Applied Logistic Regression

Week 2

1. Homework week 1: highlights
2. Likelihood ratio test
3. Finding a confidence interval for β and π
4. The multiple logistic regression model
5. Fitting the multiple logistic model: low birth weight study I
6. Fitting the multiple logistic model: low birth weight study II
7. Obtaining logistic regression coefficients using discriminant analysis
8. Homework

Stanley Lemeshow, Professor of Biostatistics

College of Public Health, The Ohio State University

THE OHIO STATE UNIVERSITY

Once we have our model, it is good statistical practice to assess the significance of the variables in the model.

- We consider the following question:
  - Does the model that includes the variable tell us more about the outcome (or response) variable than does a model that does not include that variable?

We answer this by comparing the observed values of the response variable to those predicted by each of two models

- the first with and the second without the variable in question.

If the predicted values with the variable in the model are better, or more accurate in some sense, than when the variable is not in the model then we feel that the variable in question is "significant".

Note that we are not considering the question of whether the predicted values are an accurate representation of the observed values in an absolute sense

- this is called GOODNESS-OF-FIT
  - we will discuss this later.

Instead, our question is posed in a relative sense.

In logistic regression comparison of observed to predicted values is based on the log likelihood function.

- To better understand how this comparison is done it is helpful conceptually if we think of an observed value of the response variable as also being a predicted value resulting from a saturated model.

A saturated model contains as many parameters as there are data points.

e.g., fitting a straight line regression to 2 data points.

The comparison of observed to predicted values using the likelihood function is based on the following expression:

$$D = -2\ln\left[\frac{\text{likelihood of the model}}{\text{likelihood of the saturated model}}\right]$$

likelihood ratio

We take $-2\ln(\text{likelihood ratio})$ in order to obtain a quantity whose distribution is known and thus can be used for hypothesis testing purposes.

- This is called the LIKELIHOOD RATIO TEST.

To assess the significance of an independent variable we compare the value of *D* with and without the independent variable in the equation.

- The change in *D* due to including the independent variable in the model is obtained as follows.

$$G = D\left(\text{for model without the variable}\right) - D\left(\text{for the model with the variable}\right)$$

Because the likelihood of the saturated model is common to both values of *D* being differenced to form *G*, its value is

$$G = -2\ln\left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right]$$

Under $H_0 : \beta_1 = 0$, the statistic $G \sim \chi^2(1)$

e.g., In the CHD/AGE data we have $n_1 = 43,\ n_0 = 57$

Thus, $G = -2\ln(\text{likelihood ratio}) = 29.3099$

with 1 df $(p < .001)$

$\Rightarrow$ AGE is a significant variable in predicting CHD

Finally, the WALD Test is obtained by comparing $\hat{\beta}$ to an estimate of its standard error

$$W = \hat{\beta}_1 \Big/ \widehat{SE}\left(\hat{\beta}_1\right)$$

and, under $H_0 : \beta_1 = 0,\ W \sim N(0,1)$.

These quantities are all routinely computed by stat software

e.g., $W = \dfrac{\hat{\beta}_1}{\widehat{SE}\left(\hat{\beta}_1\right)} = \dfrac{.11092}{.02406} = 4.61$ and $p < .001$

Some research has indicated that *W* behaved in an aberrant manner, often failing to reject H_0 when the coefficient was significant.

$\Rightarrow$ recommend using likelihood ratio test.

These methods require the computation of the maximum likelihood estimate for β_1 .

- For a single variable this is not a difficult or costly computational task.
  - However for large data sets with many variables the iterative computation needed to obtain the MLEs can be considerable.

From the STATA output we have the following variance/covariance matrix for the model parameters:

```
. vce
             |      AGE            _cons
-------------+------------------------------
         AGE |    .000579
       _cons |   -.026677         1.28517
```

An approximate 95% C.I. for β_1 is:

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\sqrt{\widehat{Var}\left(\hat{\beta}_1\right)}$$

or $0.1109 \pm 1.96\sqrt{.000579}$

or $0.1109 \pm 1.96 \times 0.02406$

or $\left(0.0638, 0.1580\right)$ ← as given by STATA

Finding a Confidence Interval for $\pi(x)$

First compute C.I. for logit$\left[\pi(x)\right] = \beta_0 + \beta_1(x)$.

- Then transform to get C.I. for $\pi(x)$.

- A large-sample $100(1-\alpha)\%$ C.I. for logit$\left[60\right]$ is:

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \pm z_{1-\alpha/2}\sqrt{\widehat{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right)}$$

where

$$\widehat{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) = \widehat{Var}\left(\hat{\beta}_0\right) + x^2\widehat{Var}\left(\hat{\beta}_1\right) + 2x\widehat{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right)$$

Substituting the endpoints of the C.I. into $\hat{\pi}(x) = \dfrac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}}$

gives an approximate $100(1-\alpha)\%$ C.I. for $\pi(x)$.

example:

For a 60 year old, we estimate the logit of CHD to be

$$g(60) = \hat{\beta}_0 + \hat{\beta}_1(60) = -5.309 + 0.1109 \times 60 = 1.345$$

$$\widehat{Var}\left(\hat{\beta}_0 + \hat{\beta}_1(60)\right) = \widehat{Var}\left(\hat{\beta}_0\right) + 60^2\widehat{Var}\left(\hat{\beta}_1\right) + 2(60)\widehat{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right)$$
$$= 1.285 + 3600(0.000579) + 120(-0.02668) = 0.16784$$

An approximate 95% C.I. for logit $[60]$ is:

$$\hat{\beta}_0 + \hat{\beta}_1(60) \pm z_{1-\alpha/2}\sqrt{\widehat{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 60\right)}$$

$$\text{or} \quad 1.345 \pm 1.96\sqrt{0.16784}$$

$$\text{or} \quad 1.345 \pm 0.80298$$

$$\text{or} \quad (0.54202,\ 2.1480)$$

Substituting the endpoints of the C.I. into

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}$$

for an approximate 95% C.I. for $\pi(60)$ gives:

$$\left(\frac{e^{0.54202}}{1 + e^{0.54202}}, \frac{e^{2.1480}}{1 + e^{2.1480}} \right) = (0.632, 0.895)$$

Notes:

- The usual interpretation of a confidence interval holds here
- The point estimate $\hat{\pi}(60) = 0.793$, is not in the center of this interval

As in the case of linear regression, the strength of a modeling technique lies in its ability to model many variables, some of which may be on different measurement scales.

It is assumed that there is a pre-determined collection of variables that is being examined.

Consider a collection of p independent variables that will be denoted by the vector

$$\underline{x} = \left(x_1, x_2, \ldots, x_p\right)$$

The MLR model states that the conditional probability

$$\Pr\left(y = 1 \middle| x_1, x_2, \ldots, x_p\right) = \Pr\left(y = 1 \middle| \underline{x}\right) = \pi\left(\underline{x}\right)$$

where $$\pi\left(\underline{x}\right) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}$$

or, writing the logit as

$$g\left(\underline{x}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

we have

$$\pi\left(\underline{x}\right) = \frac{e^{g\left(\underline{x}\right)}}{1 + e^{g\left(\underline{x}\right)}}$$

If some of the independent variables are discrete, nominal scaled, variables such as race, treatment group, etc., then it is inappropriate to include them in the model as if they were interval scaled.

In this situation we will use a collection of design (or dummy) variables.

In general, if a variable has k possible outcomes, there will be need for k - 1 design variables.

The notation to indicate design variables is:

Suppose that the j^{th} independent variable, x_j, has k_j levels. The $k_j - 1$ design variables will be denoted as D_{ji} and the coefficients β_{jl}, $l = 1, \ldots, k_j - 1$

Then the logit is

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} + \cdots + \beta_p x_p$$

Assume we have n independent observations $\left(\underline{x}_i, y_i\right), i = 1,\ldots,n$.

As in the univariate case, we seek maximum likelihood estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_p$.

The likelihood function for the MLR model is

$$l\left(\underline{\beta}\right) = \prod_{i=1}^{n} \pi\left(\underline{x}_i\right)^{y_i} \left[1 - \pi\left(\underline{x}_i\right)\right]^{1-y_i}$$

In this case there will be $p + 1$ likelihood equations that are obtained by differentiating the log likelihood function with respect to the $p + 1$ coefficients. The resulting likelihood equations may be expressed as

$$\sum\left(y_i - \pi\left(\underline{x}_i\right)\right) = 0$$

$$\sum x_{ij}\left(y_i - \pi\left(\underline{x}_i\right)\right) = 0 \quad \text{for } j = 1,2,\ldots,p.$$

As with the univariable model, the solution of the likelihood equations requires special purpose software that may be found in many packaged programs.

Let $\hat{\underline{\beta}}$ denote the solution to these equations. Thus the fitted values for the multiple logistic regression model are the $\hat{\pi}\left(\underline{x}_i\right)$, the value of

$$\pi\left(\underline{x}_i\right) = \frac{e^{g\left(\underline{x}\right)}}{1+e^{g\left(\underline{x}\right)}}$$

computed using $\hat{\underline{\beta}}$ and $\underline{x}_i$.

Standard errors for the coefficients $\widehat{SE}\left(\hat{\beta}_j\right)$, $j = 1,\ldots,p$ are given along with the $\hat{\beta}_j$ by any of the good computer packages.

These will be used for estimating adjusted confidence intervals.

Example

Let us consider data from a study designed to identify risk factors associated with giving birth to a baby weighing less than 2500 grams.

Four variables are believed to be important:

AGE (in years)

Weight at last menstrual period (in pounds) [LWT]

RACE White 1

Black 2

Other 3

Number of physician visits during first trimester [FTV]

Dummy Coding Used for RACE:

	Design Variable	
RACE	D_1	D_2
White	0	0
Black	1	0
Other	0	1

The results of fitting the MLR model to these data are:

Variable	$\hat{\beta}_l$	$\widehat{SE}(\hat{\beta}_l)$	$\hat{\beta}_l / \widehat{SE}(\hat{\beta}_l)$
AGE	-0.024	0.034	-.70631
LWT	-0.014	0.0065	-2.1780
RACE (1)	1.004	0.498	2.0165
(2)	0.433	0.362	1.1957
FTV	-0.049	0.167	-.2948
CONSTANT	1.295	1.069	1.2090

log likelihood = -111.286

The logit is calculated as

$$\hat{g}\left(\underline{x}\right) = 1.295 - 0.024 \times AGE - 0.014 \times LWT + 1.004 \times D_1 + 0.433 \times D_2 - 0.049 \times FTV$$

and we can estimate the probability of low birth weight as

$$\Pr\left(y = 1 \middle| \underline{x}\right) = \frac{e^{\hat{g}\left(\underline{x}_i\right)}}{1 + e^{\hat{g}\left(\underline{x}_i\right)}}$$

Now that we have our model, we can test for the overall significance of the p coefficients associated with the independent variables by using the likelihood ratio test.

The test is based on

$$G = -2\ln\left[\frac{\text{likelihood without the } p \text{ independent variables}}{\text{likelihood with the } p \text{ independent variables}}\right]$$

Now the fitted values, $\hat{\pi}$, under the model are based on the vector containing $p+1$ parameters $\underline{\hat{\beta}}$.

Under H_0 that the p "slope" coefficients for the covariates in the model are equal to zero, the distribution of G will be $\chi^2(p)$.

note: Deviance $= D = -2 \times$ log likelihood

$\therefore$ log likelihood $= D/-2$

The computer gives us this value

$G = -2$ log likelihood ratio $= 12.0991$

and, with 5 d.f., $P(\chi^2(5) > 12.0991) = .033$ $\therefore$ significant

The interpretation of this rejected H_0 is analogous to that in multiple linear regression:

i.e., at least one, and perhaps all p coefficients, are different from zero.

Before concluding that any or all of the coefficients are non-zero, we may wish to look at the univariable Wald Test Statistics

$$W_j = \hat{\beta}_j \Big/ \widehat{SE}\left(\hat{\beta}_j\right)$$

In SYSTAT these are labeled "T-RATIO" and, in Stata, "z".

Under H_0 that an individual coefficient = 0, these statistics will follow a standard normal distribution.

Thus the value of these statistics may give us an indication of which variables in the model may or may not be significant.

Using 2 as a critical value ($p \approx .05$) we conclude that weight at last menstrual period (LWT) and possibly race are significant while AGE and first trimester visits (FTV) are not.

If our goal is to find the best fitting model with the fewest parameters, the next logical step is to fit a reduced model containing only those variables thought to be significant and to compare it to the full model containing all the variables.

This was done and the results were:

Variable	$\hat{\beta}_i$	$\widehat{SE}(\hat{\beta}_i)$	$\hat{\beta}_i / \widehat{SE}(\hat{\beta}_i)$
LWT	-0.015	0.006	-2.364
RACE (1)	1.081	0.488	2.215
(2)	0.481	0.357	1.348
CONSTANT	0.806	0.845	0.953

log likelihood = -111.63

The comparison of the two models is obtained by taking

$$G = -2\ln\left[\frac{\text{likelihood with 3 variables}}{\text{likelihood with 5 variables}}\right]$$

$$= -2\left[-111.63 - \left(-111.286\right)\right] = .688$$

which, with 2 df, has a *p*-value of $P\left(\chi^2\left(2\right) > .688\right) = .709$

Since the *p*-value is large, we conclude that the reduced model is as good as the full model.

$\Rightarrow$ there is no advantage to including AGE and FTV in the model.

```
. sort low

. by low: sum age lwt ftv
```

-> low= 0

Variable	Obs	Mean	Std. Dev.	Min	Max
age	130	23.66154	5.584522	14	45
lwt	130	133.3	31.72402	85	250
ftv	130	.8384615	1.069694	0	6

-> low= 1

Variable	Obs	Mean	Std. Dev.	Min	Max
age	59	22.30508	4.511496	14	34
lwt	59	122.1356	26.55928	80	200
ftv	59	.6949153	1.038139	0	4

```
. tab low race,col
```

low	race 1	2	3	Total
0	73	15	42	130
	76.04	57.69	62.69	68.78
1	23	11	25	59
	23.96	42.31	37.31	31.22
Total	96	26	67	189
	100.00	100.00	100.00	100.00

```
. xi:logit low age lwt i.race ftv

i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -111.41656
Iteration 2:   log likelihood = -111.28677
Iteration 3:   log likelihood = -111.28645

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(5)      =      12.10
                                                  Prob > chi2     =     0.0335
Log likelihood = -111.28645                       Pseudo R2       =     0.0516

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.023823   .0337295    -0.71   0.480    -.0899317    .0422857
         lwt |  -.0142446   .0065407    -2.18   0.029    -.0270641   -.0014251
    _Irace_2 |   1.003898   .4978579     2.02   0.044     .0281143    1.979681
    _Irace_3 |   .4331084   .3622397     1.20   0.232    -.2768684    1.143085
         ftv |  -.0493083   .1672386    -0.29   0.768    -.3770899    .2784733
       _cons |   1.295366   1.071439     1.21   0.227    -.8046157    3.395347
------------------------------------------------------------------------------

. estimates store A
```

```
. xi:logit low lwt i.race

i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood =  -111.7491
Iteration 2:   log likelihood = -111.62983
Iteration 3:   log likelihood = -111.62955

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(3)      =      11.41
                                                  Prob > chi2     =     0.0097
Log likelihood = -111.62955                       Pseudo R2       =     0.0486

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |  -.0152231   .0064393    -2.36   0.018    -.0278439   -.0026023
    _Irace_2 |   1.081066   .4880512     2.22   0.027     .1245034    2.037629
    _Irace_3 |   .4806033   .3566733     1.35   0.178    -.2184636     1.17967
       _cons |   .8057535   .8451625     0.95   0.340    -.8507345    2.462241
------------------------------------------------------------------------------

. lrtest A

likelihood-ratio test                             LR chi2(2)  =      0.69
(Assumption: . nested in A)                       Prob > chi2 =    0.7096


Alternative procedure:

. estimates store B

. lrtest A B

likelihood-ratio test                             LR chi2(2)  =      0.69
(Assumption: B nested in A)                       Prob > chi2 =    0.7096
```

Discriminant Function Coefficients

If it is assumed that $\Pr\left\{x_i \middle| y = i\right\}$ has a multivariate normal distribution with mean vector $\underline{\mu}_i$ and covariance matrix Σ then,

$$\underline{\beta} = \sum^{-1}\left(\underline{\mu}_1 - \underline{\mu}_0\right),$$

which can be recognized as the vector of population discriminant function coefficients.

$$\tilde{\beta} = S^{-1}\left(\bar{x}_1 - \bar{x}_0\right)$$

are the discriminant functions estimates.

$\Rightarrow$ under these assumptions, ML estimation of $\underline{\beta}$ follows directly by using normal theory discriminant analysis for two groups.

Important note: This multivariate normality assumption is generally not realistic (e.g., when some of the x_j's are dichotomous) and use of such discriminant function-based estimates can lead to considerable bias.

In the past, researchers used discriminant function estimators of the coefficients since the software was far more economical to use.

Today, such estimates should be regarded with caution and avoided in favor of maximum likelihood estimation whenever possible.