Introduction to Logistic Regression

Content

- **Simple logistic regression**
  - **The logistic function**
  - **Estimation of parameters**
  - **Interpretation of coefficients**
- **Multiple logistic regression**
  - **Interpretation of coefficients**
  - **Coding of variables**

Logistic regression

- **Models relationship between set of variables x_i**
  - **dichotomous (yes/no)**
  - **categorical (social class, ...)**
  - **continuous (age, ...)**

and

  - **dichotomous (binary) variable Y**

- **Dichotomous outcome most common situation in biology and epidemiology**

Example

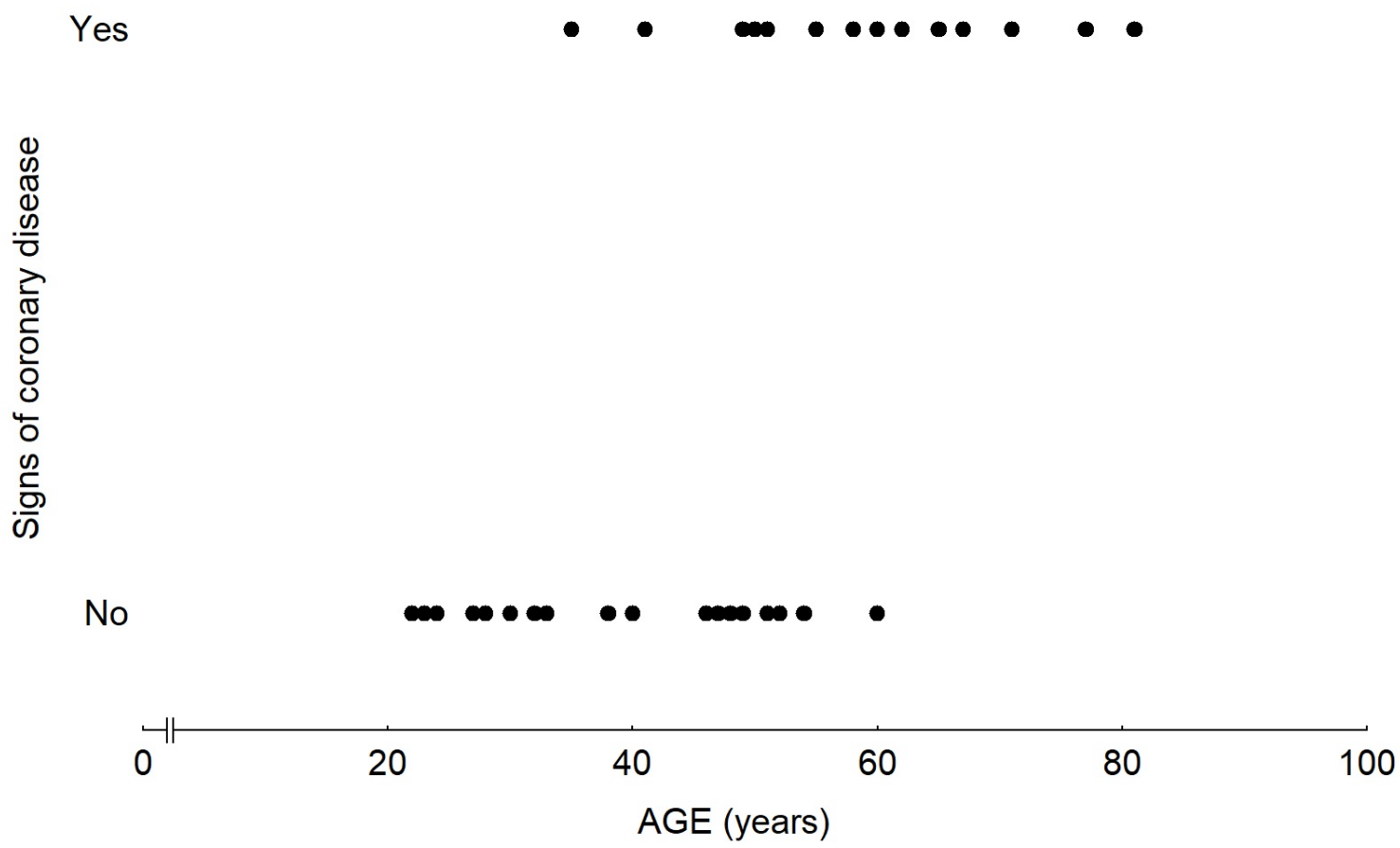
Table 1 Age and signs of coronary heart disease (CD)

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 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?**

Dot-plot: Data from Table 1

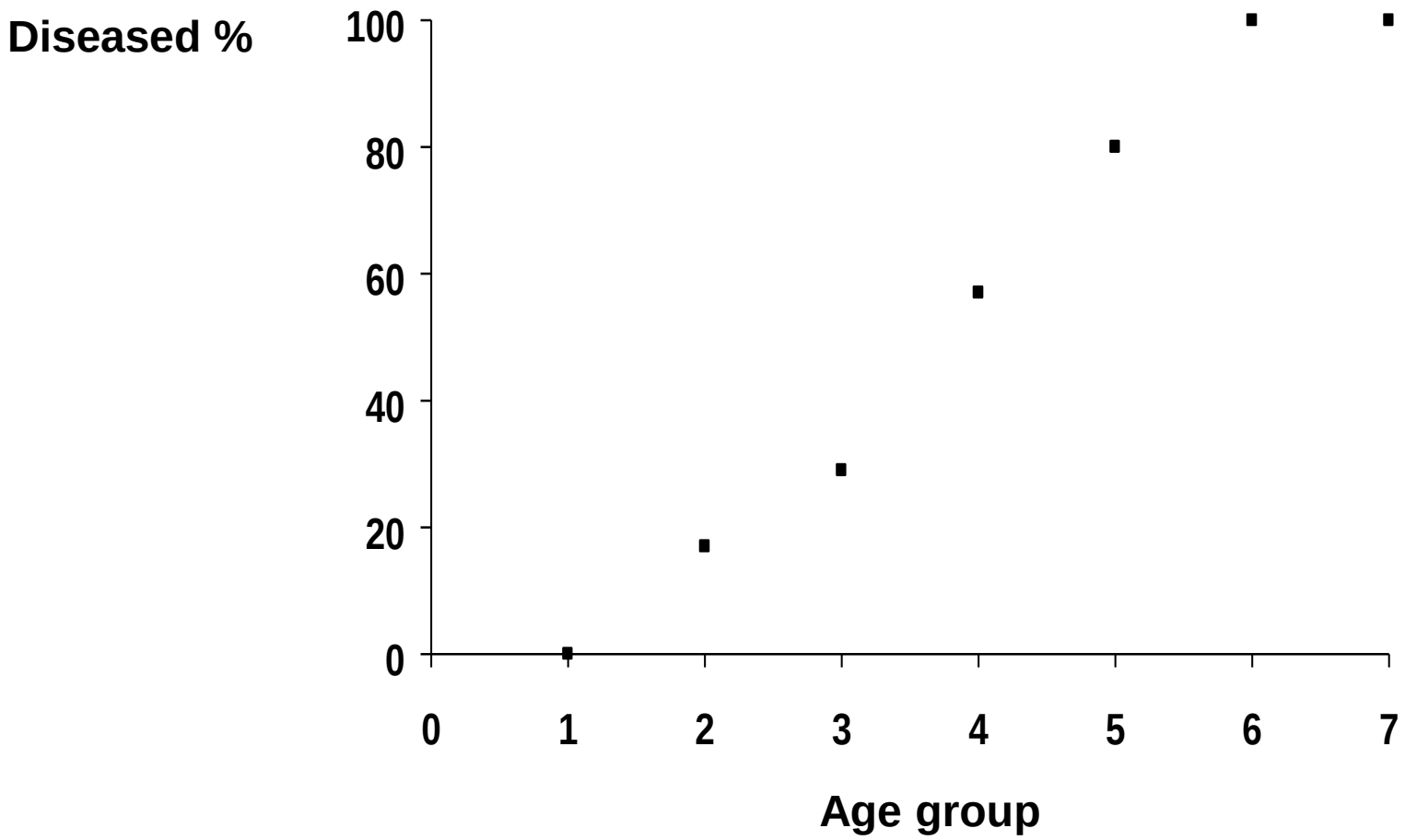


Example cont.

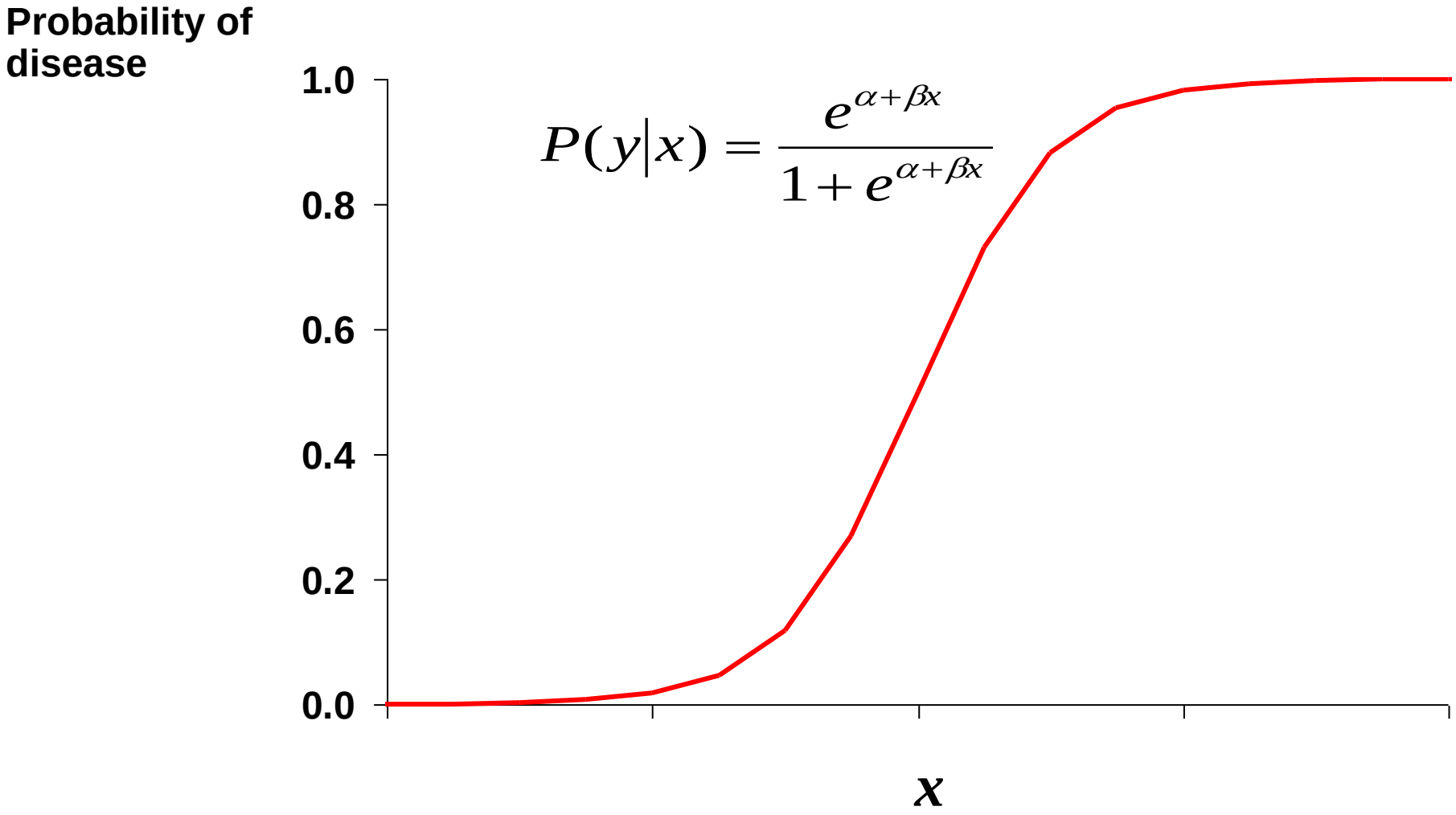
Table 2 Prevalence (%) of signs of CD according to age group

		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 2



Logistic function



Logit transformation

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

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

logit of $P(y|x)$

$$\ln\left(\frac{P}{1\text{-}P}\right) = \alpha + \beta x \quad \longrightarrow \quad \frac{P}{1\text{-}P} = e^{\alpha+\beta x}$$

Interpretation of coefficient β

	Exposure x	
Disease y	**yes**	**no**
yes	$P(y=1\|x=1)$	$P(y=1\|x=0)$
no	$1-P(y=1\|x=1)$	$1-P(y=1\|x=0)$

$$\frac{P}{1\text{-}P} = e^{\alpha+\beta x}$$

$$Odds_{d|e} = e^{\alpha+\beta}$$

$$Odds_{d|\bar{e}} = e^{\alpha}$$

$$OR = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$$

$$\ln(OR) = \beta$$

Interpretation of coefficient β

- β= increase in logarithm of odds ratio for a one unit increase in x
- Test of the hypothesis that β=0 (Wald test)

$$\chi^2 = \frac{\beta^2}{\text{Variance }(\beta)} \qquad (1\,\text{df})$$

- Interval testing

$$95\%\ \text{CI} = e^{(\beta \pm 1.96 SE_\beta)}$$

Example

- **Risk of developing coronary heart disease (CD) by age (<55 and 55+ years)**

	CD	
Age	**Present (1)**	**Absent (0)**
55+ (1)	**21**	**6**
< 55 (0)	**22**	**51**

Odds of disease among exposed = 21/6
Odds of disease among unexposed = 22/51

Odds ratio = 8.1

Logistic Regression Model

$$\ln\left(\frac{P}{1\text{-}P}\right) = \alpha + \beta_1 \times Age \quad = \quad \text{-}0.841 \; + \; 2.094 \times Age$$

	Coefficient	SE	Coeff/SE
Age	2.094	0.529	3.96
Constant	-0.841	0.255	-3.30

$$OR = e^{2.094} = 8.1$$

$$\text{Wald Test} = 3.96^2 \text{ with 1df } (p < 0.05)$$

$$95\%\ CI = e^{(2.094 \pm 1.96 \times 0.529)} = 2.9, 22.9$$

Fitting equation to the data

- **Linear regression: Least squares**
- **Logistic regression: Maximum likelihood**
- **Likelihood function**
  - **Estimates parameters α and β with property that likelihood (probability) of observed data is higher than for any other values**
  - **Practically easier to work with log-likelihood**

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

Multiple logistic regression

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

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

- **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**

Effect modification

- **Effect modification**
  - **Can be modelled by including interaction terms**

$$\ln\left(\frac{P}{1\text{-}P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2$$

Statistical testing

- **Question**
  - **Does model including given independent variable provide more information about dependent variable than model without this variable?**
- **Three tests**
  - **Likelihood ratio statistic (LRS)**
  - **Wald test**
  - **Score test**

Likelihood ratio statistic

- **Compares two nested models**

  Log(odds) = $\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ **(model 1)**

  Log(odds) = $\alpha + \beta_1 x_1 + \beta_2 x_2$ **(model 2)**

- **LR statistic**

  -2 log (likelihood model 2 / likelihood model 1) =

  -2 log (likelihood model 2) ***minus*** **-2log (likelihood model 1)**

  LR statistic is a χ^2 **with DF = number of extra parameters in model**

Example

P	**Probability for cardiac arrest**
Exc	**1= lack of exercise, 0 = exercise**
Smk	**1= smokers, 0= non-smokers**

$$\ln\left(\frac{P}{1\text{-}P}\right) = \alpha + \beta_1\ \text{Exc} \ + \ \beta_2\ \text{Smk}$$

$$= \ 0.7102 \ + \ 1.0047\ \text{Exc} \ + \ 0.7005\ \text{Smk}$$

$$(\text{SE}\ 0.2614) \qquad (\text{SE}\ 0.2664)$$

OR for lack of exercise $= \mathbf{e}^{1.0047} =$ **2.73** *(adjusted for smoking)*

95% CI $= \mathbf{e}^{(1.0047 \pm 1.96 \text{ x } 0.2614)}$ $=$ **1.64 - 4.56**

- **Interaction between smoking and exercise?**

$$\ln\left(\frac{P}{1\text{-}P}\right) = \alpha + \beta_1\, Exc + \beta_2\, Smk + \beta_3\, Smk \times Exc$$

- **Product term β_3 = -0.4604 (SE 0.5332)**

  Wald test = 0.75 (1df)

  -2log(L) = 342.092 with interaction term
  = 342.836 without interaction term

  $\Rightarrow$ LR statistic = 0.74 (1df), p = 0.39
  $\Rightarrow$ No evidence of any interaction

Coding of variables (1)

- **Dichotomous variables: yes = 1, no = 0**
- **Continuous variables**
  - **Increase in OR for a one unit change in exposure variable**
  - **Logistic model is multiplicative $\Rightarrow$ OR increases exponentially with x**
    - **If OR = 2 for a one unit change in exposure and x increases from 2 to 5: OR $= 2 \times 2 \times 2 = 2^3 = 8$**

Coding of variables (2)

- **Nominal variables or ordinal with unequal classes:**
  - **Tobacco smoked: no=0, grey=1, brown=2, blond=3**
  - **Model assumes that OR for blond tobacco = OR for grey tobacco^3**
  - **Use indicator variables (dummy variables)**

Indicator variables: Type of tobacco

Tobacco consumption	Dummy variables		
	Dark	Light	Both
Dark	1	0	0
Light	0	1	0
Both	0	0	1
None	0	0	0

- Neutralises artificial hierarchy between classes in the variable "type of tobacco"
- No assumptions made
- 3 variables (3 df) in model using same reference
- OR for each type of tobacco adjusted for the others in reference to non-smoking

Low Birth Weight Study

- 189 observations
- Low Birth Weigth
  - yes = birth weight < 2500g
  - no = birth weight >2499g
- Age of mother in years
- Weight of mother in pounds
- Race (1,2,3)
- Number of doctor's visit in last trimester

LBW
Age
Weight
Race
Visits

Risk of death from bacterial meningitis according to treatment

- 161 observations
- Death (yes, no)
- Treatment
  - 1=Chloramphenicol, 2=Ampicillin
- Delay before treatment (onset, in days)
- Convulsions (1,0)
- Level of consciousness (1-3)
- Severity of dehydration (1-3)
- Age in years
- Pathogen
  - 1 Others, 2 HiB, 3 Streptococcus pneumoniae

Reference

- **Hosmer DW, Lemeshow S. Applied logistic regression. Wiley & Sons, New York, 1989**