Basic Biostatistics in Medical Research

Lecture 3: Introduction to Linear Regression and Logistic Regression

Jungwha “Julia” Lee, PhD, MPH
Biostatistics Collaboration Center (BCC)
Department of Preventive Medicine
NU Feinberg School of Medicine
Outline

• **Multiple linear regression**
  – estimating parameters
  – testing hypotheses about parameters
  – confidence and prediction interval

• **Logistic regression**
  – odds ratio vs. relative risk
  – interpretation of coefficient
  – extensions of logistic regression
Data Analysis Overview

• All variables you encounter in data analysis can be classified according as
  [independent | dependent] and [categorical | continuous].
Data Analysis Overview

- The classification of the data dictates the tool you can use:

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Dependent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>Continuous</td>
<td>Linear Regression</td>
</tr>
<tr>
<td>Categorical</td>
<td>Chi-squared test</td>
</tr>
<tr>
<td>Categorical</td>
<td>ANOVA</td>
</tr>
</tbody>
</table>
Scatter Plots and Correlation

• A scatter plot is used to show the relationship between two variables.

• Correlation analysis is used to measure the strength of the association (linear relationship) between two variables.
  – Only concerned with the strength of the relationship
  – No causal effect is implied
Scatter Plot Examples

**Linear relationships**

**Curvilinear relationships**
Scatter Plot Examples

Strong relationships

Weak relationships
Scatter Plot Examples

No relationships

\begin{itemize}
\item \textbf{No relationships}:
\begin{itemize}
\item \textbf{No relationship}:
\item \textbf{No relationship}:
\end{itemize}
\end{itemize}
Correlation Coefficient

• The population correlation coefficient $\rho$ (rho) measures the strength of the association between the variables.

• The sample correlation coefficient $r$ is an estimate of $\rho$ and is used to measure the strength of the linear relationship in the sample observations.
Features of $\rho$ and $r$

- Unit free
- Range: $-1 \sim +1$
- The closer to $-1$
  - the stronger negative linear relationship
- The closer to $+1$
  - the stronger positive linear relationship
- The closer to $0$
  - the weaker the linear relationship
Examples of $r$ Values

- $r = -1$
- $r = -0.6$
- $r = 0$
- $r = +0.3$
- $r = +1$
Sample Correlation Coefficient

\[ r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \]

where:
- \( r \) = Sample correlation coefficient
- \( n \) = Sample size
- \( x \) = Value of the independent variable
- \( y \) = Value of the dependent variable
Sample Correlation Coefficient Example

\[ r = 0.886 \rightarrow \text{relatively strong positive linear association between } x \text{ and } y \]
Significance Test for Correlation

- Hypotheses
  - $H_0: \rho = 0$ (no correlation)
  - $H_A: \rho \neq 0$ (correlation exists)
- Test statistic ($df=n-2$)

\[
t = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{\frac{1}{n-2}}
\]
Regression Analysis

Regression analysis is used to:

- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the dependent variable
Simple Linear Regression

- Only one independent variable, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be caused by changes in x
Types of Regression Models

Positive Linear Relationship

Negative Linear Relationship

Relationship NOT Linear

No Relationship
Population Linear Regression Model

\[ y = \beta_0 + \beta_1 x + \varepsilon \]

- **Dependent Variable**
- **Population y intercept**
- **Population Slope Coefficient**
- **Independent Variable**
- **Random Error term, or residual**

**Linear component**

**Random Error component**
Linear Regression Assumptions

\[ y = \beta_0 + \beta_1 x + \varepsilon \]

- Error values (\( \varepsilon \)) are statistically independent
- Error values are *normally distributed* for any given value of \( x \)
- The probability distribution of the errors is *normal*
- The probability distribution of the errors has *constant variance*
- The underlying relationship between the \( x \) variable and the \( y \) variable is *linear*
Population Linear Regression

\[ y = \beta_0 + \beta_1 x + \varepsilon \]

- Observed Value of \( y \) for \( x_i \)
- Predicted Value of \( y \) for \( x_i \)
- Random Error for this \( x \) value
- Intercept = \( \beta_0 \)
- Slope = \( \beta_1 \)
Estimated Regression Model

The sample regression line provides an estimate of the population regression line.

\[ \hat{y}_i = b_0 + b_1 x \]

- Estimated (or predicted) y value
- Estimate of the regression intercept
- Estimate of the regression slope
- Independent variable

The individual random error terms \( e_i \) have a mean of zero.
Least Squares Criterion

- $b_0$ and $b_1$ are obtained by finding the values of $b_0$ and $b_1$ that minimize the sum of the squared residuals.

\[
\sum e^2 = \sum (y - \hat{y})^2
\]

\[
= \sum (y - (b_0 + b_1 x))^2
\]
The Least Squares Equation

- The formulas for $b_1$ and $b_0$ are:

\[ b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \]

algebraic equivalent:

\[ b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \]

\[ b_0 = \bar{y} - b_1 \bar{x} \]
Interpretation of Slope and Intercept

- $b_1$ is the estimated change in the average value of $y$ as a result of a one-unit change in $x$
- $b_0$ is the estimated average value of $y$ when the value of $x$ is zero
Finding the Least Squares Equation

• The coefficients $b_0$ and $b_1$ will usually be found using computer software, such as Stata or SAS.

• Other regression measures will also be computed as part of computer-based regression analysis.
Simple Linear Regression

Table 1. Age and systolic blood pressure (SBP) among 33 adult women

<table>
<thead>
<tr>
<th>Age</th>
<th>SBP</th>
<th>Age</th>
<th>SBP</th>
<th>Age</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>131</td>
<td>41</td>
<td>139</td>
<td>52</td>
<td>128</td>
</tr>
<tr>
<td>23</td>
<td>128</td>
<td>41</td>
<td>171</td>
<td>54</td>
<td>105</td>
</tr>
<tr>
<td>24</td>
<td>116</td>
<td>46</td>
<td>137</td>
<td>56</td>
<td>145</td>
</tr>
<tr>
<td>27</td>
<td>106</td>
<td>47</td>
<td>111</td>
<td>57</td>
<td>141</td>
</tr>
<tr>
<td>28</td>
<td>114</td>
<td>48</td>
<td>115</td>
<td>58</td>
<td>153</td>
</tr>
<tr>
<td>29</td>
<td>123</td>
<td>49</td>
<td>133</td>
<td>59</td>
<td>157</td>
</tr>
<tr>
<td>30</td>
<td>117</td>
<td>49</td>
<td>128</td>
<td>63</td>
<td>155</td>
</tr>
<tr>
<td>32</td>
<td>122</td>
<td>50</td>
<td>183</td>
<td>67</td>
<td>176</td>
</tr>
<tr>
<td>33</td>
<td>99</td>
<td>51</td>
<td>130</td>
<td>71</td>
<td>172</td>
</tr>
<tr>
<td>35</td>
<td>121</td>
<td>51</td>
<td>133</td>
<td>77</td>
<td>178</td>
</tr>
<tr>
<td>40</td>
<td>147</td>
<td>51</td>
<td>144</td>
<td>81</td>
<td>217</td>
</tr>
</tbody>
</table>

Simple Linear Regression

\[ SBP = 81.54 + 1.22 \times \text{Age} \]
Least Squares Regression Properties

• The sum of the residuals from the least squares regression line is 0 \( \sum (y - \hat{y}) = 0 \)

• The sum of the squared residuals is a minimum (minimized \( \sum (y - \hat{y})^2 \))

• The simple regression line always passes through the mean of the y variable and the mean of the x variable

• The least squares coefficients are unbiased estimates of \( \beta_0 \) and \( \beta_1 \)
Explained and Unexplained Variation

- Total variation is made up of two parts:

$$\text{SST} = \text{SSR} + \text{SSE}$$

where:

- $\bar{y}$ = Average value of the dependent variable
- $y$ = Observed values of the dependent variable
- $\hat{y}$ = Estimated value of $y$ for the given $x$ value
Explained and Unexplained Variation

- **SST = total sum of squares**
  - Measures the variation of the $y_i$ values around their mean $y$

- **SSR = regression sum of squares**
  - Explained variation attributable to the relationship between $x$ and $y$

- **SSE = error sum of squares**
  - Variation attributable to factors other than the relationship between $x$ and $y$
Explained and Unexplained Variation

$\text{SST} = \sum (y_i - \bar{y})^2$

$\text{SSE} = \sum (y_i - \hat{y}_i)^2$

$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2$
The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.

The coefficient of determination is also called R-squared and is denoted as $R^2$.

\[ R^2 = \frac{SSR}{SST} \]

where $0 \leq R^2 \leq 1$
## Coefficient of Determination, $R^2$

<table>
<thead>
<tr>
<th>$R^2 = \frac{SSR}{SST}$</th>
<th>sum of squares explained by regression</th>
<th>total sum of squares</th>
</tr>
</thead>
</table>

**Note:** In the single independent variable case, the coefficient of determination is

$$R^2 = r^2$$

where:

- $R^2 = \text{Coefficient of determination}$
- $r = \text{Simple correlation coefficient}$
Examples of $R^2$ Values

- **$R^2 = 1$**
  - Perfect linear relationship between $x$ and $y$:
  - 100% of the variation in $y$ is explained by variation in $x$
Examples of $R^2$ Values

$0 < R^2 < 1$

Weaker linear relationship between $x$ and $y$:

Some but not all of the variation in $y$ is explained by variation in $x$
Examples of $R^2$ Values

$R^2 = 0$

No linear relationship between $x$ and $y$:

The value of $Y$ does not depend on $x$. (None of the variation in $y$ is explained by variation in $x$)
Standard Error of Estimate

The standard deviation of the variation of observations around the regression line is estimated by

\[ s_\varepsilon = \sqrt{\frac{\text{SSE}}{n - k - 1}} \]

Where
- SSE = Sum of squares error
- n = Sample size
- k = number of independent variables in the model
The Standard Deviation of the Regression Slope

• The standard error of the regression slope coefficient \( (b_1) \) is estimated by

\[
S_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \left(\frac{\sum x}{n}\right)^2}}
\]

where:

- \( S_{b_1} \) = Estimate of the standard error of the least squares slope
- \( s_\varepsilon \) = Sample standard error of the estimate

\[
s_\varepsilon = \sqrt{\frac{SSE}{n - 2}}
\]
Comparing Standard Errors

Variation of observed $y$ values from the regression line

- Small $s_\varepsilon$
- Large $s_\varepsilon$

Variation in the slope of regression lines from different possible samples

- Small $s_{b_1}$
- Large $s_{b_1}$
Inference about the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between x and y?
- Null and alternative hypotheses
  - $H_0$: $\beta_1 = 0$ (no linear relationship)
  - $H_1$: $\beta_1 \neq 0$ (linear relationship does exist)
- Test statistic (df=n-2)
  \[
t = \frac{b_1 - \beta_1}{s_{b_1}}
  \]
  where:
  - $b_1$ = Sample regression slope coefficient
  - $\beta_1$ = Hypothesized slope
  - $s_{b_1}$ = Estimator of the standard error of the slope
Inference about the Slope: t Test

Estimated Regression Equation:

\[
\hat{SBP} = 81.54 + 1.22 \times \text{Age}
\]

The slope of this model is 1.22

Does age of the house affect SBP?
Inference about the Slope: t Test

- Estimated Regression Equation: \( \hat{SBP} = 81.54 + 1.22 \times \text{Age} \)
- Hypothesis:
  - \( H_0: \beta_1 = 0 \) (slope)
  - \( H_1: \beta_1 \neq 0 \)
- Estimate of \( \beta_1 \)
  - 1.22240
- Test Statistic
  - \( t = 5.74 \)
- P-value <0.0001
- Conclude: \( \beta_1 \neq 0 \)
  There is a statistically significant linear trend between age and SBP.
Regression Analysis for Description

Confidence Interval Estimate of the Slope

(95% CI of $\beta$) :

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

$$\text{d.f.} = n - 2$$

At 95% level of confidence, the confidence interval for the slope is (0.79, 1.66)
Confidence Interval for the Average y, Given x

Confidence interval estimate for the mean of y given a particular $x_p$

Size of interval varies according to distance away from mean, $x$

$$
\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}
$$
Confidence Interval for an Individual \( y \), Given \( x \)

Confidence interval estimate for an Individual value of \( y \) given a particular \( x_p \)

\[
\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}
\]

This extra term adds to the interval width to reflect the added uncertainty for an individual case.
Interval Estimates for Different Values of x

Prediction Interval for an individual $y$, given $x_p$

Confidence Interval for the mean of $y$, given $x_p$

$\hat{y} = b_0 + b_1 x$

$y = b_0 + b_1 x$

$x_p$
Residual Analysis

• Purposes
  – Examine for *linearity* assumption
  – Examine for *constant variance* for all levels of x
  – Evaluate *normal* distribution assumption

• Graphical Analysis of Residuals
  – Can plot residuals vs. x
  – Can create histogram of residuals to check for normality
Residual Analysis for Linearity
Residual Analysis for Constant Variance

Non-constant variance vs. Constant variance.
### Multiple Linear Regression

The equation for a multiple linear regression is given by:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \]

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>Predictor variables</td>
</tr>
<tr>
<td>Response variable</td>
<td>Explanatory variables</td>
</tr>
<tr>
<td>Outcome variable</td>
<td>Covariates</td>
</tr>
<tr>
<td>Dependent variable</td>
<td>Independent variables</td>
</tr>
</tbody>
</table>
Multiple Linear Regression

- Relation between a continuous variable and a set of $k$ continuous variables
  \[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \]
- Partial regression coefficients $\beta_p$
  - Amount by which $y$ changes on average when $x_p$ changes by one unit and all the other $x_p$’s remain constant
  - Measures association between $x_p$ and $y$ adjusted for all other $x_p$
- Example
  - SBP vs. age, weight, height, etc
Logistic Regression
Logistic Regression

Table 2. Age and signs of coronary heart disease (CHD)

<table>
<thead>
<tr>
<th>Age</th>
<th>CHD</th>
<th>Age</th>
<th>CHD</th>
<th>Age</th>
<th>CHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
<td>41</td>
<td>1</td>
<td>55</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>46</td>
<td>0</td>
<td>58</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>0</td>
<td>47</td>
<td>0</td>
<td>60</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>49</td>
<td>1</td>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>49</td>
<td>0</td>
<td>65</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>0</td>
<td>50</td>
<td>1</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td>33</td>
<td>0</td>
<td>51</td>
<td>0</td>
<td>71</td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>1</td>
<td>51</td>
<td>1</td>
<td>77</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td>0</td>
<td>52</td>
<td>0</td>
<td>81</td>
<td>1</td>
</tr>
</tbody>
</table>

How to analyze these data?

• Compare mean age of diseased and non-diseased.
  – Diseased       57.7 years
  – Non-diseased   38.6 years
  – Hypothesis:    \( H_0: \mu_1 = \mu_2 \) vs. \( H_1: \mu_1 \neq \mu_2 \)
  – t-test:        \( p<0.0001 \)

• Linear regression?
Scatter plot: Data from Table 2

Signs of coronary heart disease vs. Age (years)
Table 3. Prevalence (%) of signs of CHD according to age group

<table>
<thead>
<tr>
<th>Age group</th>
<th>No. in group</th>
<th>Diseased</th>
<th>No.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30-39</td>
<td>6</td>
<td>1</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>40-49</td>
<td>7</td>
<td>2</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>50-59</td>
<td>7</td>
<td>4</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>60-69</td>
<td>5</td>
<td>4</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>70-79</td>
<td>2</td>
<td>2</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>80-89</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>
Scatter plot: Data from Table 3

Diseased %

Age group
Objective of the Analysis

• How is a risk factor related to disease or death?
• To model the relationship between the risk of the outcome and the predictors of interest, where
  – The outcome is binary (Yes vs No): Y=1 or 0
  – The predictors are X=(X_1, X_2, …, X_p)
  – The risk of the outcome is Pr(Y=1) or E(Y)
• The statistical model is
  \[ Pr(Y=1|X) = E(Y|X) = f(X) \]
Construction of f(X)

• Linear regression: \( E(Y) = f(X) = \alpha + \beta X \)

• Can we use the same model? e.g.,

\[
\Pr(\text{CHD}) = \alpha + \beta \cdot \text{BMI}
\]

– \( \Pr(\text{CHD}) \) is a probability and thus will always be between 0 and 1

– Need to select \( f(X) \) so that it is always between 0 and 1
Logistic Regression Model

\[ \Pr(Y = 1 \mid X) = f(X) = \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}} \]
Logistic Regression

- Relate a binary outcome to one or more predictors.
- Describes the magnitude of effects using odds ratios.
- Let
  
  $Y=1$ for disease, $Y=0$ for non-disease,
  
  Race=1 for white, Race=0 for minority,
  
  Age.
Review: Odds and Odds Ratio

<table>
<thead>
<tr>
<th>Disease</th>
<th>White</th>
<th>Non-White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>No Disease</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Risk: \( \frac{a}{a+c} \quad \frac{b}{b+d} \)

Odds: \( \frac{a}{c} \quad \frac{b}{d} \)

Relative Risk (RR): \( \frac{a}{a+c} \times \frac{b+d}{b} \)

Odds Ratio (OR): Odd for White / Odd for Non-White

\[ = \frac{a/c}{b/d} \]
### Review: Odds and Odds Ratio

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Non-White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>130</td>
<td>101</td>
</tr>
<tr>
<td>No Disease</td>
<td>12,079</td>
<td>12,890</td>
</tr>
<tr>
<td><strong>Risk</strong></td>
<td>130/12,209</td>
<td>101/12,991</td>
</tr>
<tr>
<td><strong>Odds</strong></td>
<td>130/12,079</td>
<td>101/12,890</td>
</tr>
</tbody>
</table>

**Relative Risk (RR)**

\[
RR = \frac{\frac{130}{12,209}}{\frac{101}{12,991}} = 1.37
\]

**Odds Ratio (OR)**

\[
OR = \frac{\text{Odds for White}}{\text{Odds for Non-White}} = \frac{130/12,079}{101/12,890} = 1.37
\]
More on OR

- Ranges from 0 to infinity
- Tends to be skewed (i.e., not symmetric):
  - OR>1: Increased risk when exposed
  - OR=1: no difference in risk btw the two groups
  - OR<1: Lower risk (protective) when exposed
- Log of OR tends to be symmetric and normally distributed.
  - Log (OR) >0: increased risk
  - Log (OR) =0: no difference in risk
  - Log (OR) <0: decreased risk
Logistic Regression Model

\[ \Pr(Y = 1 \mid X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \]

\[ \ln \left( \frac{\Pr(Y = 1 \mid X)}{\Pr(Y = 0 \mid X)} \right) = \beta_0 + \beta_1 X \]

\[ \exp(\beta_1) = \text{Odds Ratio} \]
Interpretation of Coefficient

- Let \( X = 1 \) if exposed and \( X = 0 \) if not exposed
- Let \( \beta \) denote the regression coefficient for \( X \), then \( e^\beta \) is the OR for the outcome, comparing the exposed to the not exposed.

\[
e^\beta = \frac{\Pr(Y = 1 \mid X = 1) / \Pr(Y = 0 \mid X = 1)}{\Pr(Y = 1 \mid X = 0) / \Pr(Y = 0 \mid X = 0)}
\]
Interpretation of Coefficient

• For continuous exposure measurements, \( e^\beta \) is the OR for the outcome, corresponding to one unit increase from the previous level.

\[
e^{\beta} = \frac{\Pr(Y = 1 \mid X = x + 1) / \Pr(Y = 0 \mid X = x + 1)}{\Pr(Y = 1 \mid X = x) / \Pr(Y = 0 \mid X = x)}
\]

• Which previous level? Any level!
Logistic Regression Example

• Applied Logistic Regression (Hosmer & Lemeshow)
• Outcome: CHD recurrence, Yes vs No
• Predictor: Age
• The logistic regression model is

\[
Pr(\text{CHD} \mid \text{Age}) = \frac{e^{\beta_0 + \beta_1 \times \text{Age}}}{1 + e^{\beta_0 + \beta_1 \times \text{Age}}}
\]
Logistic Regression Example

<table>
<thead>
<tr>
<th>id</th>
<th>age</th>
<th>chd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>30</td>
<td>1</td>
</tr>
</tbody>
</table>

...
Logistic Regression Example

\[ \hat{Pr} (\text{CHD} \mid \text{Age}) = \frac{e^{b_0 + b_1 \times \text{Age}}}{1 + e^{b_0 + b_1 \times \text{Age}}} \]

- The estimate of \( b_1 = 0.111 \) with s.e. = 0.024
- The p-value for \( H_0: b_1 = 0 \) is <0.0005
- The 95% conf. interval for \( b_1 \) is (0.064, 0.158)
- OR corresponding to one year increase of age:
  \[ e^{\hat{b}_1} = e^{0.111} = 1.117 \]
  - with the 95% confidence interval of \( (e^{0.064}, e^{0.158}) = (1.066, 1.171) \).
Multiple Logistic Regression

• How to make adjustment for potential confounders?

• In linear regression: \( E(Y|X_1,X_2) = b_0 + b_1 X + b_2 X_2 \)

• In logistic regression:

\[
P(Y \mid X_1, X_2) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}, \quad \text{or}
\]

\[
\text{logit}(p) = \ln \left( \frac{\Pr(Y = 1)}{\Pr(Y = 0)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2
\]
Multiple Logistic Regression

- The exponential of the regression coefficient, $e^{b_1}$ can be interpreted as the OR corresponding to one unit increase in $X_1$, while the level of $X_2$ is fixed at $x_2$.

$$e^{b_1} = \frac{\Pr(Y = 1 \mid X_1 = x_1 + 1, X_2 = x_2) / \Pr(Y = 0 \mid X_1 = x_1 + 1, X_2 = x_2)}{\Pr(Y = 1 \mid X_1 = x_1, X_2 = x_2) / \Pr(Y = 0 \mid X_1 = x_1, X_2 = x_2)}$$
Multiple Logistic Regression Example

- **Dependent variable**
  - whether the patient has had a second heart attack within 1 year (binary: 1= Yes v.s. 0 = No).

- **Two independent variables:**
  - whether the patient completed a treatment consisting of anger control practice (binary: 1=Yes v.s. 0=No).
  - score on a trait anxiety scale (a higher score means more anxious).
Multiple Logistic Regression
Example: Data

<table>
<thead>
<tr>
<th>ID</th>
<th>Heart Attack</th>
<th>Treatment</th>
<th>Anxiety</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>70</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>65</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>75</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>60</td>
</tr>
</tbody>
</table>
Univariate Analysis Result

- The parameter estimate is -1.2528 with a 95% confidence interval of (-3.1048, 0.5992).
- The OR of the second heart attack comparing those received treatment to those did not is 0.2857 with a 95% CI of (0.0448, 1.8207).
Multiple Analysis Result

- The OR for the second heart attack comparing those received treatment to those did not, while controlling for the anxiety level, is 0.36 with a 95% CI of [0.04, 3.56] and a p-value of 0.38.

- Every unit increase in Anxiety score increases the odds of the second heart attack by 12.6% (95% CI: [1.1%, 25.5%], p=0.03) while treatment status remains the same.
Obtaining Risk, Risk Difference, and Relative Risk

• What if we want to know the risk difference or relative risk since odds ratio is too difficult to explain?

• You can still use logistic regression!
Obtaining Risk, Risk Difference, and Relative Risk

\[
\Pr(\text{Heart Attack}) = \frac{e^{(-6.36 - 1.02 \times \text{Treatment} + 0.12 \times \text{Anxiety})}}{1 + e^{(-6.36 - 1.02 \times \text{Treatment} + 0.12 \times \text{Anxiety})}}
\]
Obtaining Risk, Risk Difference, and Relative Risk

• Anxiety level = 50:
  • Pr(Heart attack | Treatment, 50)=0.19
  • Pr(Heart attack | No Treatment, 50)=0.40
  • Risk Difference (RD) = 0.19 - 0.40 = -0.21
  • Relative Risk (RR) = 0.19 / 0.40 = 48%

• Anxiety level = 80:
  • Pr(Heart attack | Treatment, 80)=0.89
  • Pr(Heart attack | No Treatment, 80)=0.96
  • RD = 0.89 - 0.96 = -0.07
  • RR = 0.89 / 0.96 = 93%
How about OR?

• What is the OR for treatment vs. no treatment?
  – Anxiety level = 50
    • $\Pr(\text{Heart attack}|\text{Treatment, 50}) = 0.19$
    • $\Pr(\text{Heart attack}|\text{No Treatment, 50}) = 0.40$
    • $OR = (0.19/0.81) / (0.40/0.60) = 0.36$
  – Anxiety level = 80
    • $\Pr(\text{Heart attack}|\text{Treatment, 80}) = 0.89$
    • $\Pr(\text{Heart attack}|\text{No Treatment, 80}) = 0.96$
    • $OR = (0.89/0.11) / (0.96/0.04) = 0.36$
  – What a surprise: the same OR???
Extensions of Logistic Regression

- Polytomous response
- Exact method
- Conditional logistic regression
- Longitudinal data
Cox Proportional Hazards Regression

• Outcome of interest is amount of time from an initial observation until occurrence of a subsequent event
  – Time from baseline examination until death

• Goal of Cox proportional hazards regression analysis is to examine predictors of time to an event
Next Time

Statistical Genetics: Classical to Modern
Useful References

• *Intuitive Biostatistics*, Harvey Motulsky, Oxford University Press, 1995
  – Highly readable, minimally technical

  – Readable, not too technical

• *Fundamentals of Biostatistics*, Bernard Rosner, Duxbury, 2000
  – Useful reference, somewhat technical
Contact BCC

• Please go to our web site at:
  http://www.feinberg.northwestern.edu/depts/bcc/

• Fill out the online request form for further collaboration.