Applied Logistic Regression

Week 3

1. Homework week 2: highlights
2. Confidence interval for β and π
3. Interpretation of coefficients
4. Dichotomous independent variable
5. Computer output: STATA
6. Homework

Stanley Lemeshow, Professor of Biostatistics
College of Public Health, The Ohio State University

THE OHIO STATE UNIVERSITY

It follows from maximum likelihood theory that the estimator of the variance of the estimated coefficients is the inverse of the observed information matrix. The observed information matrix is the matrix of second partial derivatives evaluated at the MLE, $\underline{\hat{\beta}}$.

In the case of a model with *p* covariates the elements in the matrix are:

$$\hat{i}\left(\hat{\beta}\right) = \begin{bmatrix} -\frac{\partial^2 L(\beta)}{\partial \beta_0^2} = \sum_{i=1}^{n} \hat{\pi}(x_i)(1-\hat{\pi}(x_i)) & -\frac{\partial^2 L(\beta)}{\partial \beta_0 \partial \beta_1} = \sum_{i=1}^{n} x_{i1}\hat{\pi}(x_i)(1-\hat{\pi}(x_i)) & \cdots & -\frac{\partial^2 L(\beta)}{\partial \beta_0 \partial \beta_p} = \sum_{i=1}^{n} x_{ip}\hat{\pi}(x_i)(1-\hat{\pi}(x_i)) \\ -\frac{\partial^2 L(\beta)}{\partial \beta_0 \partial \beta_1} = \sum_{i=1}^{n} x_{i1}\hat{\pi}(x_i)(1-\hat{\pi}(x_i)) & -\frac{\partial^2 L(\beta)}{\partial \beta_1^2} = \sum_{i=1}^{n} x_{i1}^2\hat{\pi}(x_i)(1-\hat{\pi}(x_i)) & \cdots & -\frac{\partial^2 L(\beta)}{\beta_1 \partial \beta_p} = \sum_{i=1}^{n} x_{i1}x_{ip}\hat{\pi}(x_i)(1-\pi(x_i)) \\ \vdots & \vdots & \ddots & \\ -\frac{\partial^2 L(\beta)}{\partial \beta_0 \partial \beta_p} = \sum_{i=1}^{n} x_{ip}\hat{\pi}(x_i)(1-\hat{\pi}(x_i)) & -\frac{\partial^2 L(\beta)}{\beta_1 \partial \beta_p} = \sum_{i=1}^{n} x_{i1}x_{ip}\hat{\pi}(x_i)(1-\hat{\pi}(x_i)) & \cdots & -\frac{\partial^2 L(\beta)}{\partial \beta_p^2} = \sum_{i=1}^{n} x_{ip}^2\hat{\pi}(x_i)(1-\hat{\pi}(x_i)) \end{bmatrix}$$

If we let $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}_{n\times(p+1)}$ and

$$\hat{V} = \begin{bmatrix} \hat{\pi}(x_1)(1-\hat{\pi}(x_1)) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\pi}(x_1)(1-\hat{\pi}(x_1)) \end{bmatrix}_{n\times n}$$

$$= \text{diag}\left(\hat{\pi}(x_i)(1-\hat{\pi}(x_i))\right)$$

then $\hat{I}(\hat{\beta}) = \left[X'VX\right]_{(p+1)\times(p+1)}$

It follows that the estimate of the covariance matrix of the estimated coefficients is the inverse of the estimated information matrix

$$\widehat{Var}(\hat{\beta}) = \left[X'\hat{V}X\right]^{-1}$$

This matrix is easily obtained from most packages.

It follows that the confidence interval estimator of a coefficient is

$$\hat{\beta}_j \pm z_{1-\alpha/2}\widehat{SE}\left(\hat{\beta}_j\right)$$

where $\widehat{SE}\left(\hat{\beta}_j\right) = \sqrt{\widehat{Var}\left(\hat{\beta}_j\right)}$

Confidence interval for the logit for a single subject

The estimated logit is $\hat{g}(x) = \sum_{j=0}^{p} \hat{\beta}_j x_j$

and the estimate of its variance is $\widehat{Var}\left(\hat{g}(x)\right) = x'\left[\widehat{Var}\left(\hat{\beta}\right)\right]x$

Note: If we choose x as one of the subjects then we can obtain this from a computer package, otherwise we can use lincom.

Hence the confidence interval for the logit is

$$\hat{g}(x) \pm z_{1-\alpha/2} \widehat{SE}\left(\hat{g}(x)\right)$$

We use this to obtain a confidence interval for the fitted value or estimated logistic probability as follows:

$$\frac{e^{\hat{g}(x) \pm z_{1-\alpha/2} \widehat{SE}\left(\hat{g}(x)\right)}}{1 + e^{\hat{g}(x) \pm z_{1-\alpha/2} \widehat{SE}\left(\hat{g}(x)\right)}}$$

As an example consider a woman who has:

lwt = 100 and race = black.

```
. lincom _b[_cons]+100*_b[lwt]+_b[_Irace_2]*1

 ( 1)  100 lwt + _Irace_2 + _cons = 0

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .3645094   .4901133     0.74   0.457     -.596095    1.325114
------------------------------------------------------------------------------
```

Hence the estimated probability and the end points of its confidence interval are:

$$\hat{\pi}\left(\text{lwt} = 100,\ \text{Race} = \text{black}\right) = \frac{e^{.3645094}}{1+e^{.3645094}} = 0.590$$

and

$$\frac{e^{-.596095}}{1+e^{-.596095}} = 0.355 \quad \text{and} \quad \frac{e^{1.325114}}{1+e^{1.325114}} = 0.790$$

Alternatively, we could use the following command:

```
. adjust lwt=100 _Irace_2=1 _Irace_3=0, pr ci format(%6.3f)

------------------------------------------------------------------------------
     Dependent variable: LOW     Command: logit
Covariates set to value: LWT = 100, _IRACE_2 = 1, _IRACE_3 = 0
------------------------------------------------------------------------------

------------------------------------------
      All |         pr           lb           ub
----------+-------------------------------------
          |      0.590       [0.355       0.790]
------------------------------------------
     Key:  pr          =  Probability
           [lb , ub]   =  [95% Confidence Interval]
```

Question: What is the interpretation of this point estimate and confidence interval?

- we focus on the estimated coefficients for the independent variables in the model.

The coefficients represent the slope or rate of change of a function of the dependent variable per unit of change in the independent variable.

Interpretation is then concerned with

- Determination of the functional relationship between independent and dependent variables
- Definition of unit of change for the independent variable.

What function of *y* yields a linear function of the independent variable?

[this function is known as the "link function"]

- In linear regression it is the identity function (i.e., $y = y$) since, by definition, *y* is linear in the parameters.

$$y = \beta_0 + \beta_1 x$$

- In logistic regression, the link function is

$$g(x) = \ln\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x$$

How do we define the slope?

- In linear regression $\beta_1 = \Delta y / \Delta x$

i.e., it is the difference $y\Big|_{x=x+1} - y\Big|_{x=x}$ for any value of x.

e.g., let $y(x) = \beta_0 + \beta_1 x$

then $y(x+1) = \beta_0 + \beta_1(x+1) = \beta_0 + \beta_1 x + \beta_1$

so

$$y(x+1) - y(x) = \beta_0 + \beta_1 x + \beta_1 - (\beta_0 + \beta_1 x) = \beta_1$$

Hence β_1 = change in y corresponding to a unit change in x.

e.g., Study of adolescent aged males

let y = weight

x = height

Suppose we find $\hat{y} = \hat{\beta}_0 + 5x$

$\Rightarrow \Delta$ in height of 1 inch is associated with a Δ in weight of 5 pounds

for logistic regression

$$g(x) = \ln\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x$$

$$g(x+1) = \ln\left\{\frac{\pi(x+1)}{1-\pi(x+1)}\right\} = \beta_0 + \beta_1(x+1) = \beta_0 + \beta_1 x + \beta_1$$

so

$$g(x+1) - g(x) = (\beta_0 + \beta_1 x + \beta_1) - (\beta_0 + \beta_1 x) = \beta_1$$

i.e., $\beta_1 = \Delta$ in logit for a 1 unit Δ in x.

To interpret the coefficient, β_1, in a logistic regression model, we must place meaning on the difference between 2 logits.

- this will depend on the nature of the independent variable

i.e., is it
- dichotomous
- polychotomous
- continuous

- we now consider each of these in turn.

We assume x is coded as 0,1

$$\Pr\left(y=1\middle|x\right)=\pi\left(x\right)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Then, under the logistic model we have

		Independent Variable (x)	
		$x=1$	$x=0$
Dependent variable (y)	$y=1$	$\pi(1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
	$y=0$	$1-\pi(1)=\frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0)=\frac{1}{1+e^{\beta_0}}$
	Total	1.0	1.0

The odds of the outcome being present among individuals with $x = 1$ is defined as

$$\frac{\Pr\left(y = 1 \middle| x = 1\right)}{\Pr\left(y = 0 \middle| x = 1\right)} = \frac{\pi\left(1\right)}{1 - \pi\left(1\right)}$$

Similarly, the odds of the outcome being present among Individuals with $x = 0$ is

$$\frac{\Pr\left(y = 1 \middle| x = 0\right)}{\Pr\left(y = 0 \middle| x = 0\right)} = \frac{\pi\left(0\right)}{1 - \pi\left(0\right)}$$

RECALL:

We have defined the "logit" as

$$\ln\left[\pi(x)/(1-\pi(x))\right]$$

Hence, the logit ≡ log of the odds or "log odds"

Specifically,

$$g(1) = \ln\left[\pi(1)/(1-\pi(1))\right]$$

$$g(0) = \ln\left[\pi(0)/(1-\pi(0))\right]$$

The "odds ratio" is defined as the ratio of the

odds for $x = 1$ to the odds for $x = 0$

i.e.,

$$OR = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}$$

The log of the odds ratio, termed "log-odds ratio" is

$$\ln(OR) = \ln\left[\frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}\right] = g(1) - g(0) \Leftarrow \text{logit difference}$$

Now:

$$OR = \frac{\left[\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right]\left[\frac{1}{1+e^{\beta_0}}\right]}{\left[\frac{e^{\beta_0}}{1+e^{\beta_0}}\right]\left[\frac{1}{1+e^{\beta_0+\beta_1}}\right]} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Hence, for a dichotomous independent variable

$$OR = e^{\beta_1} \quad **$$

and the logit difference, or log odds ratio is

$$\ln(OR) = \ln\left(e^{\beta_1}\right) = \beta_1$$

**** This is why logistic regression has proven such a powerful tool in epidemiologic research.**

e.g., let us consider again the CHD/AGE data.

We create a new variable as

$$\text{let } x = \begin{cases} 1 \text{ if } AGE \geq 55 \text{ years} \\ 0 \text{ if } AGE < 55 \text{ years} \end{cases}$$

This results in the following 2 ×2 table:

		AGE		
		≥ 55 (1)	< 55 (0)	
CHD	1	21	22	43
	0	6	51	57
		27	73	100

The likelihood for these data is

$$l\left(\underline{\beta}\right) = \pi\left(1\right)^{21}\left[1-\pi\left(1\right)\right]^{6}\pi\left(0\right)^{22}\left[1-\pi\left(0\right)\right]^{51}$$

Use of a logistic regression program provides the following:

VARIABLE	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	$\hat{\beta} / \widehat{SE}(\hat{\beta})$	$\widehat{OR}$
AGE	2.0935	0.5285	3.961	8.1
CONSTANT	-0.8408	0.2551	-3.296	

The value of $\widehat{OR} = e^{2.0935} = 8.1$

NOTE : we could have computed $\widehat{OR}$ directly as:

$$\widehat{OR} = \frac{21 \times 51}{22 \times 6} = 8.11$$

Also : $\hat{\beta}_1 = \ln\left(\widehat{OR}\right) = \ln(8.11) = 2.0935$

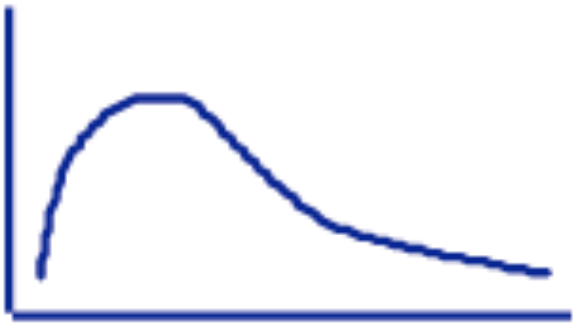
This will be an enormously useful fact when there is more than 1 independent variable where

$$e^{\hat{\beta}_i} = \text{adjusted } OR$$

where the adjustment is for all other independent variables currently in the model.

Confidence Interval Estimation

Because the distribution of $\widehat{OR}$ tends to be skewed to the right, it is clearly not normally distributed. Thus, confidence intervals are usually based on $\hat{\beta}_1$



$\left(\text{since } \ln(\text{OR}) \text{ is more nearly normally distributed}\right)$.

For sufficiently large samples

$$\hat{\beta}_1 \underset{approx}{\sim} N\left(\beta_1, Var\left(\hat{\beta}_1\right)\right)$$

We can estimate the variance in the case when x is dichotomous as

$$\widehat{Var}\left(\hat{\beta}_1\right) = \left[\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right] \quad \text{and} \quad \widehat{SE}\left(\hat{\beta}_1\right) = \sqrt{\widehat{Var}\left(\hat{\beta}_1\right)}$$

cell frequencies in the 2×2 table of $y \times x$

Thus, we compute

$$e^{\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{SE}\left(\hat{\beta}_1\right)}$$

Example:

$$\hat{\beta}_1 = 2.0935 \qquad \widehat{SE}\left(\hat{\beta}_1\right) = \left[\frac{1}{21}+\frac{1}{22}+\frac{1}{6}+\frac{1}{51}\right]^{1/2} = 0.5285$$

and a 95% CI is

$$e^{2.0935 \pm 1.96 \times 0.5285}$$

$$2.88 \leq OR \leq 22.86$$

An alternative method of coding design variables for a Dichotomous risk factor is as follows:

If we use the deviation from means coding with a dichotomous variable then we have

X	D
0	–1
1	1

Can we obtain the odds ratio with this coding?

$$\ln(OR) = g(x=1) - g(x=0)$$
$$= g(D=+1) - g(D=-1)$$
$$= (\beta_0 + \beta_1) - (\beta_0 - \beta_1) = 2\beta_1$$
$$\therefore \widehat{OR} = e^{2\hat{\beta}_1}$$

When we use deviation from means coding the confidence interval is: $e^{2\hat{\beta}_1 \pm 1.96 \times 2 \times \widehat{SE}(\hat{\beta}_1)}$

```
. gen aged=age>=55

. logit chd aged

Iteration 0:   log likelihood = -68.331491
Iteration 1:   log likelihood = -59.020453
Iteration 2:   log likelihood = -58.979594
Iteration 3:   log likelihood = -58.979565

Logit estimates                                   Number of obs   =        100
                                                  LR chi2(1)      =      18.70
                                                  Prob > chi2     =     0.0000
Log likelihood = -58.979565                       Pseudo R2       =     0.1369

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aged |    2.093546   .5285335     3.961   0.000       1.057639    3.129453
       _cons |   -.8407832   .2550733    -3.296   0.001      -1.340718   -.3408487
------------------------------------------------------------------------------

. logistic chd aged

Logit estimates                                   Number of obs   =        100
                                                  LR chi2(1)      =      18.70
                                                  Prob > chi2     =     0.0000
Log likelihood = -58.979565                       Pseudo R2       =     0.1369

------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aged |   8.113636   4.288329     3.961   0.000       2.879566    22.86147
------------------------------------------------------------------------------
```