Topics in Biostatistics
Categorical Data Analysis and Logistic Regression

B. Rosner, 5/02/17
Outline

1. Contingency table analysis
   (a) Analysis of 2 X 2 tables
   (b) Analysis of 2 x k tables

2. Logistic regression analysis
   (a) Estimation of odds ratios
   (b) Indicator variable methods
   (c) Modelling of interaction effects
Example. A case-control study of ovarian cancer is performed.

50 ovarian cancer cases and 100 controls are identified and are asked about their reproductive history.

Goal is to compare the probability of early menarche (age < 11 years) \((E)\) for cases vs. controls.
Ovarian Cancer Case-Control Study Data

1. In this study, there were 50 ovarian cancer cases and 100 controls all of whom were age 50-54.

2. Ten of the ovarian cancer cases and 12 of the controls had an age at menarche (age when periods begin) of < 11 years old.

3. Is there a significant association between early menarche and ovarian cancer?
Contingency Table \((2 \times 2)\)

1. A contingency table is a method for displaying categorical data where

(a) the categories of one of the variables (say Variable 1) are given in the rows of the table

(b) the categories of Variable 2 are given in the columns of the table

and

(c) a count is provided of the number of subjects corresponding to the combination of each row and column in the table.
### Display of Ovarian Cancer Data in a $2 \times 2$ Contingency Table

1. The ovarian cancer data is displayed in the form of a $2 \times 2$ contingency table as follows:

<table>
<thead>
<tr>
<th>age at menarche</th>
<th>&lt; 11</th>
<th>11 +</th>
<th>cases</th>
<th>controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>cases</td>
<td>10</td>
<td>40</td>
<td>50</td>
<td>12</td>
</tr>
<tr>
<td>controls</td>
<td>12</td>
<td>88</td>
<td>100</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>128</td>
<td>150</td>
<td>150</td>
</tr>
</tbody>
</table>

2. This is referred to as the **observed contingency table**.

3. The row (50,100) and column (22, 128) totals are referred to as the row and column margins.

4. The overall number of subjects (150) is referred to as the grand total.
Analysis of Contingency Table Data

1. We now must obtain an expected contingency table under the null hypothesis that row and column classifications are independent.
Analysis of Contingency Table Data (cont.)

2. We wish to test the hypothesis:

\[ H_0 : p_{ij} = a_i b_j \ vs. \ H_1 : p_{ij} \neq a_i b_j, \ where \]
\[ p_{ij} = \Pr(\text{subject is in the } i\text{th row and } j\text{th column}) \]
\[ a_i = \Pr(\text{subject is in the } i\text{th row}), \text{ estimated by } R_i / N, \]
\[ b_j = \Pr(\text{subject is in the } j\text{th column}), \text{ estimated by } C_j / N, \]
where \( R_i = \text{ith row total}, \ C_j = \text{jth column total}, \)
\( N = \text{grand total}. \)
Analysis of Contingency Table Data (cont.)

• 3. Under $H_0$, 

\[
(a) \hat{p}_{ij} = \frac{R_i}{N} \times \frac{C_j}{N},
\]

\[
(b) E_{ij} = \text{expected count in the (i,j) cell}
\]

\[
= N\hat{p}_{ij} = R_i \times C_j / N.
\]
## Calculation of Expected Table for Ovarian Cancer Data

<table>
<thead>
<tr>
<th>Observed Table</th>
<th>Expected Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>age at menarche</td>
<td>age at menarche</td>
</tr>
<tr>
<td>&lt; 11</td>
<td>&lt; 11</td>
</tr>
<tr>
<td>11 +</td>
<td>11 +</td>
</tr>
<tr>
<td>cases</td>
<td>cases</td>
</tr>
<tr>
<td>10</td>
<td>7.3</td>
</tr>
<tr>
<td>40</td>
<td>42.7</td>
</tr>
<tr>
<td>12</td>
<td>14.7</td>
</tr>
<tr>
<td>88</td>
<td>85.3</td>
</tr>
<tr>
<td>controls</td>
<td>controls</td>
</tr>
<tr>
<td>22</td>
<td>150</td>
</tr>
<tr>
<td>128</td>
<td>150</td>
</tr>
</tbody>
</table>

For example,

\[ E_{11} = \frac{50(22)}{150} = 7.3 \]

\[ E_{12} = \frac{50(128)}{150} = 42.7 \]

etc.
Hypothesis Testing, Contingency Table Method

• 1. We compute the test statistic:

\[ X_{corr}^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}} \]

which is called a Yates-corrected chi-square statistic.

2. Under \( H_0 \), it can be shown that \( X_{corr}^2 \) follows a chi-square distribution with 1 df.

3. Criterion for using this test: all \( E_{ij} \) should be \( \geq 5 \).
<table>
<thead>
<tr>
<th>d</th>
<th>.005</th>
<th>.01</th>
<th>.025</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
<th>.975</th>
<th>.99</th>
<th>.995</th>
<th>.999</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.999*</td>
<td>0.9998</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>0.01</td>
<td>0.025</td>
<td>0.05</td>
<td>0.10</td>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
<td>0.90</td>
<td>0.95</td>
<td>0.975</td>
<td>0.99</td>
<td>0.995</td>
<td>0.999</td>
</tr>
<tr>
<td>3</td>
<td>0.0015</td>
<td>0.0010</td>
<td>0.0006</td>
<td>0.0010</td>
<td>0.0014</td>
<td>0.0019</td>
<td>0.0025</td>
<td>0.0030</td>
<td>0.0035</td>
<td>0.0040</td>
<td>0.0045</td>
<td>0.0050</td>
<td>0.0055</td>
<td>0.0060</td>
</tr>
<tr>
<td>4</td>
<td>0.0010</td>
<td>0.0007</td>
<td>0.0004</td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0012</td>
<td>0.0015</td>
<td>0.0018</td>
<td>0.0020</td>
<td>0.0022</td>
<td>0.0024</td>
<td>0.0026</td>
<td>0.0028</td>
<td>0.0030</td>
</tr>
<tr>
<td>5</td>
<td>0.0006</td>
<td>0.0004</td>
<td>0.0002</td>
<td>0.0007</td>
<td>0.0008</td>
<td>0.0011</td>
<td>0.0014</td>
<td>0.0016</td>
<td>0.0018</td>
<td>0.0020</td>
<td>0.0021</td>
<td>0.0023</td>
<td>0.0024</td>
<td>0.0025</td>
</tr>
</tbody>
</table>

* = 0.0000393  † = 0.000157  ‡ = 0.000982  § z_{.02} = u^{th} percentile of a $z^2$ distribution with d degrees of freedom.

Use of the chi-square table

• To evaluate statistical significance using a chi-square table, we use the 95\textsuperscript{th} percentile of the table as the critical value if $\alpha=0.05$.

• Thus, with 1 df, the Yates-corrected chi-square must be $\geq 3.84$ for results to be significant at the 5% level.
Use of the chi-square distribution with computer programs

- One can obtain p-values and percentiles from the chi-square distribution using the
  - CHIDIST AND CHIINV functions of Excel
  - or the
  - chi2tail and invchi2tail functions of Stata
  - or the
  - pchi and qchi functions of R
Solution to Ovarian Cancer Problem

1. We have the following observed and expected contingency tables

<table>
<thead>
<tr>
<th>Observed Table</th>
<th>Expected Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>age at menarche</td>
<td>age at menarche</td>
</tr>
<tr>
<td>&lt; 11</td>
<td>11 +</td>
</tr>
<tr>
<td>cases</td>
<td>10</td>
</tr>
<tr>
<td>controls</td>
<td>12</td>
</tr>
<tr>
<td>22</td>
<td>128</td>
</tr>
</tbody>
</table>

2. We calculate the chi-square statistic as follows:
Thus, there was no significant difference between the % of women with early menarche between ovarian cancer cases and controls.
Use of computer programs to implement the chi-square test for 2 x 2 tables

• The chi-square test for 2 x 2 tables can be implemented using the
  • tabi command of Stata
  • or
  • PROC FREQ of SAS
  • or
  • the chisq.test command of R.
• An example with Stata is given on the next slide.
Use of Stata to analyze the ovarian cancer data

- `. tabi 10 40\12 88, chi2`

<table>
<thead>
<tr>
<th>col</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>row</strong></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>88</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>22</td>
<td>128</td>
</tr>
</tbody>
</table>

Pearson chi2(1) = 1.7045  Pr = 0.192
Use of Stata to analyze the ovarian cancer data

• Notice that Stata calculates the uncorrected chi-square statistic given by:

\[ X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_1^2 \text{ under } H_0, \]

\[ p-value = \Pr(\chi_1^2 > X^2). \]

I don’t recommend using uncorrected chi-square statistics, although this is controversial.

SAS provides both corrected and uncorrected chi-square statistics with its PROC FREQ output.
Use of R to perform the chi-square test for 2 x 2 tables

1. Suppose we have counts of a and b in the 1st row and c and d in the 2nd row of our 2 x 2 table.

2. We first form the matrix corresponding to the 2 x 2 table with the command:
   ```
   table <- matrix(c(a,c,b,d), nrow = 2)
   ```
   where c(a,c,b,d) is the concatenation function of R forming the vector (a,c,b,d).
Use of R to perform the chi-square test for 2 x 2 tables (cont.)

• The matrix command converts the one-dimensional vector to a two-dimensional matrix called table.

• **Note:** the counts have to be entered column-wise to form the matrix.
Use of R to perform the chi-square test for 2 x 2 tables (cont.)

• For example, for the ovarian cancer example,

• > counts< - c(10,12,40,88)
• > table<- matrix(counts, nrow = 2)
• > table
• [,1] [,2]
• [1,] 10 40
• [2,] 12 88
Use of R to perform the chi-square test for 2 x 2 tables (cont.)

3. Use the chisq.test command to calculate the Yates-corrected chi-square statistic using the syntax:

- chisq.test(table) for the Yates-corrected chi-square statistic
- and
- chisq.test(table, correct = FALSE) for the uncorrected chi-square statistic
Use of R to perform the chi-square test for 2 x 2 tables (cont.)

- \( \texttt{> chisq.test(table)} \)
- Pearson's Chi-squared test with Yates' continuity correction
- data: table
- \( X\text{-squared} = 1.1253, \; \text{df} = 1, \; \text{p-value} = 0.2888 \)

- \( \texttt{> chisq.test(table, correct = FALSE)} \)
- Pearson's Chi-squared test
- data: table
- \( X\text{-squared} = 1.7045, \; \text{df} = 1, \; \text{p-value} = 0.1917 \)

The results are the same as on slides 16 and 18.
Analysis of $R \times C$ Contingency Tables

• 1. In general, we form an $R \times C$ contingency table relating a row categorical variable ($R$ rows) to a column categorical variable ($C$ columns).

• 2. The hypotheses to be tested are:

$$H_0 : p_{ij} = a_i b_j \text{ for all } i,j \text{ vs. } H_1 : p_{ij} \neq a_i b_j \text{ for at least one } i,j,$$

where $p_{ij} = \Pr(\text{subject is in the } i\text{th row and } j\text{th column})$,

$a_i = \Pr(\text{ith row})$, $b_j = \Pr(\text{jth column})$. 
Analysis of $R \times C$ Contingency Tables (cont.)

1. Under the null hypothesis we assume that there is no association between row and column classifications.

2. Corresponding to the observed $R \times C$ table we compute an expected $R \times C$ table, where $E_{ij} = R_iC_j/N$, $i=1,..., R$; $j=1,...,C$; and $R_i = ith$ row total, $C_j = jth$ column total, $N = grand total$. 
Analysis of $R \times C$ Contingency Tables (cont.)

- 3. The test statistic is given by:

\[
X^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(R-1)(C-1)} \text{ under } H_0.
\]

4. The p-value = $\Pr(\chi^2_{(R-1)(C-1)} > X^2)$.

5. We only use this test if no more than 1/5 of the expected values are < 5 and no expected value is < 1.
Analysis of 2 x k contingency tables

Nurses’ Health Study Dietary Analyses

(1) In 1980, the Nurses’ Health Study subjects filled out a dietary questionnaire (FFQ) where they reported their frequency of consumption per week over the past year of 61 individual food items.

(2) A nutritional database was used to convert intake of individual foods into overall nutrient intake.

(3) One nutrient of interest was alcohol consumption (grams/day). One drink of alcohol contains approximately 11-14 grams of alcohol.

(4) An analysis was undertaken to relate the consumption of alcohol in 1980 to breast cancer incidence from 1980-1984.

The following data were reported for 50-54 year old women

<table>
<thead>
<tr>
<th>Alcohol Consumption</th>
<th>None</th>
<th>&lt; 1.5</th>
<th>1.5 - 4.9</th>
<th>5.0 -14.9</th>
<th>15.0 +</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>43</td>
<td>15</td>
<td>22</td>
<td>42</td>
<td>24</td>
</tr>
<tr>
<td>No</td>
<td>5901</td>
<td>2054</td>
<td>3427</td>
<td>3528</td>
<td>2893</td>
</tr>
<tr>
<td>Total</td>
<td>5944</td>
<td>2069</td>
<td>3449</td>
<td>3570</td>
<td>2917</td>
</tr>
</tbody>
</table>

4-yr Incidence Rate (per 10^5)

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-yr Incidence Rate (per 10^5)</td>
<td>723</td>
<td>725</td>
<td>638</td>
</tr>
</tbody>
</table>

(1) We could use the chi-square test for R x C tables to analyze the data

(2) In the special case of a 2 x 5 table this is equivalent to testing the hypothesis $H_0$: $p_1 = p_2 = \ldots = p_5$ vs. $H_1$: at least two of the $p_i$’s are different

• We use the test statistic:

\[ X^2 = \sum_{i=1}^{2} \sum_{j=1}^{5} (O_{ij} - E_{ij})^2 / E_{ij} \sim \chi^2_4 \text{ under } H_0 \]

with p-value = Pr(\(\chi^2_4 > X^2\)).
Chi-square test for heterogeneity applied to breast cancer-alcohol data

<table>
<thead>
<tr>
<th>Breast Cancer</th>
<th>Alcohol Consumption (g/day)</th>
<th>0</th>
<th>&lt; 1.5</th>
<th>1.5 - 4.9</th>
<th>5.0 - 14.9</th>
<th>15.0 +</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Observed</td>
<td>43</td>
<td>15</td>
<td>22</td>
<td>42</td>
<td>24</td>
<td>146</td>
</tr>
<tr>
<td></td>
<td>Expected</td>
<td>(48.3)</td>
<td>(16.8)</td>
<td>(28.1)</td>
<td>(29.0)</td>
<td>(23.7)</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>Observed</td>
<td>5901</td>
<td>2054</td>
<td>3427</td>
<td>3528</td>
<td>2893</td>
<td>17803</td>
</tr>
<tr>
<td></td>
<td>Expected</td>
<td>(5895.7)</td>
<td>(2052.2)</td>
<td>(3420.9)</td>
<td>(3541.0)</td>
<td>(2893.3)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>5944</td>
<td>2069</td>
<td>3449</td>
<td>3570</td>
<td>2917</td>
<td>17949</td>
</tr>
</tbody>
</table>
Chi-square test for heterogeneity applied to breast cancer-alcohol data (cont.)

• The chi-square statistic is given by:

\[ X^2 = \frac{(43 - 48.3)^2}{48.3} + \ldots + \frac{(2893 - 2893.3)^2}{2893.3} = 7.95 \sim \chi^2_4 \text{ under } H_0. \]

p-value = \Pr(\chi^2_4 > 7.95) = 0.093.
Stata output for chi-square test for heterogeneity for breast-cancer vs. alcohol data

- `tabi 43 15 22 42 24\5901 2054 3427 3528 2893, chi2`

<table>
<thead>
<tr>
<th></th>
<th>col</th>
</tr>
</thead>
<tbody>
<tr>
<td>row</td>
<td>1</td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
</tr>
<tr>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>2</td>
<td>5,901</td>
</tr>
<tr>
<td>Total</td>
<td>5,944</td>
</tr>
</tbody>
</table>

- **Pearson chi2(4) = 7.9502  Pr = 0.093**
Testing for trend in 2 x k tables

(1) One problem with the chi-square test for heterogeneity is that the alternative hypothesis is very non-specific.

(2) Thus, the test for heterogeneity may have low power to detect the hypothesis of interest which is that increasing alcohol consumption is associated with an increase in the risk of breast cancer.
(3) Instead, if we let $p_i = \text{risk of breast cancer for women in the } i^{th} \text{ alcohol exposure group}$ and $S_i = \text{score variable that represents the median grams of alcohol intake per day in the } i^{th} \text{ group}$, then we can test the hypothesis:

$$H_0: \beta = 0 \text{ vs. } H_1: \beta \neq 0$$

where

$$p_i = \alpha + \beta S_i$$

(4) These hypotheses can be tested using the chi-square test for trend for 2 x k tables.
Chi-square test for trend for 2 x k tables

- Suppose we have a 2 x k table and $p_i =$ event rate for the $i^{th}$ group which is estimated by

$$\hat{p}_i = \frac{x_i}{n_i},$$

where $x_i =$ # events, $n_i =$ # subjects in the $i^{th}$ group.

Let $S_i =$ score variable for the $i^{th}$ group.

We wish to test the hypothesis $H_0: \beta = 0 \ vs. \ H_1: \beta \neq 0$,

where $p_i = \alpha + \beta S_i$.
Chi-square test for trend for 2 x k tables (cont.)

• The test statistic is given by:

\[ X_1^2 = \frac{A^2}{B} \sim \chi_1^2 \text{ under } H_0, \text{ where} \]
\[ A = \text{ observed score among cases} \]
\[ \text{minus expected score among cases,} \]
\[ B = \text{ variance of observed score among cases} \]
\[ \text{under } H_0. \]

The p-value = \( \Pr(\chi_1^2 > X_1^2) \).
(5) This test should only be used when
\[ n\bar{pq} \geq 5, \]
where \( \bar{p} = \) overall \% of cases, \( \bar{q} = 1 - \bar{p}. \)
Note that this test can often be used when individual expected values are small, provided that \( n\bar{pq} \geq 5. \)

6. This test can be implemented using the Stata command `tabodds`. 
Use of Stata tabodds function to implement chi-square test for trend (cont.)

- . tabodds case alcohol [fweight=freq]

<table>
<thead>
<tr>
<th>alcohol</th>
<th>cases</th>
<th>controls</th>
<th>odds</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>43</td>
<td>5901</td>
<td>0.00729</td>
<td>0.00540  0.00984</td>
</tr>
<tr>
<td>.75</td>
<td>15</td>
<td>2054</td>
<td>0.00730</td>
<td>0.00439  0.01214</td>
</tr>
<tr>
<td>3.25</td>
<td>22</td>
<td>3427</td>
<td>0.00642</td>
<td>0.00422  0.00976</td>
</tr>
<tr>
<td>10</td>
<td>42</td>
<td>3528</td>
<td>0.01190</td>
<td>0.00878  0.01614</td>
</tr>
<tr>
<td>15</td>
<td>24</td>
<td>2893</td>
<td>0.00830</td>
<td>0.00555  0.01240</td>
</tr>
</tbody>
</table>

Test of homogeneity (equal odds): chi2(4) = 7.95
Pr>chi2 = 0.0934

Score test for trend of odds: chi2(1) = 2.73
Pr>chi2 = 0.0988
SAS implementation of the chi-square test for trend

• You can use PROC FREQ of SAS to implement the chi-square test for trend.

• You have to use the Statistics option and refer to the Mantel-Haenszel chi-square.

• One limitation is that the score variable is assumed to be 1,2,..., k for categories 1,2,..., k.

• In the alcohol example, a more appropriate score is the median alcohol intake within each category.
R implementation of the chi-square test for trend

• In R, you can use the prop.trend.test command to perform the chi-square test for trend.
Alternative version of test statistic in terms of slopes

(1) We can also re-express the test statistic for the chi-square test for trend in terms of the slope of \( \hat{p}_i \) on \( S_i \)

\[
\hat{\beta} = \text{estimated slope of } \hat{p}_i \text{ on } S_i,
\]

\[
\text{var}(\hat{\beta}) = \text{variance of estimated slope}.
\]

\[
X^2_1 = \frac{\hat{\beta}^2}{\text{var}(\hat{\beta})} \sim \chi^2_1 \text{ under } H_0.
\]

This test is referred to as the Cochran-Armitage test.
We should only use this test if \( N \bar{p}(1 - \bar{p}) \geq 5 \).

This version allows for a 95% CI for \( \beta \).
Cochran-Armitage test applied to NHS Breast cancer-alcohol data

(1) \( \hat{\beta} = 19.5 \times 10^5 \)
   \[ \text{var}(\hat{\beta}) = 1.40 \times 10^{-8} \]
   \[ \text{se}(\hat{\beta}) = 1.18 \times 10^{-4} = 11.8 \times 10^5 \]
   \[ X_1^2 = (19.5/11.8)^2 = 2.72 \sim \chi_1^2, \text{p-value} = 0.099. \]

(2) 95% CI for \( \beta = 19.5 \pm 1.96 \times (11.8) \times 10^5 = (-3.6, 42.6) \times 10^5 \)

(3) Interpretation of slope
   Increase in 4-year cumulative incidence for 1 drink/day
   \( \approx 11 \) grams alcohol/day
   \[ = 11 \times (19.5) / 10^5 = 214.5 / 10^5 \]
The Cochran-Armitage test can also be used in genetic association studies.

For example, SNP RS1061170 has often been associated with the progression of age-related macular degeneration in genetic association studies.

The following data were obtained from the AREDS study:
We can use the Cochran-Armitage test to compare the progression rate according to the number of C alleles in the RS 1061170 genotype.
Cochran-Armitage test (cont.)

• The model would take the form:

\[ p_i = \alpha + \beta S_i, \]

where

\[ p_i = \text{proportion of progressors among subjects with } i \text{ C alleles}, \]
\[ S_i = \text{number of C alleles } = i, i = 0,1,2. \]

The analysis of these data using Stata is shown on the next 2 slides.
. insheet using amd.csv, names comma
clear
(3 vars, 6 obs)

. list

|-----------------------------+|
| c_alle~s | progress | freq |
|-----------------------------|
1. | 0 | 0 | 366 |
2. | 0 | 1 | 39 |
3. | 1 | 0 | 521 |
4. | 1 | 1 | 116 |
5. | 2 | 0 | 280 |
6. | 2 | 1 | 124 |
|-----------------------------+|
The results indicate a significant association between the number of C alleles and progression of AMD.
Confounding variables

• A confounding variable is a variable that is associated with both disease and exposure.

• If we don’t control for confounding variable(s), then we may obtain biased estimates of disease-exposure relationships.
Example of confounding variable

• Earlier in this lecture, we used the chi-square test for trend to study the association between breast cancer incidence and alcohol intake.

• However, age is associated with breast cancer and may also be associated with alcohol intake.

• Hence, we should control for age in studying the association between breast cancer incidence and alcohol intake.
Limitations of contingency table analysis

- One approach to control for confounding is to stratify the dataset by a confounding variable(s) and use contingency table analyses within individual strata.

- For example, in the breast cancer-alcohol example, we could perform separate analyses for women age < 50 vs. age >= 50.
Limitations of contingency table analysis (cont.)

• However, we will lose power as a result and won’t completely control for confounding if the age strata are broad.

• A better approach is to use a regression-type model that simultaneously adjusts for both the primary exposure (e.g. alcohol) and confounders (e.g., age) in predicting breast cancer incidence.
Choices for Regression models for binary outcome data

• If we have covariates $x_{i1},...,x_{ir}$ and a probability of disease $= p_i$, then we could try to fit a model

  \[ p_i = \alpha + \beta_1 x_{i1} + \ldots + \beta_r x_{ir} \]

• However, the predicted probability from this model could be $< 0$ or $> 1$, which is impossible.
Choices for Regression models for binary outcome data (cont.)

• Similar concerns exist for an exponential model:

\[
p_i = \exp(\alpha + \beta_1 x_{i1} + \ldots + \beta_r x_{ir})
\]

• where the predicted \( p_i \) can be > 1.
Logits

• We define logit(p) = ln[p/(1-p)] = ln(odds)
• Thus, if p = 0.2, logit(0.2) = ln(1/4) = -ln(4) = -1.39,
• if p = 0.5, logit(0.5)= ln(1) = 0,
• if p = 0.8, logit(0.8) = ln(4) = 1.39,
• if p = 0, then logit(0) = ln(0) = -∞,
• if p = 1, then logit(1) = ln(∞) = ∞.
Logistic Regression Model

- A logistic regression model is defined by
  \[
  \ln \left( \frac{p_i}{1 - p_i} \right) = \logit(p_i) = \alpha + \beta_1 x_{i1} + \ldots + \beta_r x_{ir}
  \]

- We are fitting a linear model in the logit scale.
Logistic Regression Model (cont.)

- If we solve for $p_i$, then we obtain

$$p_i = \frac{\exp(\alpha + \beta_1 x_{i1} + \ldots + \beta_r x_{ir})}{1 + \exp(\alpha + \beta_1 x_{i1} + \ldots + \beta_r x_{ir})}$$
Logistic Regression Model (cont.)

• Thus, probabilities under this model must be between 0 and 1.
Interpretation of Logistic Regression Intercept with a single binary covariate

- Let $x_i = 1$ if exposed, = 0 if unexposed.
- Let $p_i = $ probability of disease for the $i$th subject
- Suppose we fit the logistic regression model:

$$\ln[p_i/(1-p_i)] = \alpha + \beta x_i$$

- If $x_i = 0$, then $\text{logit}(p_i) = \alpha = \log(\text{odds})$, and $\text{odds} = \exp(\alpha)$. 
Interpretation of Logistic Regression Slope with a single binary covariate

• Suppose we compare logit($p_i$) for an exposed vs. an unexposed person.

• $\beta = \logit(p_i \mid x_i = 1) - \logit(p_i \mid x_i = 0)$
  
  • $= \ln(\text{odds} \mid x_i = 1) - \ln(\text{odds} \mid x_i = 0)$
  
  • $= \ln(\text{OR})$.

• Thus, $\text{OR} = \exp(\beta) = \text{odds ratio relating disease to exposure}$
Interpretation of logistic regression slope with a single continuous covariate

• We have $\text{logit}(p_i) = \alpha + \beta x_i$
• where $x_i$ is continuous.

• Suppose we compare 2 people A and B with levels of $(x+1)$ and $x$ for the continuous covariate $X$. We have:

• $\text{logit}_A = \alpha + \beta (x + 1)$
• $\text{logit}_B = \alpha + \beta x$
Interpretation of logistic regression slope with a single continuous covariate (cont.)

- Hence,
- \[ \text{logit}_A - \text{logit}_B = \beta \]
- or
- \[ \frac{\text{odds}_A}{\text{odds}_B} = \exp(\beta) = \text{OR}_{A \text{ vs. } B} \]

- Thus, the odds in favor of disease for person A vs. person B = \exp(\beta).
Effect of the use of surfactant on in-hospital mortality

• One of the leading causes of death for low birthweight babies is the respiratory distress syndrome (RDS)

• During the period 1985-1990, surfactant was introduced to wide-scale clinical use
Effect of the use of surfactant on in-hospital mortality (cont.)

• Surfactant is introduced intratracheally (i.e. through the windpipe) to infants with RDS

• A study was performed comparing in-hospital mortality in 14 hospitals before and after surfactant use (Schwartz, et al NEJM, 1991).
Relationship of Surfactant use to in-hospital mortality

• Before surfactant use, there were 3922 births with birthweight 500-1500 g, of which 960 died in the hospital (group 2).

• After surfactant use, there were 1707 births with birthweight 500-1500 g, of which 335 died in the hospital (group 1).
Relationship of Surfactant use to in-hospital mortality (cont.)

This yields:

\[ \hat{p}_2 = 0.245, \hat{p}_1 = 0.196, \]

where

\[ \hat{p}_1 = \text{mortality risk after surfactant was introduced}, \]
\[ \hat{p}_2 = \text{mortality risk before surfactant was introduced}. \]

We could also display the data in the form of a 2 x 2 contingency table as shown on the next slide.
### Relationship of Surfactant use to in-hospital mortality

<table>
<thead>
<tr>
<th>Surfactant use</th>
<th>In-hospital Mortality</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>yes</td>
<td>1707</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>1372</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>335</td>
</tr>
<tr>
<td>No</td>
<td>yes</td>
<td>3922</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2962</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>960</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>5629</td>
</tr>
</tbody>
</table>
Relationship of Surfactant use to in-hospital mortality (cont.)

• We could analyze these data using Stata as shown on the next slide.
Using Stata to obtain the OR and test of significance with the chi-square test for 2 x 2 tables for the surfactant data

- . csi 335 960 1372 2962, or woolf

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>335</td>
<td>960</td>
<td>1295</td>
</tr>
<tr>
<td>Noncases</td>
<td>1372</td>
<td>2962</td>
<td>4334</td>
</tr>
<tr>
<td>Total</td>
<td>1707</td>
<td>3922</td>
<td>5629</td>
</tr>
</tbody>
</table>

| Risk          | .1962507| .2447731  | .2300586|

<table>
<thead>
<tr>
<th></th>
<th>Point estimate</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk difference</td>
<td>-.0485223</td>
<td>-.0716748  -.0253699</td>
</tr>
<tr>
<td>Risk ratio</td>
<td>.801766</td>
<td>.717798    .8955566</td>
</tr>
<tr>
<td>Prev. frac. ex.</td>
<td>.198234</td>
<td>.1044434  .282202</td>
</tr>
<tr>
<td>Prev. frac. pop</td>
<td>.0601147</td>
<td></td>
</tr>
<tr>
<td>Odds ratio</td>
<td>.7533634</td>
<td>.6550237  .866467 (Woolf)</td>
</tr>
</tbody>
</table>

\[ \text{chi}^2(1) = 15.81 \quad \text{Pr}>\text{chi}^2 = 0.0001 \]

Thus, there is a significant association (p < 0.001), with estimated OR = 0.75.
Using indicator variables in logistic regression models

• We continue with the study of the association between surfactant use and infant mortality.

• We now wish to explore the association of mortality with the confounding variable birthweight.

• We can treat birthweight either as a continuous or categorical variable.
Using indicator variables in logistic regression models (cont.)

• We will treat birthweight as a categorical variable, since it is likely that there is a non-linear relationship between birthweight and infant mortality.

• To include birthweight as categorical in logistic regression models, we need to represent it as a set of indicator variables.
Using indicator variables in logistic regression models (cont.)

• We need to select 1 category as a reference group and enter a set of indicator (or dummy variables) into the model

• If we choose 500-749 g (group 1) as the reference group, then we create:
  • \( x_1 = 1 \) if birthweight = 2 (750-999 g)/=0 else
  • \( x_2 = 1 \) if birthweight = 3 (1000-1249 g)/=0 else
  • \( x_3 = 1 \) if birthweight = 4 (1250-1499 g)/= 0 else
  • and fit the logistic regression model:
Using indicator variables in logistic regression models (cont.)

\[
\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,
\]

where we drop the subscript \(i\) for simplicity.
Using indicator variables in logistic regression models (cont.)

We can interpret $\beta_1$, $\beta_2$ and $\beta_3$ as follows:

$OR_{2 \text{ vs. } 1} = OR$ for 1 yr mortality for group 2 vs. group 1 = $\exp(\beta_1)$

$OR_{3 \text{ vs. } 1} = OR$ for 1 yr mortality for group 3 vs. group 1 = $\exp(\beta_2)$

$OR_{4 \text{ vs. } 1} = OR$ for 1 yr mortality for group 4 vs. group 1 = $\exp(\beta_3)$
Using indicator variables in logistic regression models (cont.)

• Stata provides an overall test of the hypothesis $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs. $H_1$: at least one of the $\beta_i$’s are different from 0, which is equivalent to the chi-square test for heterogeneity for 2 x 4 tables.

• In addition, it provides estimates of $OR_1$, $OR_2$ and $OR_3$, where the reference group = group 1 for each odds ratio.
Using indicator variables in logistic regression models (cont.)

- Stata will construct the dummy variables for us in the model if we specify `i.birthweight` as a predictor in the model (rather than `birthweight`) and use the logistic command

  - `xi: logistic death i.birthweight [fweight = freq]`

- where the `xi:` prefix is needed because of the indicator variables for `birthweight`. The results are as follows:
Using indicator variables in logistic regression models (cont.)

- `. xi: logistic death i.birthweight [fweight = freq]
  i.birthweight     _Ibirthweig_1-4    (naturally coded; _Ibirthweig_1 omitted)
- Logistic regression
  Number of obs    =     5629
  LR chi2(3)       =    1114.73
  Prob > chi2      =     0.0000
  Log likelihood   =  -2478.6223           Pseudo R2       =     0.1836

|                                            | Odds Ratio | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|-------------------------------------------|------------|-----------|-------|-------|----------------------|
| _Ibirthweig_2    | 0.2303739  | 0.0203267 | -16.64| 0.000 | 0.1937889            |
| _Ibirthweig_3    | 0.097544   | 0.0097884 | -23.19| 0.000 | 0.0801279            |
| _Ibirthweig_4    | 0.0499619  | 0.0056048 | -26.71| 0.000 | 0.0401005            |
Using indicator variables in logistic regression models (cont.)

• We see that the estimated OR for group 2, 3 and 4 vs. group 1 are 0.23, 0.10 and 0.05 and all are highly significant vs. group 1.

• Comparable analyses can be performed using PROC LOGISTIC of SAS by representing birthweight as a CLASS variable.
Surfactant Analysis

• In our previous analyses, we have showed that the OR for mortality in the 1\textsuperscript{st} year = 0.75 after surfactant use vs. before surfactant use.

• Also, birthweight was an important predictor of mortality in the 1\textsuperscript{st} year among infants with RDS and birthweight < 1500 g, with higher mortality as birthweight decreases.

• We now consider them jointly as follows:
Multiple Logistic Regression Analysis

- We consider the model
  \[
  \text{logit}(p_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}
  \]

- where
  - \(x_{1i}\) = surfactant use (1=yes/0=no)
  - \(x_{2i}\) = 1 if birthweight = 750-999 g/ = 0 else
  - \(x_{3i}\) = 1 if birthweight = 1000-1249 g/ = 0 else
  - \(x_{4i}\) = 1 if birthweight = 1250-1500g/= 0 else

- How do we interpret the parameters of this model?
Partial Regression Coefficients

• If there is more than 1 predictor in a logistic regression model, then the coefficients in the model are referred to as partial regression coefficients.
Partial Regression Coefficients (cont.)

• Suppose we consider two individuals (A, B), where subject A was born after surfactant was used \((x_1 = 1)\), while subject B was born before surfactant was used \((x_1 = 0)\).

• Suppose also that both individuals are in the same birthweight group (e.g., group 3 = 1000-1249 g).
Partial Regression Coefficients (cont.)

- It follows that:

  \[ \text{logit}(p_A) = \alpha + \beta_1 + \beta_3 \]
  \[ \text{logit}(p_B) = \alpha + \beta_3 \]
  and
  \[ \text{logit}(p_A) - \text{logit}(p_B) = \beta_1, \]
  or
  \[ OR_{A \text{ vs. B}} = \exp(\beta_1), \]
  which is interpreted as the effect of surfactant adjusted for birthweight group.
Similarly, let us consider two individuals C, D, where individual C is in birthweight group 3, individual D is in birthweight group 1 and both individuals are born after surfactant use was introduced.

It follows that:
Partial Regression Coefficients (cont.)

\[
\logit(p_C) = \alpha + \beta_1 + \beta_3, \\
\logit(p_D) = \alpha + \beta_1, \\
\text{and} \\
\logit(p_C) - \log it(p_D) = \beta_3, \\
or \\
\text{OR}_{C \text{ vs. } D} = \exp(\beta_3),
\]

which is interpreted as the effect of birthweight group 1000-1249 g vs. birthweight group 500-749 g after accounting for differences in surfactant use.

The Stata analyses are given in the next slide.
Stata multiple logistic model of 1 year mortality on surfactant use and birthweight

- . xi: logistic death surfactant i.birthweight [fweight=freq]
- i.birthweight  _Ibirthweig_1-4  (naturally coded; _Ibirthweig_1 omitted)

Logistic regression  
  Number of obs   =       5629
  LR chi2(4)      =    1134.94
  Prob > chi2     =     0.0000
  Log likelihood =  -2468.517  
  Pseudo R2       =     0.1869

|                | Odds Ratio | Std. Err. | z    | P>|z|    | [95% Conf. Interval] |
|----------------|------------|-----------|------|-------|----------------------|
| surfactant     | .7005633   | .056131   | -4.44| 0.000 | .5987519    .8196866 |
| _Ibirthwei-2  | .2294394   | .0203125  | -16.63| 0.000 | .1928902    .2729144 |
| _Ibirthwei-3  | .0959767   | .0096688  | -23.26| 0.000 | .0787798    .1169275 |
| _Ibirthwei-4  | .0494499   | .0055618  | -26.73| 0.000 | .0396668    .0616457 |

.
We see that there is still a significant effect of surfactant even after controlling for birthweight group (OR = 0.70, 95% CI = 0.60-0.82, p < 0.001), which is actually slightly stronger than for the crude analyses (OR = 0.75, 95% CI = 0.66-0.87, p < 0.001).
Summary

1. Chi-square methods are useful for assessing the association between a binary outcome variable and a single categorical exposure variable either coded as binary or as categorical with more than 2 categories.

2. The chi-square test for trend is useful when the exposure variable is ordinal and one can define an average level of exposure within a specific category.
Summary (cont.)

• 3. Chi-square methods are of limited utility when one has several exposure variables to be considered in the same model.

• 4. Logistic regression is an extension of chi-square methods and allows one to predict a binary outcome variable as a function of one or more predictor variables that may be either continuous or categorical.

• 5. Also, effect of predictors can be interpreted in terms of OR’s which makes for easy interpretation.