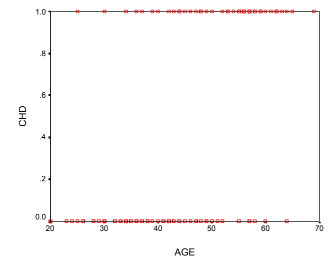
Why do we use Logistic Regression?

- Binary dependent variable
- Several independent variables
  - too many to stratify
  - want to assess role of suspected cause and confounding factors including EM
- Provide simple, interpretable result (inference)

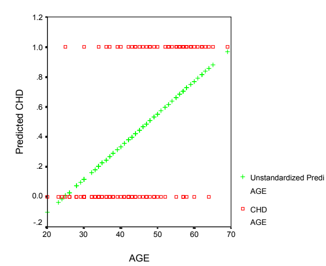
Example: Plot of CHD vs. Age



Example: Interpretation

- Plot of binary values
  - Hard to summarize
  - Appears that 0's are younger than 1's
  - Large variability at all ages
  - Overall relationship unclear

Example: Linear Regression



Example: Interpretation of Linear Regression

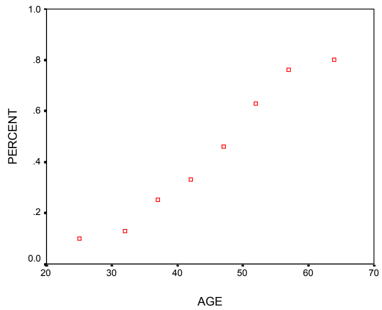
- Probability of CHD increases with subject's age
  - Probability is less than 0 for those under 25
  - Probability is greater than 1 for those over 70
- Substantive interpretation problematic if probability is less than 0 or greater than 1

Example: Grouped Data

- Plot means for 5 or 10 year age groups

Age Group	N	CHD Absent	CHD Present	Proportion
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
Total	100	57	43	0.43

Example: Plot of Mean CHD Risk vs. Age



Example: Interpretation

- As age increases, proportion increases
- Note
  - CHD (y) ranges between 0 and 1
  - Relationship is "s-shaped"
  - As age gets large, incremental change in CHD decreases
  - Errors are binomial (not normally) distributed

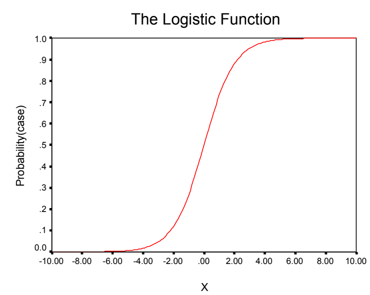
Example: Apply Logit Model

$$odds = \frac{\pi(x)}{1-\pi(x)}; where \quad \pi(x) = \Pr(case)$$

$$\ln(odds) = \ln[\frac{\pi(x)}{1-\pi(x)}] = \beta^0 + \beta^1 x$$

$$solving, \pi(x) = \frac{1}{\;{}_{-x}} = \frac{e^{x}}{\;{}_{x}}$$

A Plot of the Logit Model



Example: Logistic Model Fit

$$\ln(odds) = \ln[\frac{\pi(x)}{\;}] = \beta^{0} + \beta^{1} x$$

- Variable / Coefficient / t-score
  - age / 0.11 / 4.6
  - constant / -5.3 / -4.7
- -2 ln likelihood = 107.4
- odds ratio = 1.12 per year

Variable	Coefficient	t-score
age	0.11	4.6
constant	-5.3	-4.7

Example: Logistic Equation

$$\ln(odds) = -5.31 + 0.11 * age$$

$$OR(\Delta age = 1) = \frac{odds(x = age)}{odds(x = age - 1)}$$

$$OR(\Delta age = 1) = \frac{e^{-5.31+0.11*1}}{e^{-5.31+0.11*0}} = e^{0.11} = 1.12$$

$$OR(\Delta age = 10) = \frac{e^{-5.31+0.11*10}}{e^{-5.31+0.11*0}} = e^{1.1} = 3.0$$

Logistic Model Limitations

- Model is linear (i.e., loglinear)
- Odds ratio is constant
- Linear change in x results in multiplicative change in y (effect)
  - size of effect is determined by coefficient
- Intercept is usually ignored (nuisance)
  - intercept is log odds of disease if x=0

Hypothesis Testing--Overview

- Goal: Assess the role of chance
- Strategy
  - Can we reject the null hypothesis (i.e., hypothesis of no association)?
  - If so, what is the most likely alternative?
- Errors
  - false positive (or alpha or Type I)
  - false negative (or beta or Type II)

Hypothesis Testing--Errors--1

- False Positive
  - say it is true when it is false
  - reject null hypothesis
  - typically use 1 in 20 (0.05) as guideline
  - typically consider two-tailed distribution

Hypothesis Testing--Errors--2

- False Negative
  - say it is false when it is true
  - accept null hypothesis
  - typically use 1 in 20 (0.05) as guideline
  - typically consider two-tailed distribution
  - Pr(Type II error)=1-Power
    - power is the ability to detect an effect given that it is present in the data

Limitations of Hypothesis Testing

- Arbitrary cutpoint (e.g., 0.05)
  - is 0.049 really different than 0.051?
- No measure of effect
  - p-values do not correspond to ORs or RRs
- No measure of sample size
  - the number of subjects can have a large effect
- Transformation of continuous result into a dichotomous result

Alternative to Hypothesis Testing

- Confidence Intervals
  - Definition
    - all parameter values within range are compatible with the data under the standard interpretation of statistical significance testing
    - contains true value x% of the time
  - Properties
    - combine effect size and sample size
    - measure of precision of estimate
    - can be used to assess null hypothesis

Multiple Comparisons

- Traditional p-value 0.05 (1 in 20)
- If there is not effects and:
  - If we conduct 100 studies, 5 statistically significant
  - If we conduct 100 tests, 5 statistically significant
  - We usually report mainly the positive test results—FALSE POSITIVES
- Options
  - Must we report all tests (incl. all cutpoints)?
  - Should we report each test in a separate publication?
  - Should we adjust p-values for all tests?
  - Should we calculate a joint distribution for all parameters?

Review of Linear Regression

- The Data
  - one dependent variable (Y)
  - several independent variables (X's)
  - error distribution is normal
- The Model
  - What is the unit change in Y for each unit change in any of the X's?