Generalized linear models: likelihood ratio test

1 Introduction

- We return to GLMs. As a reminder, we model the data as being observed values of independently sampled $(X_1, Y_1), \ldots, (X_n, Y_n)$, such that

$$Y_i | X_i = x \sim P_{Y|X}, \quad \mathbb{E}[Y_i | X_i = x] := m(x) = g^{-1}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)$$

  In other words, the conditional distribution of $Y_i|X_i$ has mean $m(x) := \mathbb{E}[Y_i|X_i = x]$ that is linear after being transformed by some known link function g. A case of particular importance is logistic regression: we suppose $Y_i|X_i = x \sim \text{Bern}(m(x))$, where the conditional mean is

$$m(x) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}.$$

- If the distribution of Y is completely determined by its mean – as is the case with Poisson, Binomial, and Bernoulli distributions, for any choice of link function – then this completely specifies the model. If not – as is the case with the Normal distribution – then we may need to introduce some additional nuisance parameters (like σ^2) to specify the conditional density $Y_i|X_i = x$. **For convenience, in today's notes I will assume there are no nuisance parameters; e.g. in the case of the Normal model, the error variance is known.**
- Today we discuss hypothesis testing for coefficients in a generalzied linear model.

2 Wald test

- Suppose we are interested in testing whether predictor j adds anything to a model that includes all of the other predictors. This corresponds to the null hypothesis

$$H_{0,j} : \beta_j = 0.$$

- Last time we worked out the asymptotic sampling distribution of the maximum likelihood estimator $\hat{\beta}$. Recall that $I(\beta) = \mathbb{E}[-\nabla^2 \ell(\beta)] \in \mathbb{R}^{p+1 \times p+1}$ is the **Fisher information** matrix. Letting β^* be the true parameters, we have that $\hat{\beta}$ converges in distribution,

$$\hat{\beta} \to N\Big(\beta^*, \big(I(\beta^*)\big)^{-1}\Big).$$

  In the special case of logistic regression,

$$I(\beta^*) = \mathbf{X}^\top \mathbf{V}_{\beta^*} \mathbf{X},$$

  where $\mathbf{V}_{\beta^*} = \text{diag}(m(x_i)(1 - m(x_i)))$ is a diagonal matrix with ith diagonal element equal to the conditional variance of $Y_i|X_i = x_i$.

- It follows that if the null hypothesis $\beta_j = 0$ is correct then

$$\frac{\hat{\beta}_j - 0}{\sqrt{[I(\beta^*)^{-1}]_{j+1,j+1}}} \to N(0,1).$$

Plugging in an estimate for the Fisher information leads to the **Wald statistic**:

$$W_j = \frac{\hat{\beta}_j - 0}{\sqrt{[I(\hat{\beta})^{-1}]_{j+1,j+1}}}.$$

The **Wald test** rejects the null hypothesis if $|W_j| > z^{1-\alpha/2}$, where as usual $z^{1-\alpha/2}$ is the $1-\alpha/2$th quantile of a Normal distribution. Asymptotically, as $n \to \infty$, this test has a $100\alpha\%$ false positive rate.

3 Likelihood ratio test

- Now suppose we are interested in testing whether all of the predictors together add anything to a model that includes just the intercept. This corresponds to the null hypothesis:

$$H_0 : \beta_1 = \ldots = \beta_p = 0.$$

As we discussed previously, this null hypothesis can equivalently be written as $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independently sampled,

$$Y_i | X_i = x \sim P_{Y|X}, \quad \mathbb{E}[Y_i | X_i = x] = g^{-1}(\gamma_0).$$

We could define a statistic based on the average magnitude of the estimated coefficients, i.e $S^2 = \sum_{j=1}^p \hat{\beta}_j^2$, and calibrate it using the asymptotic Normality of $\hat{\beta}$. This is called a **chi-squared test**. Instead we will discuss another test for the same null hypothesis called the **likelihood ratio test**.

- Remember that the **likelihood** is the probability of the data as a function of the parameters. Let $f(y_i|x_i;\beta)$ represent the probability of $y_i|x_i$ at a given value of parameters β. If we observe $(x_1, y_1), \ldots, (x_n, y_n)$ then the likelihood is

$$\prod_{i=1}^n f(y_i|x_i;\beta) \tag{1}$$

- To measure how much the predictors improve the model, we compare the likelihood of the full model to the likelihood of the null model, resulting in the **likelihood ratio**:

$$R = \frac{\max_{\beta_0,\beta_1,\ldots,\beta_p} \prod_{i=1}^n f(y_i|x_i;\beta_0,\beta_1,\ldots,\beta_p)}{\max_{\gamma} \prod_{i=1}^n f(y_i|x_i;\gamma,0,\ldots,0)}$$

Notice that in both numerator and denominator we are actually finding the maximum likelihood of the data over all possible values of the parameters, in the full and null models, respectively. Hence this should really be called the maximum likelihood ratio statistic. It is also sometimes called **Wilks' statistic**.

- Larger values of the likelihood ratio give more evidence against the null hypothesis. But adding more parameters to the model always increases the maximum value of the likelihood, and so the ratio of the maximum likelihood of the full model over the maximum likelihood of the null model will always be at least 1. We need to compare the observed value to what we would expect to see if the null hypothesis were correct.

- It can be shown that if the null hypothesis is correct then twice the log of the likelihood ratio is asymptotically chi-squared distributed with p degrees of freedom:

$$2 \log R \to \chi^2_p. \tag{2}$$

We use (2) to calibrate our test: that is, we reject the null hypothesis if $2 \log R$ is greater than the $(1-\alpha)$th quantile of a χ^2_p distribution.

The proof of (2) relies on the asymptotic normality of $\hat{\theta}$ along with a second-order Taylor expansion. From this proof it can be seen that the asymptotic approximation begins to break down when $n - p$ is too small (say, less than 20.) We won't go into the details beyond that.

- **Special case**: Consider the case $Y_i | X_i = x_i \sim N(m(x_i), \sigma^2)$ of linear regression with a known variance σ^2. The log-likelihood of the full model and null model are respectively

$$\ell_n(\hat{\beta}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 + const, \quad \ell_n(\hat{\gamma}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \bar{y})^2 + const$$

where const refers to a term that is the same across models. So twice the difference in log-likelihood is

$$2 \log R = 2(\ell_n(\hat{\theta}) - \ell_n(\hat{\theta}_\omega)) = \frac{\left(\|Y - \hat{Y}\|^2 - \|Y - \bar{Y}\mathbf{1}\|^2 \right)}{\sigma^2}.$$

We recognize the numerator as being the numerator of the F-statistic multiplied by a factor p. We worked out that this was exactly distributed χ^2_p. So we see that for the usual linear model (2) is exact.

- The likelihood ratio test extends in a relatively straightforward way to testing whether a group of predictors adds to a model with other predictors included. This corresponds to the null hypothesis

$$H_0 : \beta_{q+1} = \ldots = \beta_p = 0.$$

Our test statistic is again the ratio between the maximum likelihood under the full model and the maximum likelihood under the null model:

$$R_{0:q} = \frac{\max_{\beta_0, \beta_1, \ldots, \beta_p} \prod_{i=1}^{n} f(y_i | x_i; \beta_0, \beta_1, \ldots, \beta_p)}{\max_{\gamma_0, \gamma_1, \ldots, \gamma_q} \prod_{i=1}^{n} f(y_i | x_i; \gamma_0, \gamma_1, \ldots, \gamma_q, 0, \ldots, 0)}$$

If the null hypothesis is correct then $2 \log R_{0:q}$ is asymptotically χ^2_{p-q}, and so we reject the null if $2 \log R_{0:q}$ is greater than the $(1-\alpha)$th quantile of the χ^2_{p-q} distribution.

- We see that there are two ways of testing whether $\beta_j = 0$, the Wald test and the likelihood ratio test. For the linear model, these two tests will be the same: either both will reject the null hypothesis, or neither will. For the generalized linear model, this is only asymptotically true. For small number of samples, the tests may differ: in this case the likelihood ratio test is usually a bit more accurate if all of the assumptions of the model are correct.

4 Deviance

- In diagnostics for linear regression the residuals $e_i = (y_i - \hat{y}_i)$ played an important role for both inference and diagnostics. In generalized linear models the corresponding quantity is called the **deviance**. To understand the deviance we need one more model, the **saturated model**, which supposes that there is one parameter per observation:

$$m(x_i) = \gamma_i.$$

Let $\hat{\gamma}$ be the MLE of the saturated model; for typical GLMs this will just be $\hat{\gamma}_i = y_i$. The deviance measures the contribution of observation i to the difference in log-likelihood between the saturated model and the full-model:

$$\text{dev}_i := \log \left(\frac{f(y_i | x_i; \hat{\gamma})}{f(y_i | x_i; \hat{\beta})} \right)$$

Notice that maximizing the likelihood is the same as minimizing the sum of the deviances

$$\text{dev}(y; x; \beta) := \sum_{i=1}^{n} \text{dev}(y_i, x_i^\top \beta)$$

Similarly, the likelihood ratio is equal to the difference in sum of deviances.

- In linear regression with Normal errors and known error variance, the deviance is

$$\text{dev}_i = \frac{(y_i - \hat{y}_i)^2}{\sigma^2}$$

In Bernoulli regression,

$$\text{dev}_i = -\Big(y_i \log(\hat{m}(x_i)) + (1 - y_i) \log(1 - \hat{m}(x_i))\Big).$$

- Diagnostics for GLMs is important but even fuzzier than diagnostics for LMs, so we will not go over it.

5 Model selection

- Remembering that AIC was defined in terms of likelihoods, it is easy to do variable selection by minimizing AIC. Any search strategy – forward, backward, stepwise, or best subsets – can be used.
- One can also run the lasso by minimizing the sum of deviances plus an ℓ^1 penalty on the coefficients. This is handled by the `glmnet` package.