Introduction to Logistic Regression

Multivariate analysis

- Multiple models
  - Linear regression
  - Logistic regression
  - Cox model
  - Poisson regression
  - Loglinear model
  - Discriminant analysis
  - ……
- Choice of the tool according to the objectives, the study, and the variables

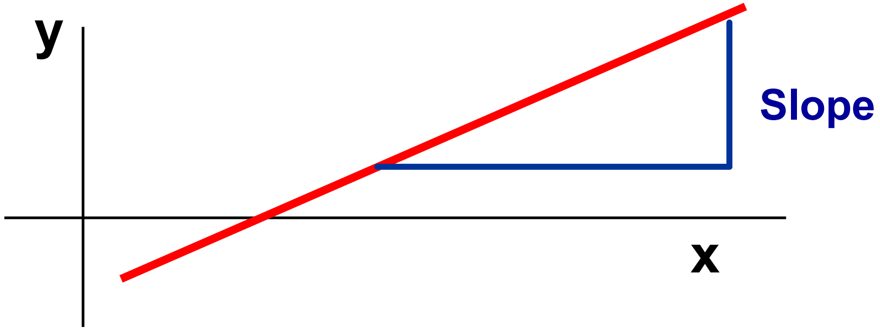
Simple linear regression

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

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

Simple linear regression

- Relation between 2 continuous variables (SBP and age)



$$y = \alpha + \beta_1 x_1$$

- Regression coefficient β_1
  - Measures association between y and x
  - Amount by which y changes on average when x changes by one unit
  - Least squares method

Multiple linear regression

- Relation between a continuous variable and a set of i continuous variables

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

- Partial regression coefficients β_i
  - Amount by which y changes on average when x_i changes by one unit and all the other x_is remain constant
  - Measures association between x_i and y adjusted for all other x_i

- Example
  - SBP *versus* age, weight, height, etc

Multiple linear regression

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i$$

Predicted	**Predictor variables**
Response variable	**Explanatory variables**
Outcome variable	**Covariables**
Dependent	**Independent variables**

Logistic regression (1)

Table 2 Age and signs of coronary heart disease (CD)

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

How can we analyse these data?

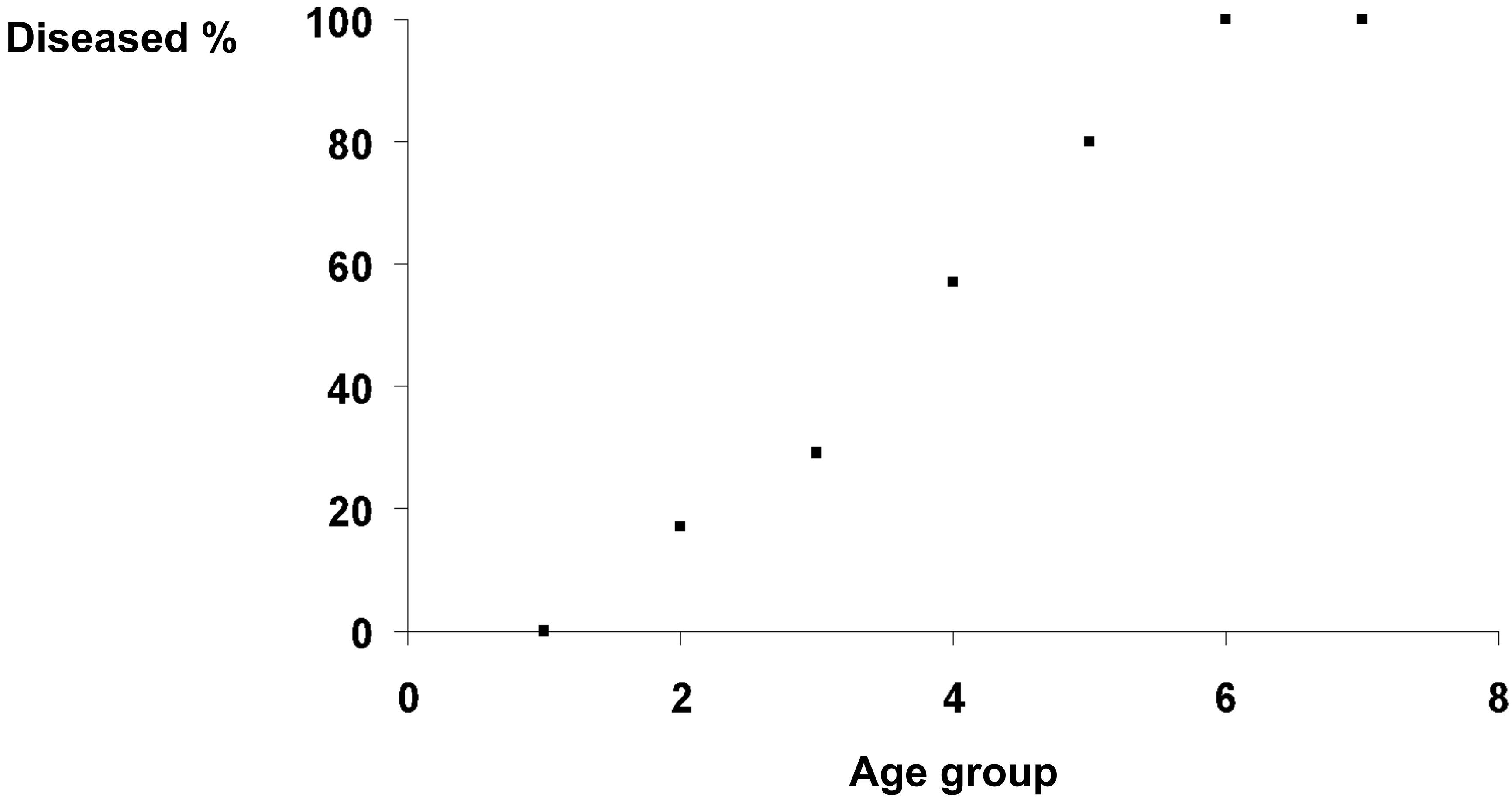
- Compare mean age of diseased and non-diseased
  - Non-diseased: 38.6 years
  - Diseased: 58.7 years ($p<0.0001$)
- Linear regression?

Logistic regression (2)

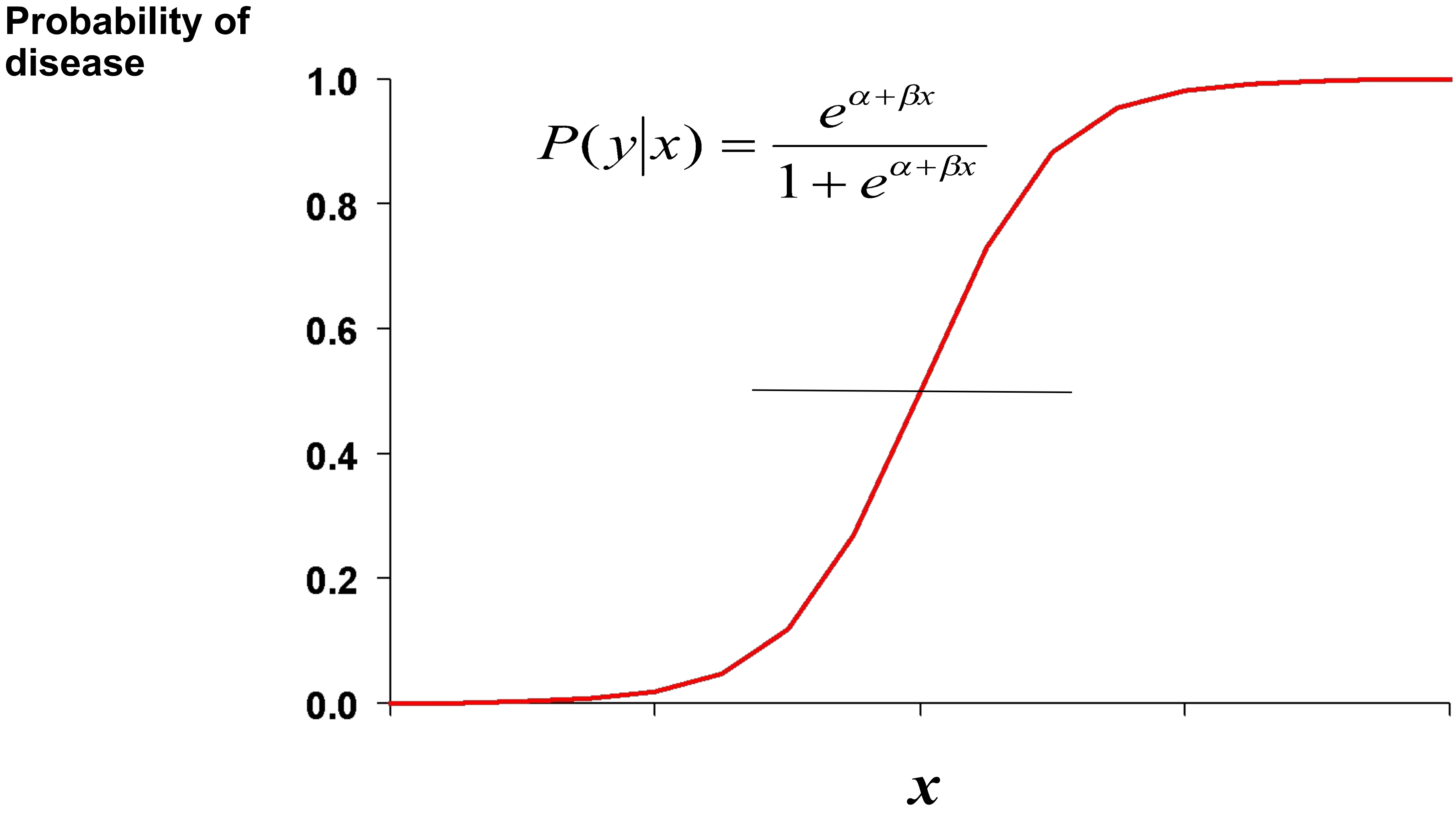
Table 3 Prevalence (%) of signs of CD according to age group

		Diseased	
Age group	# in group	#	%
20 - 29	5	0	0
30 - 39	6	1	17
40 - 49	7	2	29
50 - 59	7	4	57
60 - 69	5	4	80
70 - 79	2	2	100
80 - 89	1	1	100

Dot-plot: Data from Table 3



Logistic function (1)



$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

Transformation

$$P(y|x)=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\frac{P(y|x)}{1-P(y|x)}$$

$$\underbrace{\ln\left[\frac{P(y|x)}{1-P(y|x)}\right]}_{\text{logit of } P(y|x)}=\alpha+\beta x$$

- ✔α = log odds of disease in unexposed
- ✔β = log odds ratio associated with being exposed
- ✔e^{β} = odds ratio

Fitting equation to the data

- Linear regression: Least squares
- Logistic regression: Maximum likelihood
- Likelihood function
  - Estimates parameters α and β
  - Practically easier to work with log-likelihood

$$L(\mathrm{B}) = \ln\left[l(\mathrm{B})\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i)\ln\left[1 - \pi(x_i)\right] \right\}$$

Maximum likelihood

- Iterative computing
  - Choice of an arbitrary value for the coefficients (usually 0)
  - Computing of log-likelihood
  - Variation of coefficients' values
  - Reiteration until maximisation (plateau)

- Results
  - Maximum Likelihood Estimates (MLE) for α and β
  - Estimates of P(y) for a given value of x

Multiple logistic regression

- More than one independent variable
  - Dichotomous, ordinal, nominal, continuous …

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_i x_i$$

- Interpretation of β_i
  - Increase in log-odds for a one unit increase in x_i with all the other x_is constant
  - Measures association between x_i and log-odds adjusted for all other x_i