `lm()`. The command `summary()` when applied to a model gives information about the residuals, the null deviance as well as p-values for the coefficients calculated using Wald tests. The p-values test the nulls $\beta_j = 0$. The command `anova()` can also be used to calculate the difference in deviance which is analogous to the

difference in the error sum of squares between two linear models. But be careful, in logistic regression, the square of the Wald test statistic does not equal the difference of deviances. If you call `glm()` without specifying a family the default `gaussian` is used. This essentially results in fitting a standard linear model.

1.2 Regression and $I \times 2$ tables

Suppose that we have an $I \times 2$ table. We can define $Y_i = N_{i1}$ for $1 \leq i \leq I$. If we have independent rows, then $Y_i \sim \mathsf{Binomal}(N_{i+}, \pi_i)$. We can treat the row index as categorical dependent variable in a logistic regression. This then gives the model $\mathsf{logit}(\pi_i) = \alpha + \beta_i$ for some parameters α and β. This is analogous to a one-way ANOVA linear regression. The model $\mathsf{logit}(\pi_i) = \alpha + \beta_i$ is non-identifiable since we have $I + 1$ parameters. We need to introduce a constraint which normally takes the form of $\beta_i = 0$ for some fixed i. This model is *saturated* since the number of parameters equals the number of data points.

We can use `glm(family = binomial)` in R to fit a logistic regression model to a contingency table. The counts N_{i1} are used as the dependent variables and give a factor of the row labels as the independent variables. We also have to specify the input `weights` which in this example will be the row totals N_{i+}. This input `weights` is what allows us to use logistic regression for binomial as well as binary data.

When fitting a logistic regression model to an $I \times 2$ table, testing against the model with just an intercept is analogous to testing for independence. The difference of deviances between the full model and the model with just an intercept equals the likelihood ratio test for independence.

If our row labels are ordinal, we can fit a simpler model than $\mathsf{logit}(\pi_i) = \alpha + \beta_i$. The simpler model works by first assigning scores to each row and then using $\mathsf{logit}(\pi) = \alpha + \beta \cdot \text{score}_i$. This ordinal model has fewer degrees on freedom than the saturated model. The difference in deviance test against independence may therefore be more powerful since the p-values are calculated using a χ^2 distribution with fewer degrees of freedom.

1.3 Deviance

In logistic regression, the *deviance* of a model takes the place of the sum of square residuals in linear regression. The deviance is defined to be,

$$\text{DEV}(\pi|Y) = -2 \log L(\pi|Y) + 2 \log L(\widehat{\pi}_s|Y),$$

where π is the MLE for π within the model and $\widehat{\pi}_s$ is an estimator fit using the saturated model. The estimator $\widehat{\pi}_s$ is fit by taking the number of parameters to equal the number of data points and without any constraints. For binary data $\widehat{\pi}_{s,i} = Y_i$ and $2 \log L(\widehat{\pi}_s|Y) = 0$. For binomial data $Y_i \sim \mathsf{Binomal}(n_i, \pi_i)$,

$$\widehat{\pi}_{s,i} = \frac{Y_i}{n_i},$$

and we typically don't have $\log L(\widehat{\pi}_s|Y) = 0$. If Y is a binary vector, then

$$\text{DEV}(\pi|Y) = -2 \log L(\pi|Y) = -2 \sum_{i=1}^{n} (Y_i \log(\pi_i) + (1 - Y_i) \log(1 - \pi_i)).$$

For logistic regression, we have

$$\text{DEV}(\pi|Y) = -2(X\beta)^T Y + 2 \sum_{i=1}^{n} \log\left(1 + \exp(X_i^T \beta)\right).$$

This is because $\mathsf{logit}(\pi_i) = X_i^T\beta$ and thus,

$$\begin{aligned}
\mathrm{DEV}(\beta|Y) &= -2\sum_{i=1}^{n} Y_i \log\left(\frac{\pi_i(\beta)}{1-\pi_i(\beta)}\right) + \log(1-\pi_i(\beta)) \\
&= -\sum_{i=1}^{n} Y_i \mathsf{logit}(\pi_i) + \log(1-\pi_i) \\
&= -\sum_{i=1}^{n} Y_i X_i^T\beta + \log\left(1 - \frac{e^{(X_i^T\beta)}}{1+e^{X_i^T\beta}}\right) \\
&= -2(X\beta)^T Y - 2\sum_{i=1}^{n} \log\left(\frac{1}{1+e^{X_i^T\beta}}\right) \\
&= -2(X\beta)^T Y + 2\sum_{i=1}^{n} \log\left(1+\exp(X_i^T\beta)\right).
\end{aligned}$$

1.4 Fitting logistic regression

How do we solve the optimization problem,

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} -\log L(\beta|Y).$$

We have previously seen that $-\log L(\beta|Y)$ is convex in β and thus has a unique minimum. To calculate this minimizer we can use the iterative Newton–Raphson method. To do this we need to calculate the Hessian of $-\log L(\beta|Y)$. Note that,

$$\begin{aligned}
\nabla(-\log L(\beta|Y)) &= \nabla\left(-(X\beta)^T Y + \sum_{i=1}^{n} \log(1+\exp(X_i^T\beta))\right) \\
&= -X^T Y + \sum_{i=1}^{n} \nabla\left(\log(1+\exp(X_i^T\beta))\right).
\end{aligned}$$

We know that $\log(1+\exp(\eta_i))$ is the cummulant generating function for y and thus $\nabla_{\eta_i}\log(1+\exp(\eta_i)) = \mathbb{E}_{\eta_i}[Y_i]$ and $\nabla^2_{\eta_i}\log(1+\exp(\eta_i)) = \mathrm{Var}_{\eta_i}[Y_i]$. The chain rule then gives,

$$\sum_{i=1}^{n} \nabla\left(\log(1+\exp(X_i^T\beta))\right) = \sum_{i=1}^{n} X_i^T \mathbb{E}_\beta[Y_i|X_i] = X^T\mathbb{E}_\beta[Y|X].$$

Thus,

$$\nabla(-\log L(\beta|Y)) = -X^T Y + X^T\mathbb{E}_\beta[Y|X] = -X^T(Y - \mathbb{E}_\beta[Y|X]).$$

Taking a further derivative gives,

$$\nabla^2(-\log L(\beta|Y)) = X^T\nabla\mathbb{E}_\beta[Y|X] = X^T\,\mathrm{Var}_\beta(Y|X)X^T = X^T\,\mathrm{diag}\left(\frac{\exp(X\beta)}{(1+\exp(X\beta))^2}\right)X.$$

Define,

$$\pi_\beta(X) = \mathbb{E}_\beta[Y|X] = \frac{\exp(X\beta)}{1+\exp(X\beta)},$$

and

$$W_\beta(X) = \mathrm{Var}_\beta(Y|X) = \mathrm{diag}(\pi_\beta(X)(1-\pi_\beta(X))).$$

The Newton–Raphson method for logistic regression is thus,

1. Begin with an initial guess $\widehat{\beta}^{(0)}$.
2. Until convergence, define,

$$\widehat{\beta}^{(t+1)} = \widehat{\beta}^{(t)} - \left(X^T W_{\widehat{\beta}^{(t)}}(X) X\right)^{-1} X^T (Y - \pi_{\widehat{\beta}^{(t)}}(X)).$$

Note that the difference $\delta^{(t+1)} = \widehat{\beta}^{(t)} - \widehat{\beta}^{(t+1)}$ is the solution to

$$\operatorname*{argmin}_{\delta} \sum_{i=1}^{m} W_{ii}^{(t)} \left(X_i^T \delta - r_i^{(t)}\right)^2,$$

where $W_{ii}^{(t)}$ are the diagonal elements of $W_{\widehat{\beta}^{(t)}}(X)$ and $r^{(t)}$ is the vector of scaled residuals with entries $r_i^{(t)} = \frac{Y_i - \pi_{\widehat{\beta}^{(t)}}(X_i)}{W_{ii}^{(t)}}$. Thus, the Newton–Raphson method is a type of iteratively re-weighted least squares. This was an important fact when people were first fitting generalized linear models.

1.5 The distribution of $\widehat{\beta}$

The MLE $\widehat{\beta}$ satisfies $\nabla - \log L(\widehat{\beta}|Y) = 0$ and thus satisfies,

$$X^T (Y - \pi_{\widehat{\beta}}(X)) = 0.$$

If the logistic regression model is true for some parameter β^*, then

$$\mathbb{E}_{\beta^*}[X^T (Y - \pi_{\beta^*}(X))] = 0,$$

A Taylor series approximation gives,

$$\pi_{\widehat{\beta}}(X) \approx \pi_{\beta^*}(X) + W_{\beta^*}(X) X (\widehat{\beta} - \beta^*),$$

since $\nabla \pi_{\beta^*}(X) = W_{\beta^*} X$. The error in the above Taylor's approximation is of the order $\left\|\widehat{\beta} - \beta^*\right\|_2^2$ which is of the order $\frac{1}{n}$. If we rearrange the above approximation and multiply by X^T, then we get

$$X^T W_{\beta^*}(X) X (\widehat{\beta} - \beta_*) \approx X^T \pi_{\widehat{\beta}}(X) - X^T \pi_{\beta^*}(X) = X^T (Y - \pi_{\beta^*}(X)),$$

by stationarity. Thus,

$$\widehat{\beta} - \beta^* \approx (X^T W_{\beta_*}(X) X)^{-1} X^T (Y - \pi_{\beta^*}(X)).$$

Thus, the expectation $\mathbb{E}_{\beta^*}[\widehat{\beta} - \beta^*]$ is asymptotically 0 and the asymptotic variance is,

$$\begin{aligned} \operatorname{Var}_{\beta^*}(\widehat{\beta} - \beta^*) &\approx (X^T W_{\beta^*}(X) X)^{-1} \operatorname{Var}_{\beta^*}(X^T (Y - \pi_{\beta^*}(X)))(X^T W_{\beta^*}(X) X)^{-1} \\ &\approx (X^T W_{\beta^*}(X) X)^{-1}. \end{aligned}$$

This is because,

$$\operatorname{Var}_{\beta^*}(X^T (Y - \pi_{\beta^*}(X))) \approx X^T W_{\beta^*}(X) X.$$

The approximation is valid due to the strong law of large numbers. For large values n we can approximate the variance with the sample variance. In practice, we do not know β^*, so instead we plug in $\widehat{\beta}$. We thus approximately have

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

This approximation can be used to construct Wald confidence intervals. In R, `confint()` gives profile likelihood intervals when applied to a logistic regression model.

We can also ask about the distribution of $\widehat{\beta}$ when the assumptions of the logistic regression model do not hold. Suppose that $(X_i, Y_i) \overset{\text{iid}}{\sim} F$ with Y_i binary. We can define $\beta^* = \beta^*(F)$ by the equation,

$$\mathbb{E}_F[X(Y - \pi_{\beta^*(F)})] = 0.$$

The parameter β^* is a population parameter. Another way to describe $\beta^*(F)$ is as the minimizer of,

$$\beta \mapsto \mathbb{E}_F[-(X^T\beta)Y + \log(1 + \exp(X^T\beta))].$$

If we perform the same Taylor expansion and rearranging around $\beta^*(F)$, we get

$$\widehat{\beta} - \beta^*(F) \approx (X^T W_{\beta^*(F)} X)^{-1} X^T (Y - \pi_{\beta^*(F)}(X)).$$

Under suitable conditions on F, we will have,

$$n^{-1/2} X^T (Y - \pi_{\beta^*(F)}) \xrightarrow{d} \mathsf{N}(0, \Sigma(F)).$$

We will also have

$$n^{-1} X^T W_{\beta^*(F)}(X) X \xrightarrow{p} \mathbb{E}_F[X^T W_{\beta^*(F)} X] =: Q(F).$$

Thus,

$$n^{1/2}(\widehat{\beta} - \beta^*(F)) = (n^{-1} X^T W_{\beta^*(F)}(X) X)^{-1} \left(n^{-1/2} X^T (Y - \pi_{\beta^*(F)}) \right) \xrightarrow{d} \mathsf{N}(0, Q(F)^{-1} \Sigma(F) Q(F)^{-1}).$$

This estimate for the variance is called the *sandwich form*. We can estimate $Q(F)$ with $\frac{1}{n} X^T W_{\widehat{\beta}}(X) X$. This gives,

$$\widehat{\beta} - \beta^*(F) \approx \mathsf{N}(0, (X^T W_{\widehat{\beta}}(X) X)^{-1} (n\Sigma(F)) (X^T W_{\widehat{\beta}}(X) X)^{-1}).$$

We now know all the terms in our sandwich estimator apart from $\Sigma(F)$. We can estimate $\Sigma(F)$ be bootstrapping, or we could use the bootstrap to estimate $\text{Var}(\widehat{\beta})$ directly. One upshot is that, when the model doesn't hold we can still get standard errors for $\widehat{\beta}$, but the expression is more complicated since we don't know $\Sigma(F)$ (if the model does hold $\Sigma(F) = Q(F) = \text{Var}_{\beta^*}(X^T(Y - \pi_{\beta^*}(X))) \approx X^T W_{\beta^*(F)} X$). When the model doesn't hold we do have to be more careful about how we interpret the deviance. The asymptotic distribution may not hold.

1.6 Testing

In linear regression, we compare a sub-model to a larger model by comparing the sum of square errors. More precisely, suppose we have models $M_R \subseteq M_F$ with residual degrees of freedom $df_R > df_F$ (the residual degrees of freedom is the number of samples minus the number of parameters, the model M_F has more parameters and thus fewer residual degrees of freedom). Under the null that M_R contains the true distribution, we know that

$$\frac{1}{\sigma^2} \left(SSE(M_R) - SSE(M_F) \right) \sim \chi^2_{df_R - df_F}.$$

In logistic regression, we have

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

These tests both assume that M_R is rich enough to contain the true distribution. If this is not true, then a Wald test with the sandwich estimator should be used. The Wald test can be used when the model is true as well, but the Wald test will give different results to the difference of deviances test.

1.7 Model diagnostics

Like in the linear model, we can assess the model assumptions by studying the residuals. Recall that Pearson's residuals are,

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

Unfortunately the Pearson residuals are not properly standardized. Their variance is smaller than that of the true errors (this is because $\widehat{\pi}$ was fit to Y_i – just like the linear model). The *standardized residuals*,

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

are a better match for the true errors. The matrix $H_{\widehat{\beta}}$ is analogous to the hat matrix from linear regression. For logistic regression it is defined to be

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

Like in the linear model, we can do model diagnostics by plotting r against the fitted values. However, since the response Y is binary, the resulting plots look very different to the plots from linear models. The response being binary means that the residuals will always have some structure. In particular, we expect the there two be two clusters of residuals. One corresponding to $Y = 1$ and another corresponding to $Y = 0$.

One way to get more meaningful plots is to group the residuals into bins and then plot the sum of the residuals against the sum of the fitted values for each bin. We could also put these binned values into a contingency table and then test for independence.