# Logistic regression

Sergey Tarima, PhD

Sponsored by the Clinical and Translational Science Institute (CTSI)
and the Department of Population Health / Division of Biostatistics

# Speaker Disclosure

# Outline

- Odds, LOGITs and Probabilities on examples
- Simple logistic regression
  - Single binary predictor
  - Single continuous predictor
  - Interpretation of regression coefficients
- Multiple Logistic regression
  - Logistic models
  - Estimation / Inference
  - Logistic model for association test
  - Logistic model for prediction/classification
- Summary

# Example 1.1: 100 participants are randomized to a new or standard treatment (50 subjects to each treatment group).

| Groups | New | Standard | Total |
|---|---|---|---|
| Success | 20 | 10 | 30 |
| Failure | 30 | 40 | 70 |
| Total | 50 | 50 | 100 |

Are chances of success equal for each treatment choice?

# Example 1.1: (cont)

How to measure the chances of success?

1) The probability of success:

$P_{new}$ = Pr ( Success | new treatment) = 20/50 = 40%

P

# Example 1.1: (cont)

Odds Ratio (OR) is a possible way to capture inequality in the chances of success:

- OR = $O_{new}/O_{st}$ = (20/30)/(10/40) = 0.67/0.25=2.67
- Obviously the odds ratio is just a RATIO OF ODDS ☺ (between the new and standard treatment groups)
- If OR =1 then the success chances are the same in each group, which means $P_{new} = P_{st}$ or $O_{new}=O_{st}$.
- In our case, obviously, the odds of success are 2.67

# Example 1.2a (independence): How does "no difference" in treatment success rates look? (one variant)

| Groups | New | Standard | Total |
|---|---|---|---|
| Success | 20 | 20 | 40 |
| Failure | 30 | 30 | 60 |
| Total | 50 | 50 | 100 |

In this case $P_{new} = P_{st} = 50\%$, and $O_{new}=O_{st}=1$, and OR=1

# Example 1.2b (independence): How does "no difference" in success rates look? (another variant)

| Groups | New | Standard | Total |
|---|---|---|---|
| Success | 10 | 10 | 20 |
| Failure | 40 | 40 | 80 |
| Total | 50 | 50 | 100 |

In this case $P_{new} = P_{st} = 20\%$, and $O_{new}=O_{st}=0.25$, OR=1

# Simple logistic regression

- The probability of success can be represented via odds or LOGITs of success.
- From Example 1.1,
  $\log(O_{new}) = -0.41$ and $\log(O_{st}) = -1.39$, so the *difference between the log odds is equal* to 0.98.
- We can combine these two log odds for different groups into one formula:

  log(odds) = -1.39 + 0.98*(treatment is new)

  (this is an example of a simple logistic regression)

# Simple logistic regression (cont)

- LOGIT=log(odds) = -1.39 + 0.98*(treatment is new)
- In this logistic regression -1.39 and 0.98 are regression coefficients…
- -1.39 is called the model intercept
- 0.98 is the treatment effect

- It is important to understand the “connection” between the regression coefficients and probabilities of success

# Simple logistic regression (cont)

- If we apply antilog to 0.98 then exp(0.98)=2.67, the odds ratio!!!
- This 2.67 is different from 1, which means we have a significant increase in odds of treatment success (chi-square O-value was < 5%)

# Example 2: Coronary Heart Disease (CHD)

- Risk factors for CHD include gender, age, smoking, high blood pressure, high cholesterol, obesity, etc.
- First we look at AGE as a continuous predictor

# Age and CHD

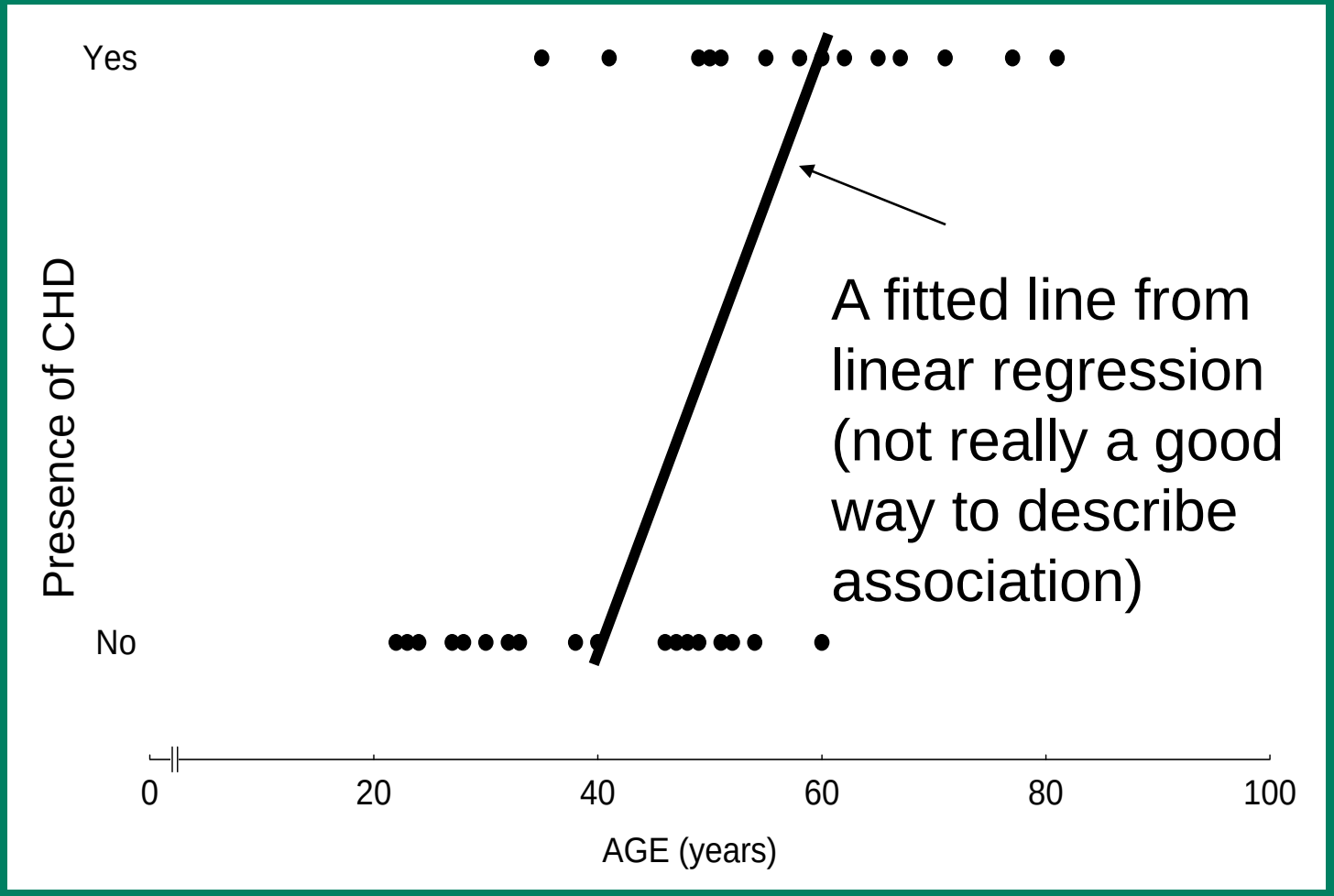
**Table 1. Age and coronary heart disease**

| Age | CD | Age | CD | Age | CD |
|---|---|---|---|---|---|
| 22 | 0 | 40 | 0 | 54 | 0 |
| 23 | 0 | 41 | 1 | 55 | 1 |
| 24 | 0 | 46 | 0 | 58 | 1 |
| 27 | 0 | 47 | 0 | 60 | 1 |
| 28 | 0 | 48 | 0 | 60 | 0 |
| 30 | 0 | 49 | 1 | 62 | 1 |
| 30 | 0 | 49 | 0 | 65 | 1 |
| 32 | 0 | 50 | 1 | 67 | 1 |
| 33 | 0 | 51 | 0 | 71 | 1 |
| 35 | 1 | 51 | 1 | 77 | 1 |
| 38 | 0 | 52 | 0 | 81 | 1 |

# How to Describe Association?

- Age is a continuous variable
- Compare mean age between diseased and non-diseased groups
  - Non-diseased (38.6 years) vs. diseased (58.7 years) $\Rightarrow p<0.0001$
  - Not informative to assess the magnitude of age effect
- Look at the relationship between age and the presence of CHD

# A Dot-Plot


Yes
No
Presence of CHD
A fitted line from linear regression (not really a good way to describe association)
0
20
40
60
80
100
AGE (years)

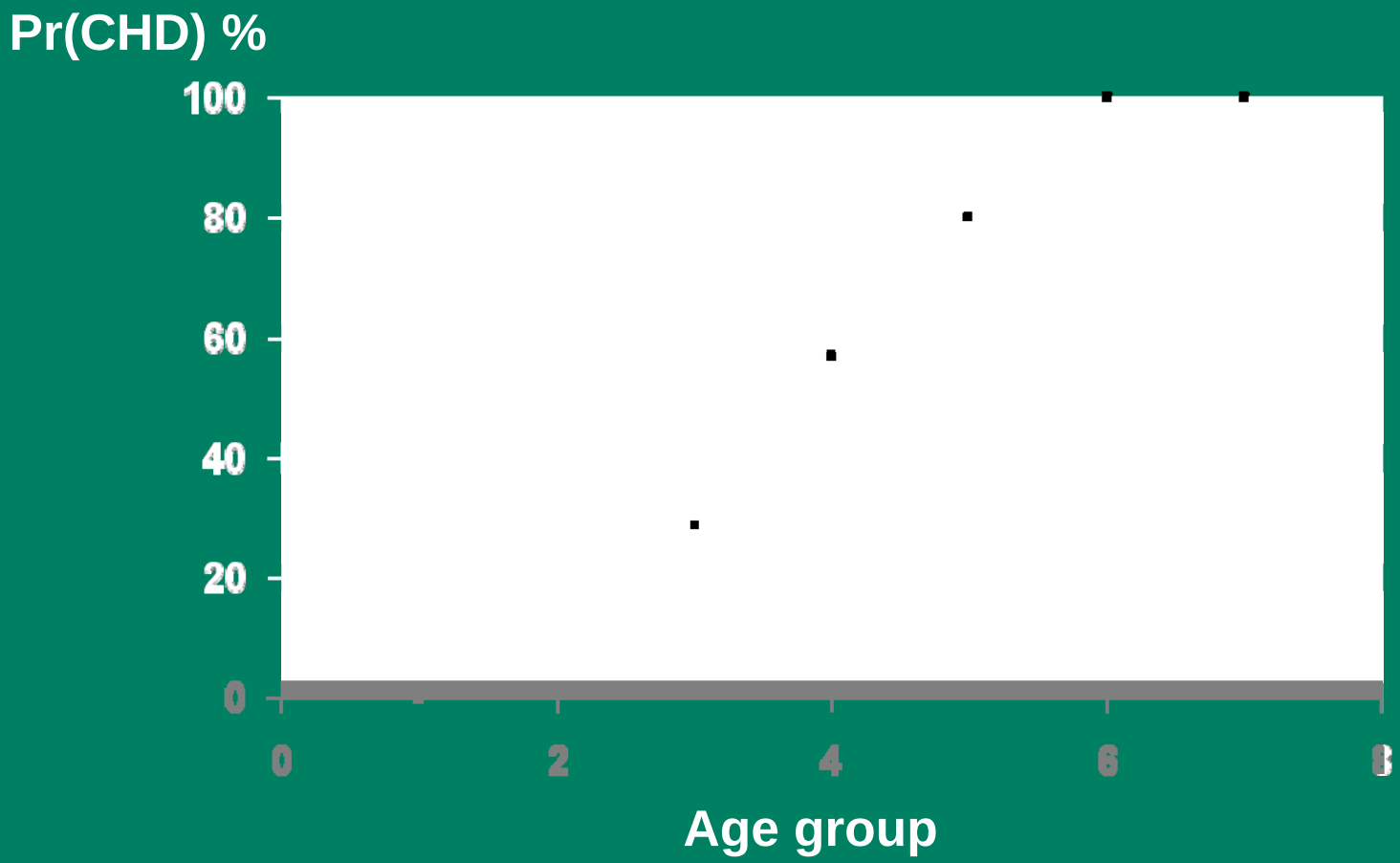
# Other Options

- When the outcome variable is binary (like the presence or absence of CHD), it makes more sense to consider *probability of having the disease* at different ages
- Categorize age into multiple groups
- Look at presence/absence of disease in each age group

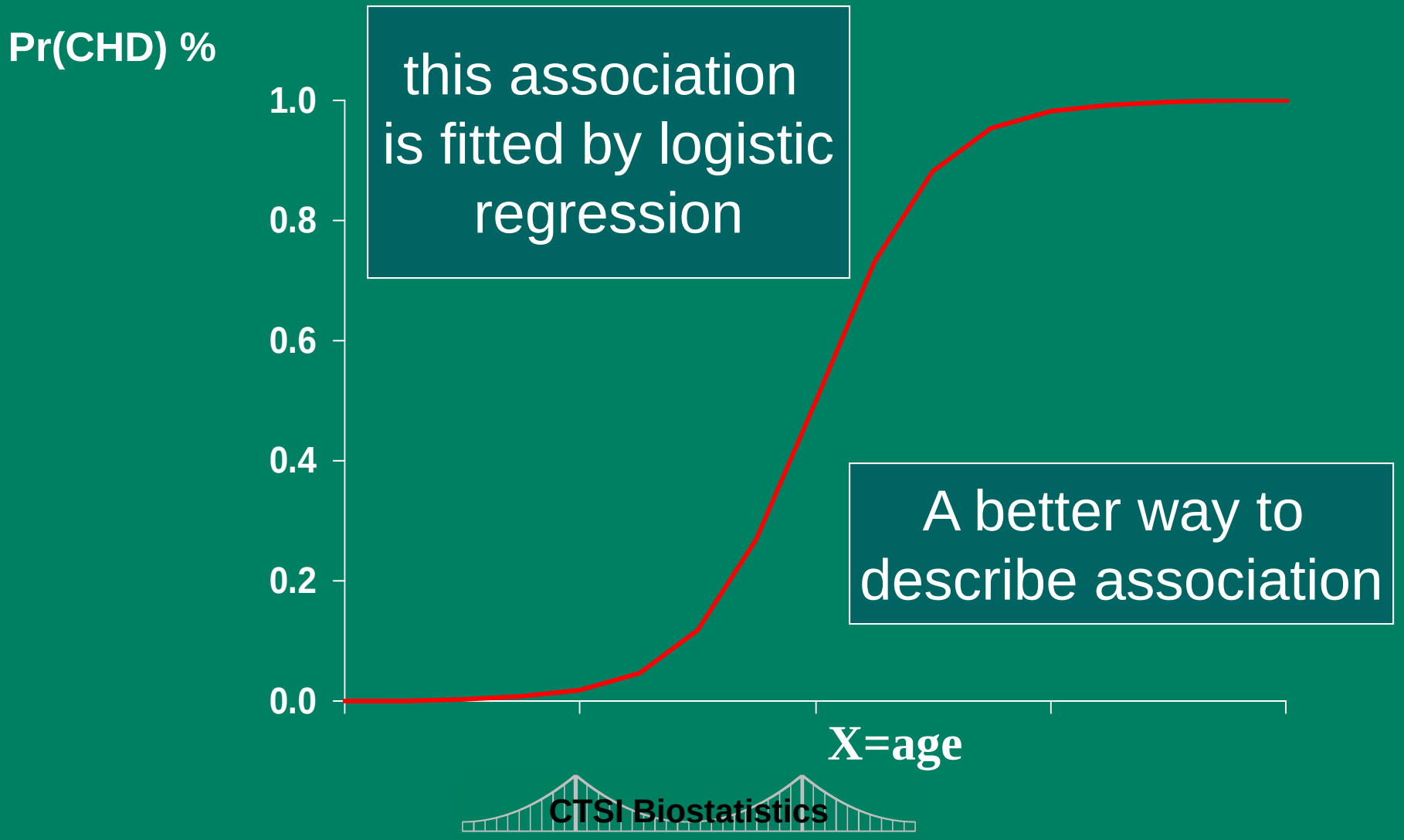
# Table. Presence (%) of CHD in different age groups

**Diseased**

# Dot plot of CHD presence (%) in different age groups



# Logistic Curve



# Interpretation of Parameters

- When I fit logistic regression (in SAS) for CHD data I have the following output:

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| *Intercept* | *1* | *-6.5820* | *2.3121* | *8.1038* | *0.0044* |
| *AGE* | *1* | *0.1309* | *0.0458* | *8.1557* | *0.0043* |

- This output leads to the following

  LOGIT(CHD)= -6.5820 + 0.1309*AGE

# Interpretation of Parameters (cont)

Note, exp(0.1309)=1.14 is also an odds ratio, but this odds ratio describes % increase in odds when AGE increases by 1 year.

For example,

(1) at AGE=40 the ODDS of CHD were 0.26, then at AGE=41 the ODDS=0.26*1.14=0.296,

(2) at AGE=60 the ODDS of CHD were 3.56, then at AGE=61 the ODDS=3.56*1.14=4.06,

(3) at AGE=60 the ODDS of CHD were 3.56, then at AGE=59 the ODDS=3.56/1.14=3.13

# Assumptions

- Independent observations
- A linear relationship between LOGIT of CHD and AGE

What if we do not have a linear relationship between the LOGIT of CHD and AGE???

In this case we can
(only one of possible solutions)
use a categorization of AGE

# Age and CHD

**Age**

# Two by three table (after AGE categorization)

- Frequency
  Percent
  Row Pct
  Col Pct

| | 20-39 | 40-53 | 54-90 | Total |
|---|---|---|---|---|
| NO | 10 | 6 | 2 | 18 |
| | 31.25 | 18.75 | 6.25 | 56.25 |
| | 55.56 | 33.33 | 11.11 | |
| | 90.91 | 60.00 | 18.18 | |
| YES | 1 | 4 | 9 | 14 |
| | 3.13 | 12.50 | 28.13 | 43.75 |
| | 7.14 | 28.57 | 64.29 | |
| | 9.09 | 40.00 | 81.82 | |
| Total | 11 | 10 | 11 | 32 |
| | 34.38 | 31.25 | 34.38 | 100.00 |

# Odds and LOGITs for the categorized AGE

- Odds of CHD in 20-39 group = 1/10
- Odds of CHD in 40-53 group = 4/6
- Odds of CHD in 54-90 group = 2/2

- LOGIT of CHD in 20-39 group = log(1/10)
- LOGIT of CHD in 40-53 group = log(4/6)
- LOGIT of CHD in 54-90 group = log(2/2)

# Multiple Logistic Regression (SAS output)

- **Analysis of Maximum Likelihood Estimates**

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -2.3026 | 1.0488 | 4.8199 | 0.0281 |
| age1 | 1 | 1.8971 | 1.2315 | 2.3730 | 0.1235 |
| age2 | 1 | 3.8066 | 1.3081 | 8.4686 | 0.0036 |

This output leads to the following:

LOGIT = -2.3026 + 1.8971*(39<age<=53) + 3.8066*(age > 53)

# Interpretation of Parameters (categorical AGE)

| AGE Cat | LOGIT | ODDS | PROBABILITY |
|---|---|---|---|
| 20-39 | -2.3036 | 0.0999 | 0.0908 |
| 40-53 | -0.4055 | 0.6666 | 0.4000 |
| 54-90 | 1.5040 | | |

# Further Improvement

- We have seen how two predictors can be incorporated in the model
- Risk factors for CD include gender, age, smoking, high blood pressure, high cholesterol, obesity, etc.
- A model including multiple risk factors
  - Adjusts for other risk factors
  - Provides better prediction

# Example 3: low birth weight data

(Hosmer & Lemeshow "Applied Logistic

Goal: to identify risk factors associated with lower birth weight (variable "low")

Dataset: 189 women (59 lower birth weight babies, and 130 – normal weight babies)

Possible Risk Factors: age ( "AGE"), subject's weight ( "LWT"), race ("race2" and "race3"), and the number of physician visits ("FTV")

# The SAS output (all four predictors):

Analysis of Maximum Likelihood Estimates

| Parameter | Estimate | Standard Error | Chi-Square | Pr >ChiSq |
|---|---|---|---|---|
| Intercept | **1.2953** | **1.0714** | **1.4616** | **0.2267** |
| AGE | **-0.0238** | **0.0337** | **0.4988** | **0.4800** |
| LWT | **-0.0142** | **0.0065** | **4.7428** | **0.0294** |
| race2 | **1.0039** | **0.4979** | **4.0660** | **0.0438** |
| race3 | **0.4331** | **0.3622** | **1.4296** | **0.2318** |
| FTV | **-0.0493** | **0.1672** | **0.0869** | **0.7681** |

Here “race2” and “race3” are indicators that RACE=2 and RACE=3 (the race categories were enumerated in the data as 1, 2 and 3)

# The SAS output (excluding AGE and FTV):

Standard

# Multiple Logistic Regression Objectives:

- To find significant predictors (risk or protective factors)
- To build a predictive model for predicting the LOGIT
- To control for effect of significant predictors (risk factors)… a way to eliminate confounding effects

# Prediction & Classification

- Prediction: From the fitted model, a predicted *probability* can be computed for each set of predictors
- Classification: If the predicted probability exceeds some cut-off point, the observation is predicted to be an *event* observation; otherwise, it is predicted as a *nonevent*.

# Summary

Logistic regression

- deals with binary outcomes

- allows multiple predictor variables, which can be continuous, categorical or ordinal

- provides estimates of adjusted odds ratios

# Resources

- The Clinical and Translation Science Institute (CTSI) supports education, collaboration, and research in clinical and translational science: www.ctsi.mcw.edu
- The Biostatistics Consulting Service provides comprehensive statistical support

  http://www.mcw.edu/biostatsconsult.htm

# Free drop-in consulting

- Froedtert:
  - Monday, Wednesday, Friday 1 – 3 PM
  - Location: Pavilion, LL772A – TRU offices
- MCW:
  - Tuesday, Thursday 1 – 3 PM
  - Location: Health Research Center, H2400
- VA:
  - 1st and 3rd Monday, 8:30-11:30 am
  - VA Medical Center, 111-B-5423
- Marquette:
  - Tuesday 9 – 5 PM
  - School of Nursing, Clark Hall