Applied Logistic Regression

Week 6

1. Homework week 5: highlights
2. Stratified analysis vs. Logistic regression
3. Confounding and effect modification
   - Mantel-Haenszel estimator and logit based estimator
   - Use of logistic regression to obtain adjusted odds ratios and confidence intervals
4. Assessing the fit of the logistic regression model
5. Introduction to goodness of fit
6. Homework

Stanley Lemeshow, Professor of Biostatistics

College of Public Health, The Ohio State University

THE OHIO STATE UNIVERSITY

Typically, epidemiologists have dealt with the issues of confounding and interaction by performing stratified analyses of 2×2 contingency tables.

- The objective of these analyses is to determine whether or not the odds ratios are consistent (or homogeneous) over the strata.

- If there is consistency, then a stratified odds ratio estimator such as the "Mantel-Haenszel estimator" or "weighted logit-based estimator" will be computed.

- As we will now see, these same analyses may be performed quite simply using our logistic regression techniques.

Example

We are interested in an analysis of the risk factor smoking on low birth weight.

The data for the 189 women in the study are as follows:

		SMOKE 0	SMOKE 1	
LOW	0	86	44	130
	1	29	30	59
		115	74	189

The crude odds ratio is computed as

$$\widehat{OR} = \frac{86 \times 30}{44 \times 29} = 2.02$$

Now, it is believed that RACE may be a confounder or effect modifier.

- To examine this, we stratify our sample on RACE and form the 2×2 table of SMOKE vs. LOW within each RACE group.

WHITE

LOW \ SMOKE	0	1	
0	40	33	73
1	4	19	23
	44	52	96

$$\widehat{OR}_W = \frac{40 \times 19}{33 \times 4} = 5.76$$

BLACK

LOW \ SMOKE	0	1	
0	11	4	15
1	5	6	11
	16	10	26

$$\widehat{OR}_B = \frac{11 \times 6}{4 \times 5} = 3.3$$

OTHER

LOW \ SMOKE	0	1	
0	35	7	42
1	20	5	25
	55	12	67

$$\widehat{OR}_O = \frac{35 \times 5}{7 \times 20} = 1.25$$

While all these $\widehat{OR}_i$ are in the same direction $\left(\text{i.e., } >1\right)$, they vary considerably.

Before we compute a summary odds ratio we must assume that the odds ratio is constant over strata.

- We may assess the validity of this assumption both visually or by performing a statistical test that compares the stratum-specific estimates to an overall estimate.
- The overall estimate is computed under the assumption that the odds ratios are, in fact, constant over strata.

There are two overall estimates we will consider here:

(1) Mantel-Haenszel estimator

(2) logit-based estimator

(1) Mantel-Haenszel estimator

- this is a weighted average of the stratum-specific odds ratios

$$\widehat{OR}_{MH} = \frac{\sum a_i \times d_i / N_i}{\sum b_i \times c_i / N_i}$$

evaluating this expression, we have

i	a_i	b_i	c_i	d_i	N_i	$a_i d_i / N_i$	$b_i c_i / N_i$
1	40	33	4	19	96	7.9	1.375
2	11	4	5	6	26	2.54	0.769
3	35	7	20	5	67	2.61	2.089
						13.05	4.233

$$\widehat{OR}_{MH} = 13.05 / 4.233 = 3.08$$

recall: crude $\widehat{OR} = 2.02$

(2) logit-based estimator

This is obtained as a weighted average of the stratum-specific odds ratios.

- The weights are the inverse of the variance of the log-odds ratios.

Specifically,

$$\widehat{OR}_{\mathrm{L}} = e^{\sum w_i \ln\left(\widehat{OR}_i\right) \Big/ \sum w_i}$$

where

$$w_i = 1 \Big/ \widehat{Var}\left[\ln\left(\widehat{OR}_i\right)\right]$$

For our data:

$$\frac{1}{a_i}+\frac{1}{b_i}+\frac{1}{c_i}+\frac{1}{d_i} \qquad \frac{1}{\widehat{Var}\left[\ln\left(\widehat{OR}_i\right)\right]}$$

Stratum	a_i	b_i	c_i	d_i	$\widehat{OR}_i$	$\ln\left(\widehat{OR}_i\right)$	$\widehat{Var}\left[\ln\left(\widehat{OR}_i\right)\right]$	w_i	$w_i\ln\left(\widehat{OR}_i\right)$
White	40	33	4	19	5.76	1.751	0.358	2.794	4.891
Black	11	4	5	6	3.30	1.194	0.708	1.413	1.687
Other	35	7	20	5	1.25	0.223	0.421	2.373	0.529
								6.580	7.107

$$\widehat{OR}_{\text{L}} = e^{7.107/6.580} = 2.95$$

recall: $\widehat{OR}_{\text{MH}} = 3.09, \quad \widehat{OR}_{\text{crude}} = 2.02$

In general, $\widehat{OR}_{\text{L}}$ and $\widehat{OR}_{\text{MH}}$ will be similar when the data are not too sparse within the strata.

- One considerable advantage of the MH estimator is that it may be computed even when some of the cell entries are 0.

A test for homogeneity of the odds ratios across strata is based on a weighted sum of squared deviations of the stratum-specific log-odds from their weighted mean.

Specifically,

$$\chi^2_{\mathrm{H}} = \Sigma\left\{w_i\left[\ln\left(\widehat{OR}_i\right) - \ln\left(\widehat{OR}_L\right)\right]^2\right\}$$

and, under $H_0 : \widehat{OR}_i$ are constant, $\chi^2_{\mathrm{H}} \sim \chi^2(\#\,\text{strata} - 1)$

In our example,

Stratum	$\ln\left(\widehat{OR}_i\right)$	$\ln\left(\widehat{OR}_L\right)$ $\ln(2.95)$	$\left[\ln\left(\widehat{OR}_i\right) - \ln\left(\widehat{OR}_L\right)\right]^2$	w_i	$w_i\left[\ln\left(\widehat{OR}_i\right) - \ln\left(\widehat{OR}_L\right)\right]^2$
White	1.751	1.082	0.448	2.794	1.250
Black	1.194	1.082	0.013	1.413	0.018
Other	0.223	1.082	0.738	2.375	1.752
					3.021

Hence, $\chi^2_{\text{H}} = 3.02$

and, comparing this to $\chi^2(2)$, we have a p – value of 0.221

That is, in spite of the apparent differences in the odds ratios, this test suggests that they are within sampling variability of each other.

SAS-PC computes the Breslow-Day test for homogeneity.

- This test computes

$$\chi^2_{\text{BD}} = \sum \frac{\left(a_i - \hat{e}_i\right)^2}{\hat{v}_i} \qquad \text{where}$$

$\hat{e}_i$ is the estimated frequency in the 1,1 cell if the odds ratio were constant.

$\hat{v}_i$ is an estimate of the variance of a_i under the assumption of a constant odds ratio.

In this example, $\chi^2_{\text{BD}} = 3.13$, $p = .21$ $\left[\text{based on } \chi^2(2)\right]$

Now let us perform this analysis through logistic regression.

- **To do this, we must fit 3 models:**

dummy variables

MODEL 1: $g(\text{SMOKE}) = \beta_0 + \beta_1(\text{SMOKE})$

MODEL 2: $g(\text{SMOKE, RACE}) = \beta_0 + \beta_1(\text{SMOKE}) + \beta_2(\text{RACE1}) + \beta_3(\text{RACE2})$

MODEL 3: $g(\text{SMOKE, RACE, S} \times \text{R}) = \beta_0 + \beta_1(\text{SMOKE}) + \beta_2(\text{RACE1}) + \beta_3(\text{RACE2}) + \beta_4(\text{SMOKE} \times \text{RACE1}) + \beta_5(\text{SMOKE} \times \text{RACE2})$

Our analysis will focus on the coefficients $\hat{\beta}_1$ under the 3 models.

Model	$\hat{\beta}_1$	$\widehat{OR}$	log-likelihood	G	df	p
1	0.704 [1]	2.02	-114.90			
2	1.116 [2]	3.05	-109.99	9.82	2	0.007
3	1.751		-108.41	3.16 [3]	2	0.206

1: The crude odds ratio is $e^{\hat{\beta}_1} = e^{.704} = 2.02,$ which is the same as we obtained earlier for the overall 2×2 table.

2: Adjusting for RACE, the stratified estimate is $\widehat{OR} = e^{1.116} = 3.05.$ This value is the maximum likelihood estimator of the estimated odds ratio and is similar in value to $\widehat{OR}_{MH} = 3.086$ and $\widehat{OR}_{L} = 2.95.$

The change in the estimate of the odds ratio from 2.02 to 3.05 suggests considerable confounding due to RACE.

3: The likelihood ratio test of model 3 to model 2 provides an assessment of the homogeneity of the odds ratios across the strata.

- Here, $G = 3.16$ and is compared to the $\chi^2(2)$ since 2 interaction terms were added to the model.

This G plays the same role and is similar in quantity to $\chi^2_H = 3.02$ and $\chi^2_{BD} = 3.13$.

Hence, logistic regression provides a fast and effective way of obtaining a stratified odds ratio estimate and to assessing the assumption of homogeneity across strata.

```
. logit low smoke
Logit estimates                                   Number of obs   =        189
                                                  LR chi2(1)      =       4.87
                                                  Prob > chi2     =     0.0274
Log likelihood =  -114.9023                       Pseudo R2       =     0.0207

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   .7040592   .3196386      2.203   0.028       .0775791    1.330539
   _cons |  -1.087051   .2147299     -5.062   0.000      -1.507914   -.6661886
------------------------------------------------------------------------------

. xi:logit low smoke i.race
i.race                 Irace_1-3    (naturally coded; Irace_1 omitted)
Logit estimates                                   Number of obs   =        189
                                                  LR chi2(3)      =      14.70
                                                  Prob > chi2     =     0.0021
Log likelihood = -109.98736                       Pseudo R2       =     0.0626

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   1.116004   .3692258      3.023   0.003       .3923346    1.839673
 Irace_2 |   1.084088   .4899845      2.212   0.027       .1237362     2.04444
 Irace_3 |   1.108563   .4003054      2.769   0.006       .3239787    1.893147
   _cons |  -1.840539   .3528633     -5.216   0.000      -2.532138   -1.148939
------------------------------------------------------------------------------
```

```
. xi:logit low i.race*smoke
i.race              Irace_1-3     (naturally coded; Irace_1 omitted)
i.race*smoke        IrXsmo_#      (coded as above)

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(5)      =      17.85
                                                  Prob > chi2     =     0.0031
Log likelihood = -108.40889                       Pseudo R2       =     0.0761

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Irace_2 |   1.514128   .7522689     2.013   0.044       .0397077    2.988548
 Irace_3 |   1.742969   .5946183     2.931   0.003       .5775389      2.9084
   smoke |   1.750517   .5982759     2.926   0.003       .5779173    2.923116
IrXsmo_2 |   -.556594   1.032235    -0.539   0.590      -2.579738     1.46655
IrXsmo_3 |  -1.527373   .8828152    -1.730   0.084      -3.257659     .202913
   _cons |  -2.302585   .5244039    -4.391   0.000      -3.330398   -1.274772
------------------------------------------------------------------------------

 lrtest /for interaction model vs. main  effects model/
Logit:  likelihood-ratio test                         chi2(2)     =       3.16
                                                      Prob > chi2 =     0.2063
```

```
. mhodds low smoke, by(race)

Maximum likelihood estimate of the odds ratio
Comparing smoke==1 vs. smoke==0
by race

-------------------------------------------------------------------------------
     race | Odds Ratio        chi2(1)         P>chi2       [95% Conf. Interval]
----------+--------------------------------------------------------------------
        1 |   5.757576           9.75         0.0018        1.65890    19.98288
        2 |   3.300000           2.00         0.1569        0.57171    19.04810
        3 |   1.250000           0.12         0.7327        0.34647     4.50975
-------------------------------------------------------------------------------

    Mantel-Haenszel estimate controlling for race
    ----------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ----------------------------------------------------------------
       3.086381       9.41        0.0022         1.445341   6.590657
    ----------------------------------------------------------------

Test of homogeneity of ORs (approx): chi2(2)   =    3.04
                                     Pr>chi2   =  0.2186
```

```
. cc LOW SMOKE,bd by(RACE)

           RACE |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               1 |   5.757576      1.657574    25.1388        1.375 (exact)
               2 |        3.3      .4865385   23.45437     .7692308 (exact)
               3 |       1.25       .273495   5.278229     2.089552 (exact)
-----------------+-------------------------------------------------
           Crude |   2.021944      1.029092   3.965864              (exact)
    M-H combined |   3.086381       1.49074   6.389949
-------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(2) =     3.03  Pr>chi2 = 0.2197
Test of homogeneity (B-D)      chi2(2) =     3.13  Pr>chi2 = 0.2095

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      9.41
                                                Pr>chi2 =    0.0022
```

After estimates of the coefficients have been obtained, an estimate of the probability of development of the outcome may be calculated for each individual in the study.

Now we would like to know how effective the model we have is in describing the outcome variable.

- This will be accomplished by comparing observed outcomes to predicted outcomes based on the logistic model.

This comparison is referred to as assessing "Goodness-of-Fit".

What does it mean to say that the model "fits"?

Let us denote the observed outcomes as $y_1, y_2, \ldots, y_n$

and

let us denote the values predicted by the model as $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$

We will conclude that the model fits if

- The summary measures of the distance between $\underline{y}$ and $\underline{\hat{y}}$

are small

and if

- The contribution of each pair $(y_i, \hat{y}_i)$, $i = 1, \ldots, n$ to these summary measures is unsystematic and is small relative to the error structure of the model

Let us concentrate on the first point, computation and evaluation of overall measures of fit.