Models In Epidemiology And Biostatistics
Gordon Hilton Fick
Session 3: Linking Stratified Analysis to Logistic Regression

Recall:

That a log(odds) is called a logit

Odds can be any positive number

log(odds) can be any number

If odds = 1, the log(odds) = 0

If odds < 1, then log(odds) < 0

If odds > 1, then log(odds) > 0

Odds Ratios in the Analysis of Case Control Studies

E – exposure D – disease

Disease and Exposure Coding: (same name may be used for both labels and codes)

0= absence : label has a “bar” $\bar{E}$ or $\bar{D}$

1= presence : label has no “bar” E or D

Strata: (can be more than 2 levels) 0, 1, 2, ...

Maybe more than one stratum variables:

Age: Young (A=0) Old (A=1)

Gender: Male (G=0) Female (G=1)

2 Probabilities : 2 Odds

	Exposed	Unexposed
Cases	p_1	$1-p_1$
Controls	p_0	$1-p_0$

$p_1 = P(E \mid D) =$ probability of exposure given case status

$p_0 = P(E \mid \bar{D}) =$ probability of exposure given control status

$\frac{p_1}{1-p_1}$ = the odds of exposure given case status

$\frac{p_0}{1-p_0}$ = the odds of exposure given control status

A logit (log (odds)) in an equation
[Model 1]

p = a conditional probability of exposure

$p/(1-p)$ = a conditional odds of exposure

This odds is conditional on disease status

We write: $\log(p/(1-p)) = \beta_0 + \beta_1 D$

So that: $\log(p_0/(1-p_0)) = \beta_0$

And: $\log(p_1/(1-p_1)) = \beta_0 + \beta_1$

We call β_0 and β_1 the regression coefficients.

Taking the difference

$$\beta_1 = \log(p_1/(1-p_1)) - \log(p_0/(1-p_0))$$

β_1 is the difference between 2 log odds

The log of the odds of exposure for those with disease minus the log of the odds of exposure for those without disease

The odds ratio appears

Lets take the exponent of β_1

$$\beta_1 = \log(p_1/(1-p_1)) - \log(p_0/(1-p_0))$$

$$\beta_1 = \log \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

$$e^{\beta_1} = \exp(\beta_1) = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

This is the odds ratio: the odds of exposure for those with disease divided by the odds of exposure for those without disease

Now let us consider smoking status

We now have 4 probabilities (4 odds)

...and 2 odds ratios:

an OR for the smokers (strata = S = 1)

an OR for the non-smokers (strata = S = 0)

	S=1		S=0	
	E=1	E=0	E=1	E=0
D=1	p_{11}	$1- p_{11}$	p_{10}	$1- p_{10}$
D=0	p_{01}	$1- p_{01}$	p_{00}	$1- p_{00}$
	OR_1		OR_0	

Modeling the conditional odds

Consider: $\log(p/(1-p)) = \beta_0 + \beta_1 D + \beta_2 S + \beta_3 DS$

So that:

Nonsmoking controls $\log(p_{00}/(1-p_{00})) = \beta_0$

Nonsmoking cases $\log(p_{10}/(1-p_{10})) = \beta_0 + \beta_1$

Smoking controls $\log(p_{01}/(1-p_{01})) = \beta_0 + \beta_2$

Smoking cases $\log(p_{11}/(1-p_{11})) = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Differences and Ratios

So for nonsmokers:

$$\log(p_{10}/(1-p_{10}))-\log(p_{00}/(1-p_{00}))=(\beta_0+\beta_1)-\beta_0=\beta_1$$

$$\log\frac{p_{10}/(1-p_{10})}{p_{00}/(1-p_{00})}=\beta_1$$

So for smokers:

$$\log(p_{11}/(1-p_{11}))-\log(p_{01}/(1-p_{01}))$$
$$=(\beta_0+\beta_1+\beta_2+\beta_3)-(\beta_0+\beta_2)=\beta_1+\beta_3$$

$$\log\frac{p_{11}/(1-p_{11})}{p_{01}/(1-p_{01})}=\beta_1+\beta_3$$

Stratified analysis via logistic regression [Model 2]

Now consider:

$$\log(p/(1-p)) = \beta_0 + \beta_1 D + \beta_2 S + \beta_3 DS$$

For S=0 $\log(p/(1-p)) = \beta_0 + \beta_1 D$

For S=1 $\log(p/(1-p)) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) D$

For S=0, we see that β_1 is the log odds ratio as before.

Now notice that for S=1, $\beta_1 + \beta_3$ is the log odds ratio.

Lets check this:

Checking:

For S=1, we have:

$$\log(p/(1-p)) = \beta_0 + \beta_2 + (\beta_1 + \beta_3)\mathrm{D}$$

So for D=0, we have:

$$\log(p/(1-p)) = \beta_0 + \beta_2$$

So for D=1, we have:

$$\log(p/(1-p)) = \beta_0 + \beta_2 + \beta_1 + \beta_3$$

So $\beta_1 + \beta_3$ is the difference.

Ratio of Odds Ratios

So the OR for strata 0 (OR_0) is $\exp(\beta_1)$

the OR for strata 1 (OR_1) is $\exp(\beta_1+\beta_3)$

$\exp(\beta_3)$ is the ratio of the 2 ORs : $\mathrm{OR}_1/\mathrm{OR}_0$

β_3 is the difference between the 2 log ORs

Ratio of Odds Ratios

We are again taking differences

$$\log \frac{p_{11}/(1-p_{11})}{p_{01}/(1-p_{01})} - \log \frac{p_{10}/(1-p_{10})}{p_{00}/(1-p_{00})} = (\beta_1 + \beta_3) - \beta_1 = \beta_3$$

We are again taking ratios

$$\left(\frac{p_{11}/(1-p_{11})}{p_{01}/(1-p_{01})}\right) / \left(\frac{p_{10}/(1-p_{10})}{p_{00}/(1-p_{00})}\right) = \exp(\beta_3)$$

Assessing Modification

What if $\beta_3 = 0\,?$

Then $\mathrm{OR}_0 = \mathrm{OR}_1$

So that smoking is not a modifier.

What if $\beta_3 \neq 0\,?$

Then $\mathrm{OR}_0 \neq \mathrm{OR}_1$

So smoking is a modifier

If smoking is a modifier...

...then we need to report that the stratum specific odds ratios are different. Both odds ratios need interpretation with separate estimates [confidence intervals and tests]

...it would misleading to attempt to report a single odds ratio

...remember the DeLury quote about “statistical decency”

...any attempt at finding a single odds ratio is inevitably a “combination” of 2 different odds ratios

If smoking is not a modifier...

...then a “combination” of the two odds ratios may make sense

...and we may be able to report a “crude” odds ratio as well

This part of the analysis process is the same as stratified analysis only now we can compare corresponding coefficients directly or we can compare the exponents of these coefficients

Lets assess a “simpler” model [Model 3]

Lets try: $\log(p/(1-p)) = \beta_0 + \beta_1 \mathrm{D} + \beta_2 \mathrm{S}$

Now specialize this equation ->

If S=0, $\log(p/(1-p)) = \beta_0 + \beta_1 \mathrm{D}$

If S=1, $\log(p/(1-p)) = \beta_0 + \beta_2 + \beta_1 \mathrm{D}$

For S=0, we see that β_1 is the log odds ratio as before.

Now notice that for S=1, again β_1 is the log odds ratio.

Lets check this:

Checking:

For S=1, we have:

$$\log(p/(1-p)) = \beta_0 + \beta_2 + \beta_1 \mathrm{D}$$

So for D=0, we have:

$$\log(p/(1-p)) = \beta_0 + \beta_2$$

So for D=1, we have:

$$\log(p/(1-p)) = \beta_0 + \beta_2 + \beta_1$$

So β_1 is again the difference.

Assumed Common Odds Ratio

We can then see that β_1 is the log odds ratio for both strata.

Now $\exp(\beta_1)$ is the "assumed common" OR

This model "forces" the stratum specific odds ratios to be the same.

Now you compare this "adjusted" OR with the "crude" OR from model 1 just like with a stratified analysis.

Notice that..

..the meaning of the coefficients changes as soon we change the model by adding terms or deleting terms.

For example, β_1 in model 1: $\log(p/(1-p)) = \beta_0 + \beta_1 D$

means something very different from β_1 in model 2:

$$\log(p/(1-p)) = \beta_0 + \beta_1 D + \beta_2 S + \beta_3 DS$$

The meaning of the coefficients must be reconsidered every time we change the model.

Interaction

With models like:

$$\log(p/(1-p)) = \beta_0 + \beta_1 D + \beta_2 S + \beta_3 DS$$

we see terms involving a product of two variables (here D times S)

We saw that including such a term enabled a specific set of interpretations for the coefficients

Some authors call $\beta_3 DS$ an 'interaction' term.

Interaction terms have several uses

We saw that in Model 2:

$$\log(p/(1-p)) = \beta_0 + \beta_1 D + \beta_2 S + \beta_3 DS$$

including such a term enabled us to assess modification.

In future sessions, we will see other uses of interaction terms that include more elaborate assessments of confounding and the consideration of one than one exposure or disease status in a model.

In this way, referring to 'interaction' is generic and we will try to be more specific in our interpretations and descriptions.

Interpreting all the coefficients

In principle, all the coefficients in a model can be interpreted.

Sometimes, the inclusion of certain terms in a model is to enable an important interpretation for a key coefficient.

For example, with Model 2, we can find useful interpretations for β_0 , β_1 and β_3

It is instructive to provide an interpretation for β_2 and to understand why it is included in this model. Try it.

R.A. Fisher

Background: The Likelihood Function

R.A. Fisher wrote the first published paper on likelihood in 1922

Fisher R.A. (1922) *On the mathematical foundations of theoretical statistics.* Phil. Trans., A, 222: 309-368.

The best introductions to this subject are:

Kalbfleisch J.G. (1985) *Probability and Statistical Inference Vol 2.* Springer Verlag

Fraser D.A.S. (1976) *Probability and Statistics: Theory and Applications.* Duxbury (GHF wrote the solutions manual!)

The Likelihood is a function that is proportional to the probability of the observed.

How Likelihood enables the estimation of the unknowns

An observed likelihood function (of the unknowns) is determined from the data and the model.

This function is maximized. The maximum value (M) of this function and the values of the unknowns that give this maximum (the maximum likelihood estimates (mle's)) are computed.

Other characteristics are determined that give us standard errors for the estimates and confidence intervals. These other characteristics are based on an approximating parabolic curve that is supposedly “close” to the actual log likelihood curve.

A brief introduction to likelihood by example:

Consider the estimation of a prevalence: p or equivalently a log odds: log(p/(1-p)) using the binomial probability function [undergrad stats]

$$\mathrm{L} = c\,\mathrm{p}^{y}(1\text{-}\mathrm{p})^{n-y} \text{ or } \log(\mathrm{L}) = y\log(\mathrm{p}) + (n-y)\log(1\text{-}\mathrm{p}) + a$$

[calculus needed] There is an approximating parabola to log(L):

$$\log(\mathrm{L}) \approx \mathrm{M} - 0.5\left(\frac{\text{logodds - mle}}{\text{se(mle)}}\right)^{2}$$

Suppose we have a sample of size n=5 and there were y=3 with disease
Then, we can coerce Stata to provide the “analysis”

```
. list suc cons
     +------------+
     | suc   cons |
     |------------|
  1. |   1      1 |
  2. |   1      1 |
  3. |   1      1 |
  4. |   0      1 |
  5. |   0      1 |
     +------------+

. logit suc cons,nocons
Iteration 0:   log likelihood = -3.4657359
Iteration 1:   log likelihood = -3.3650763
Iteration 2:   log likelihood = -3.3650583

Logistic regression                               Number of obs   =          5
                                                  LR chi2(1)      =          .
Log likelihood = -3.3650583                       Prob > chi2     =          .

------------------------------------------------------------------------------
         suc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cons |   .4054651   .9128707     0.44   0.657    -1.383729    2.194659
------------------------------------------------------------------------------
```

The equations look like:

log(L) = 5*log(p) + 2*log(1-p)

$$\log(L) \approx -3.365 - 0.5\left(\frac{\text{logodds} - 0.4055}{0.9129}\right)^2$$

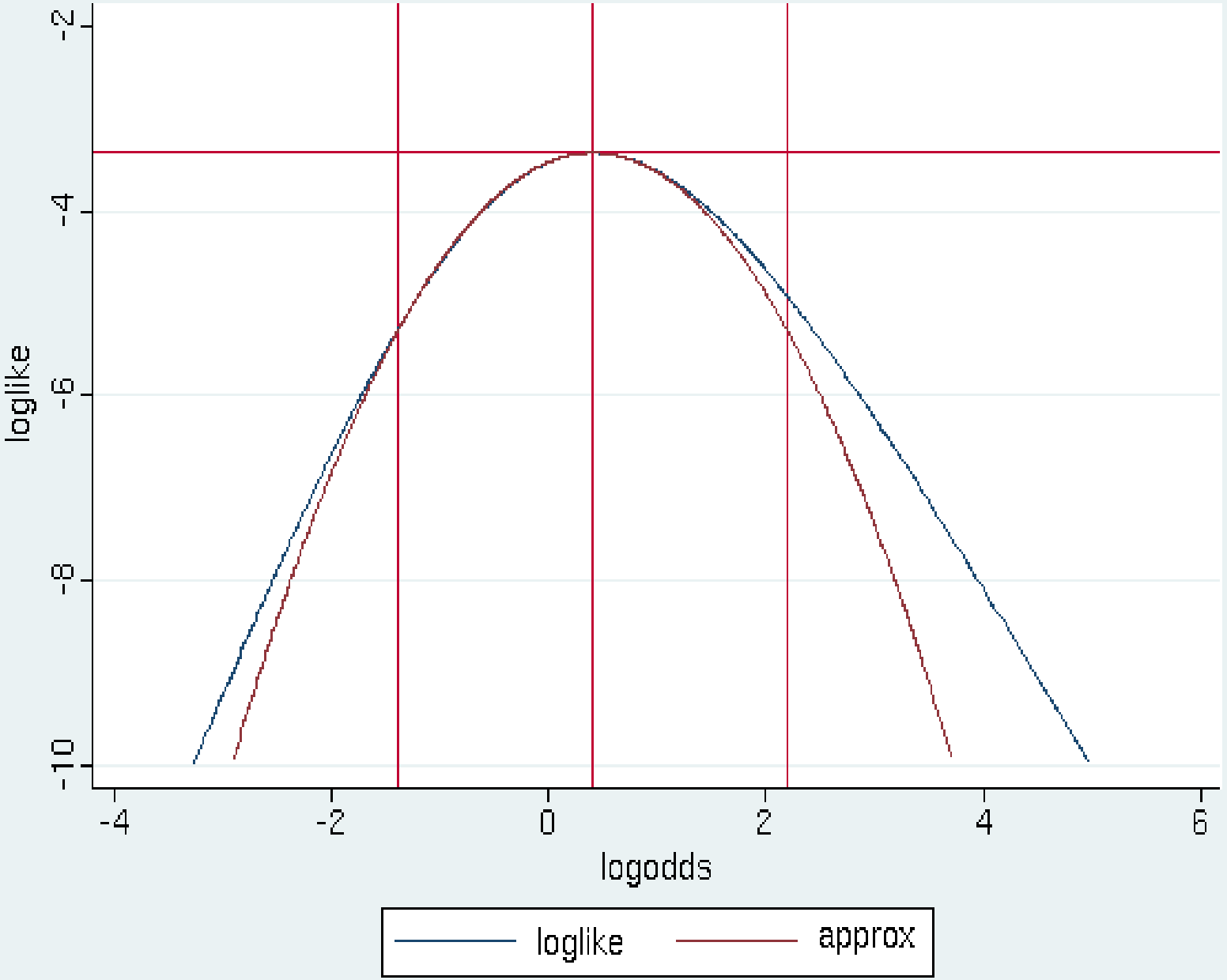
Notice that the mle is 0.4055 = log(1.5) = log(3/2)

and the maximum of the log likelihood is -3.365 = 5*log(0.6) + 2*log(0.4)

The formula for se(mle) turns out [calculus needed] to be

= sqrt(1/(np(1-p))) = sqrt(5/6) = 0.9129

A picture “helps” to put all the pieces together


-2
-4
-6
-8
-10
loglike
-4
-2
0
2
4
6
logodds
loglike
approx

The logistic model and the fit

Model:

$$\log\left(\frac{p}{1-p}\right)=\sum_{i=0}^{k}\beta_i x_i=\beta_0 x_0+\beta_1 x_1+\ldots+\beta_k x_k$$

Fit:

$$\log\left(\frac{\hat{\mathrm{p}}}{1-\hat{\mathrm{p}}}\right)=\sum_{i=0}^{k}\hat{\beta}_i x_i=\sum_{i=0}^{k} b_i x_i=b_0 x_0+b_1 x_1 \ldots b_k x_k$$

$$\hat{\beta}_i=b_i=[\text{maximum likelihood}]\ \text{estimate of}\ \beta_i$$

or: $$\hat{\mathrm{p}}=\frac{1}{1+e^{-\sum_{i=0}^{k} b_i x_i}}$$

More on interpretation

Interpreting the fit: Given any sensible set of values for the x's, the fit gives an estimate of the [conditional] log odds

Take the exponent to get an estimate of the [conditional] odds or “solve” the equation to get an estimate of a [conditional] probability

Interpreting a coefficient from a fit: Add “an estimate of” to the description of the model coefficient

Estimation and confidence intervals via maximum likelihood

The odds ratio estimates in models 1 and 2 yield the familiar estimates from stratified analysis.

Model 3 does not give the exact same odds ratio estimate as the Mantel-Haentzel estimate. If they are meaningfully different, you will need to understand “why”

Confidence intervals can be different. The formulae used are not quite the same.

The Likelihood Ratio Test

Testing nested hypothesis can be accomplished using a Likelihood Ratio (LR) test.

This test is based on the fitting of 2 models. The likelihood function is maximized with each fit. The maximum values are compared in a ratio (actually a difference on a logarithmic scale)

The distribution theory is based on an approximating χ^2 distribution. The degrees of freedom is the difference in the number of unknowns.

The Wald test

The “familiar” z-test based on the ratio of estimate to standard error is called the Wald test in this setting.

This test is dependent on the approximating parabolic curve and the suitability of the computed se(mle). As such, it is typically inferior to the LR test. Both the LR test and the Wald test are “approximate”.

Approximate confidence intervals can be obtained from the estimate and standard error.