This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.

Copyright 2009, The Johns Hopkins University and John McGready. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided “AS IS”; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.
Section B

The Case for Logistic Regression
Example

- Relation between age and coronary heart disease

- Next slide: (Excerpt from) table of age and coronary heart disease
  - Evidence status (CHD) of 58 subjects (average age 45 years, range 20 to 64, 43% showed evidence of CHD) selected from a hospital population and screened for evidence CHD
Example

- Excerpt of the data

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>CHD</th>
<th>ID</th>
<th>Age</th>
<th>CHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>0</td>
<td>10</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>0</td>
<td>11</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>0</td>
<td>12</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>0</td>
<td>13</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>1</td>
<td>14</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>26</td>
<td>0</td>
<td>15</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>
Example

- **Goal**
  - To determine whether age is a risk factor for CHD and estimate the magnitude of this outcome exposure relationship in this patient population

- **Options for analysis**
  - Categorize age into two groups, do a comparison of proportions
  - Categorize age into multiple groups (three or four) and do multiple comparisons of proportions
  - Analyze using age as a continuous measure
Outcome

- Presence/absence of CHD evidence from screening result
  - $y = 1$ if there is CHD evidence
  - $y = 0$ if there is no CHD evidence

- Notice that $y$ only takes on two values: 1 ("yes") or 0 ("no")
Examples

- Options for analysis
  - Categorize age into two groups, do a comparison of proportions
  - Categorize age into multiple groups (three or four) and do multiple comparisons of proportions
  - Analyze using age as a continuous measure
Examples

- Could we use linear regression? Here is a scatterplot of $y$ versus $x$
Examples

- Could we use linear regression?
Examples

- Could we use linear regression?

\[ \hat{y} = .40? \]
How about creating age intervals as initially suggested?

<table>
<thead>
<tr>
<th>Age Group</th>
<th>n</th>
<th>Absent</th>
<th>Present</th>
<th>Proportion with CHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>0.20</td>
</tr>
<tr>
<td>30-34</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>35-39</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>40-44</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>0.13</td>
</tr>
<tr>
<td>45-49</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>0.44</td>
</tr>
<tr>
<td>50-54</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>0.60</td>
</tr>
<tr>
<td>55-59</td>
<td>12</td>
<td>3</td>
<td>9</td>
<td>0.75</td>
</tr>
<tr>
<td>60-69</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Example: CHD and Age

- Notice, each of the age intervals contain very few observations

<table>
<thead>
<tr>
<th>Age Group</th>
<th>n</th>
<th>Absent</th>
<th>Present</th>
<th>Proportion with CHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>0.20</td>
</tr>
<tr>
<td>30-34</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>35-39</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>40-44</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>0.13</td>
</tr>
<tr>
<td>45-49</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>0.44</td>
</tr>
<tr>
<td>50-54</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>0.60</td>
</tr>
<tr>
<td>55-59</td>
<td>12</td>
<td>3</td>
<td>9</td>
<td>0.75</td>
</tr>
<tr>
<td>60-69</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Example: CHD and Age

- There appears to be some structure/pattern here (percentage with CHD tends to increase with age)
Logistic Regression

- Wouldn’t it be nice if we could model this structure without having to categorize age?

- Logistic regression will allow for such a curve relating (equation) the proportion with outcome to age

- It can do it without actually dividing up age into intervals
Logistic Regression

- Logistic regression will allow for the estimation of an equation that fits a curve that is the probability of CHD relationship/age.
Logistic Regression

- A regression method to deal with the case when the dependent (outcome) variable $y$ is binary (dichotomous)

- There can be many predictor variables ($x$s)
Objectives of Logistic Regression

- Estimating a magnitude of outcome/exposure relationships
  - To evaluate the association of a binary outcome with a set of predictors

- Prediction
  - Develop an equation to determine the probability or likelihood that the individual has the condition \( y = 1 \) that depends on the independent variables (the xs)
Linear vs. Logistic Regression

- Linear regression
  - Outcome variable $y$ is continuous

- Logistic regression
  - Outcome variable $y$ is binary (dichotomous)

- The only (data type) question a researcher need ask when choosing a regression method is . . .
  - “What does my outcome look like?”
  - Either regression method allows for many $x$s (independent variables)
  - These $x$s can be either continuous or discrete
The Logistic Regression Model

- Equation for Pr(y = 1) - the proportion of subjects with y = 1

\[ p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots}} \]

- \( e \) is the “natural constant” \( \approx 2.718 \)
- \( p \) = probability (proportion) of \( y = 1 \)
Why is this equation appropriate?

\[ 0 < e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots} \leq \infty \]

And so it follows:

\[ 0 < \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots}} \leq 1 \]
The Logistic Regression Model

- $0 < p \leq 1$

- This formulation for $p$ ensures that our estimates of the probability of having the condition “$y$” is between 0 and 1
The Logistic Regression Model

- Can be transformed as follows

\[ \ln\left( \frac{p}{1 - p} \right) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... \]

- Sometimes written as:

\[ \log\left( \frac{p}{1 - p} \right) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... \]

- Where \( \ln \) (or \( \log \)) is the natural logarithm (base \( e \))
The Logistic Regression Model

- Recall, the odds of an event is defined as:

  \[
  \text{odds} = \left[ \frac{p}{1 - p} \right]
  \]

- Where \( p \) = probability of having the event “\( y \),” i.e., the proportion of persons with \( y = 1 \)
Logistic Regression Model

- For the CHD-age data set, we could try to estimate the following:

\[ \ln\left( \frac{p}{1 - p} \right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 \]

- \( p \) = probability of CHD evidence (proportion of persons with CHD evidence), \( x_1 \) = age

- \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are called regression coefficients

- Another way to write the above equation:

\[ \log(ODDS \ of \ CHD) = \hat{\beta}_0 + \hat{\beta}_1 x_1 \]
Logistic Regression Model

- Recall, from 611, the higher the odds of an event, the larger the probability of an event.

- A predictor $x_i$ that is positively associated with the odds will also be positively associated with the probability of the event (i.e., estimated slope $\hat{\beta}_i$ will be positive).

- A predictor $x_i$ that is negatively associated with the odds will also be negatively associated with the probability of the event (i.e., estimated slope $\hat{\beta}_i$ will be negative).
Example: CHD and Age

- Results from logistic regression of log odds of CHD evidence on age:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimated Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.135</td>
<td>0.036</td>
</tr>
<tr>
<td>Constant</td>
<td>-6.54</td>
<td>1.73</td>
</tr>
</tbody>
</table>
Example: CHD and Age

- The resulting equation

\[ \ln\left( \frac{p}{1-p} \right) = -6.54 + 0.135 \times \text{Age} \]

- Where \( p \) is the estimated probability of evidence (i.e., the estimated proportions of persons with CHD evidence) amongst persons of a given age
The estimated coefficient \( \hat{\beta}_1 \) of age \( x_1 \) is positive; hence, we have estimated a positive association between age and log odds of CHD.

Therefore, we have estimated a positive association between age and probability of coronary heart disease evidence.

How can we actually interpret the value .135, though?

Let's write out the equation comparing two groups of individuals who differ in age by one year:
- Group 1, age = k years
- Group 2, age = k + 1 years
Example: CHD and Age

- The resulting equations estimating the \textit{ln odds} of CHD evidence in each age group

\[
\ln(\text{Odds of CHD}; x_1 = k + 1) = \hat{\beta}_0 + \hat{\beta}_1 (k + 1)
\]

\[
\ln(\text{Odds of CHD}; x_1 = k) = \hat{\beta}_0 + \hat{\beta}_1 k
\]
Example: CHD and Age

- Multiplying out, and taking the difference (subtracting)

\[
\ln(\text{Oddsof CHD}; x_1 = k + 1) = \hat{\beta}_0 + \hat{\beta}_1 k + \hat{\beta}_1
\]

\[
\ln(\text{Oddsof CHD}; x_1 = k ) = \hat{\beta}_0 + \hat{\beta}_1 k
\]
Example: CHD and Age

- Multiplying out, and taking the difference (subtracting)

\[
ln(Odds of CHD; x_1 = k + 1) = \hat{\beta}_0 + \hat{\beta}_1 k + \hat{\beta}_1
\]

\[
ln(Odds of CHD; x_1 = k) = \hat{\beta}_0 + \hat{\beta}_1 k
\]

\[\hat{\beta}_1\]

- So, when the dust settles:

\[
\hat{\beta}_1 = ln(Odds of CHD; x_1 = k + 1) - ln(Odds of CHD; x_1 = k)
\]
Example: CHD and Age

- Now . . .

\[ \hat{\beta}_1 = \ln(\text{Odds of CHD}; x_1 = k + 1) - \ln(\text{Odds of CHD}; x_1 = k) \]

- “Reversing” one of the famous properties of logarithms:

\[ \hat{\beta}_1 = \ln\left( \frac{\text{Odds of CHD}; x_1 = k + 1}{\text{Odds of CHD}; x_1 = k} \right) = \ln(\hat{OR}) \]

- So \( \hat{\beta}_1 \), the estimated slope for \( x_1 \) is the natural log of an estimated odds ratio:

- To get the estimated odds ratio, exponentiate \( \hat{\beta}_1 \), i.e.:

\[ \hat{OR} = e^{\hat{\beta}_1} \]
In our example, recall $\hat{\beta}_1 = 0.135$

Here, $\hat{OR} = e^{\hat{\beta}_1} = e^{0.135} \approx 1.14$

The estimated odds ratio of CHD evidence for a one-year age difference is 1.14, older to younger

If we were to compare two groups of people who differ by one year of age, the estimated odds ratio for CHD evidence of the older group to the younger group is 1.14 (this is valid for age comparisons within our original range of data, 20-69 years)

- 60 years old to 59 years old
- 45 years old to 44 years old
- 27 years old to 26 years old
General Interpretation: Slope in Logistic Regression

- $\hat{\beta}_1$ is the estimated change in the log odds of the outcome for a one unit increase in $x_1$
  - “Change in the log odds of CHD for a one year increase in age”

- It estimates the log odds ratio for comparing two groups of observations:
  - One group with $x_1$ one unit higher than the other

- This estimated slope can be exponentiated to get the corresponding estimated odds ratio
What about the Intercept?

- The resulting equation

\[
\ln\left(\frac{p}{1-p}\right) = -6.54 + .135 \times \text{Age}
\]

- Here, the intercept estimate \( \hat{\beta}_0 \) is again just a “placeholder”—it is the estimated \( \ln \text{odds} \) of CHD evidence for persons of age 0.

- The intercept is mathematically necessary to specify the entire equation and use the entire equation to estimate the \( \ln \text{odds} \) of the outcome for any group given \( x_1 \) (we’ll show how to do this in a later section).
The estimated regression coefficients are not the true population parameter regression coefficients. We will need to estimate a range of plausible values which takes into account error associated with an imperfect sample. We will need to test for a statistical significant association in the population.

We will need tools for doing inference (coming in a future section).