Lecture 13
Estimation and hypothesis testing for logistic regression

BIOST 515

February 19, 2004

Outline

- Review of maximum likelihood estimation
- Maximum likelihood estimation for logistic regression
- Testing in logistic regression

Maximum likelihood estimation

Let's begin with an illustration from a simple bernoulli case.

In this case, we observe independent binary responses, and we wish to draw inferences about the probability of an event in the population. Sound familiar?

- Suppose in a population from which we are sampling, each individual has the same probability, p, that an event occurs.
- For each individual in our sample of size n, $Y_i = 1$ indicates that an event occurs for the ith subject, otherwise, $Y_i = 0$.
- The observed data is $Y_1, \ldots, Y_n$.

The joint probability of the data (the **likelihood**) is given by

$$\begin{aligned} L &= \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i} \\ &= p^{\sum_{i=1}^{n} Y_i}(1-p)^{n-\sum_{i=1}^{n} Y_i}. \end{aligned}$$

For estimation, we will work with the **log-likelihood**

$$l = log(L) = \sum_{i=1}^{n} Y_i \log(p) + (n - \sum_{i=1}^{n} Y_i) log(1-p).$$

The maximum likelihood estimate (MLE) of p is that value that maximizes l (equivalent to maximizing L).

The first derivative of l with respect to p is

$$U(p) = \frac{\partial l}{\partial p} = \sum_{i=1}^{n} Y_i/p - (n - \sum_{i=1}^{n} Y_i)/(1-p)$$

and is referred to as the **score funcion**. To calculate the MLE of p, we set the score function, $U(p)$ equal to 0 and solve for p. In this case, we get an MLE of p that is

$$\hat{p} = \sum_{i=1}^{n} Y_i/n.$$

Information

Another important function that can be derived from the likelihood is the **Fisher information** about the unknown parameter(s). The information function is the negative of the curvature in $l = \log L$. For the likelihood considered previously, the information is

$$\begin{aligned} I(p) &= E\left[-\frac{\partial^2 l}{\partial p^2}\right] \\ &= E\left[\sum_{i=1}^{n} Y_i/p^2 + (n - \sum_{i=1}^{n} Y_i)/(1-p)^2\right] \\ &= \frac{n}{p(1-p)} \end{aligned}$$

We can estimate the information by substituting the MLE of p into $I(p)$, yielding $I(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})}$.

Our next interest may be in making inference about the parameter p. We can use the the inverse of the information evaluated at the MLE to estimate the variance of $\hat{p}$ as

$$\widehat{\text{var}}(\hat{p}) = I(\hat{p})^{-1} = \frac{\hat{p}(1-\hat{p})}{n}.$$

For large n, $\hat{p}$ is approximately normally distributed with mean p and variance $p(1-p)/n$. Therefore, we can construct a $100 \times (1-\alpha)\%$ confidence interval for p as

$$\hat{p} \pm Z_{1-\alpha/2}[\hat{p}(1-\hat{p})/n]^{1/2}.$$

Hypothesis tests

- Likelihood ratio tests
- Wald tests
- Score tests

Likelihood ratio tests

The likelihood ratio test (LRT) statistic is the ratio of the likelihood at the hypothesized parameter values to the likelihood of the data at the MLE(s). The LRT statistic is given by

$$\begin{aligned} LR &= -2\log\left(\frac{L \text{ at } H_0}{L \text{ at } MLE(s)}\right) \\ &= -2l(H_0) + 2l(MLE). \end{aligned}$$

For large n, $LR \sim \chi^2$ with degrees of freedom equal to the number of parameters being estimated.

For the binary outcome discussed above, if the hypothesis is $H_0 : p = p_0$ vs $H_A : p \neq p_0$, then

$$l(H_0) = \sum_{i=1}^{n} Y_i \log(p_0) + (n - \sum_{i=1}^{n} Y_i) \log(1 - p_0),$$

$$l(MLE) = \sum_{i=1}^{n} Y_i \log(\hat{p}) + (n - \sum_{i=1}^{n} Y_i) \log(1 - \hat{p})$$

and the LRT statistic is

$$LR = -2 \left[\sum_{i=1}^{n} Y_i \log(p_0/\hat{p}) + (n - \sum_{i=1}^{n} Y_i) \log\{(1 - p_0)(1 - \hat{p})\} \right],$$

where $LR \sim \chi^2_1$.

Wald test

The **Wald test statistic** is a function of the difference in the MLE and the hypothesized value, normalized by an estimate of the standard deviation of the MLE. In our binary outcome example,

$$W = \frac{(\hat{p} - p_0)^2}{\hat{p}(1 - \hat{p})/n}.$$

For large n, $W \sim \chi^2$ with 1 degree of freedom. In R, you will see $\sqrt{W} \sim N(0,1)$ reported.

Score test

If the MLE equals the hypothesized value, p_0, then p_0 would maximize the likelihood and $U(p_0) = 0$. The score statistic measures how far from zero the score function is when evaluated at the null hypothesis. The test statistic for the binary outcome example is

$$S = U(p_0)^2/I(p_0),$$

and $S \sim \chi^2$ with 1 degree of freedom.

How does this all relate to logistic regression?

So far, we've learned how to estimate p and to test p in the one-sample bernoulli case. How can we use this with logistic regression.

- MLEs for βs
- Testing hypotheses about the βs
- Constructing confidence intervals for the βs

MLEs for coefficients in logistic regression

Remember:

$$E[Y_i] = \pi_i$$

and

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = x_i'\beta,$$

where $x_i = (1, x_{i1}, \ldots x_{ip})'$ and $\beta = (\beta_0, \ldots, \beta_p)'$. Therefore

$$E[Y_i] = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}.$$

The likelihood for n observations is then

$$L = \prod_{i=1}^{n} \left(\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \right)^{\sum_{i=1}^{n} Y_i} \left(1 - \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \right)^{n-\sum_{i=1}^{n} Y_i},$$

and the log-likelihood is

$$l = \sum_{i=1}^{n} \left[Y_i \log\left(\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \right) + \right.$$
$$\left. (1-Y_i)\log\left(1 - \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \right) \right].$$

The $p + 1$ score functions of β for the logistic regression model cannot be solved analytically. It is common to use a numerical algorithm, such as the Newton-Raphson algorithm, to obtain the MLEs.

The information in this case will be a $(p + 1) \times (p + 1)$ matrix of the partial second derivatives of l with respect to the parameters, β. The inverted information matrix is the covariance matrix for $\hat{\beta}$.

Testing a single logistic regression coefficient in R

To test a single logistic regression coefficient, we will use the Wald test,

$$\frac{\hat{\beta}_j - \beta_{j0}}{\hat{se}(\hat{\beta})} \sim N(0,1),$$

where $\hat{se}(\hat{\beta})$ is calculated by taking the inverse of the estimated information matrix. This value is given to you in the R output for $\beta_{j0} = 0$. As in linear regression, this test is conditional on all other coefficients being in the model.

Example

Let's revisit the catheterization example from last class.

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 cad.dur_i + \beta_2 gender_i.$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−0.3203	0.0579	−5.53	0.0000
cad.dur	0.0074	0.0008	9.30	0.0000
sex	−0.3913	0.1078	−3.63	0.0003

The column labelled "z value" is the Wald test statistic. What conclusions can we make here?

Confidence intervals for the coefficients and the odds ratios

In the model

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

a $(1-\alpha/2) \times 100\%$ confidence interval for β_j, $j = 1, \ldots, p$, can easily be calculated as

$$\hat{\beta}_j \pm Z_{1-\alpha/2} \hat{se}(\hat{\beta}_j).$$

The $(1-\alpha/2) \times 100\%$ confidence interval for the odds ratio over a one unit change in x_j is

$$\left[\exp(\hat{\beta}_j - Z_{1-\alpha/2} \hat{se}(\hat{\beta}_j)), \exp(\hat{\beta}_j + Z_{1-\alpha/2} \hat{se}(\hat{\beta}_j))\right].$$

What about for a 10 unit change in x_j?

Example

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 cad.dur_i + \beta_2 gender_i.$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−0.3203	0.0579	−5.53	0.0000
cad.dur	0.0074	0.0008	9.30	0.0000
sex	−0.3913	0.1078	−3.63	0.0003

95% CI for a one-unit change in $cad.dur$:

$$[\exp(.0074 - 1.96 \times 0.0008), \exp(.0074 + 1.96 \times 0.0008)]$$

$$= \left[e^{0.0058}, e^{0.0090}\right] = [1.006, 1.009]$$

How can we construct a similar confidence interval for males vs. females?

Testing a single logistic regression coefficient using LRT

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

We want to test $H_0 : \beta_2 = 0$ vs. $H_A : \beta_2 \neq 0$

Our model under the null hypothesis is

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{1i}.$$

What is our LRT statistic?

$$LR = -2\left(l(\hat{\beta}|H_0) - l(\hat{\beta}|H_A)\right)$$

To get both $l(\hat{\beta}|H_0)$ and $l(\hat{\beta}|H_A)$, we need to fit two models: the full model and the model under H_0. Then $l(\hat{\beta}|H_0)$ is the log-likelihood from the model under H_0, and $l(\hat{\beta}|H_A)$ is the log-likelihood from the full model.

If we are testing just one coefficient, $LR \sim \chi^2_1$.

Example

Our full model is

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 cad.dur_i + \beta_2 gender_i.$$

We wish to test $H_0 : \beta_2 = 0$ vs $H_A : \beta_2 \neq 0$. The reduced model is:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 cad.dur_i.$$

Full model:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.3203125  0.0579127  -5.531 3.19e-08
cad.dur      0.0074097  0.0007965   9.303  < 2e-16
sex         -0.3913392  0.1077709  -3.631 0.000282
---

    Null deviance: 3230.5  on 2331  deg of freedom
Residual deviance: 3117.9  on 2329  deg of freedom
AIC: 3123.9
```

Reduced model:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.3965755  0.0541846  -7.319  2.5e-13
cad.dur      0.0074037  0.0007951   9.312  < 2e-16
---

    Null deviance: 3230.5  on 2331  deg of freedom
Residual deviance: 3131.3  on 2330  deg of freedom
AIC: 3135.3
```

We can get the $-2 \log L$ from the output of $summary()$. It is listed as the residual deviance. For the full model, $-2 \log L = 3117.9$. For the reduced model, $-2 \log L = 3131.3$. Therefore,

$$LR = 3131.3 - 3117.9 = 13.4 > 0.53 = \chi^2_{1,1-\alpha}.$$

This is very similar to the Wald test.

Testing groups of variables using the LRT

Suppose instead of testing just variable, we wanted to test a group of variables. This follows naturally from the likelihood ratio test. Let's look at it by example.
Again suppose our full model is

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 cad.dur_i + \beta_2 gender_i,$$

and we test $H_0 : \beta_1 = \beta_2 = 0$ vs. $H_A : \beta_1 \neq 0$ or $\beta_2 \neq 0$.
The $-2\log L$ from the full model is 3117.9. For the reduced model, $-2\log L = 3230.5$. Therefore,

$$LR = 3230.5 - 3117.9 = 112.6 > 5.99 = \chi^2_2.$$

Why 2 degrees of freedom?

Analysis of deviance table

We can get this same information from the analysis of deviance table. We can get this in R, by sending a glm object to the $anova()$ function.
For the model

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 cad.dur_i + \beta_2 gender_i,$$

the (edited) analysis of deviance table is:

```
Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev
NULL                       2331     3230.5
cad.dur    1     99.2      2330     3131.3
sex        1     13.4      2329     3117.9
```

You read the analysis of deviance table similarly to the ANOVA table.

- The 1st row is the intercept only model, and the 5th column is the $-2\log L$ for the intercept only model.
- In the jth row, $j = 2, \ldots, p+1$ row, the 5th column, labeled "Resid. Dev" is the $-2\log L$ for the model with the variables labeling rows $1, \ldots, j$.
  - So, in the 2nd row of the table above, the $-2\log L$ for the model, logit$(\pi_i) = \beta_0 + \beta_1 cad.dur$, is 3131.3.
  - In the third row, the $-2\log L$ for the model, logit$(\pi_i) = \beta_0 + \beta_1 cad.dur + \beta_2 sex_i$, is 3117.9.
- The second column, labeled "Deviance", lists the LRT statistics for the model in the jth row compared to the reduced model in the $j-1$th row.

We can get all the LRT statistics we've already calculated from the analysis of deviance table. (We will look at this in class.)

Can we test that the coefficient for $cad.dur$ is equal to 0 from the analysis of deviance table?