INTREC - INDEPTH Training & Research Centres of Excellence

Lecture 7

Logistic regression

Multivariate analysis

Model	Outcome
Linear regression	continuous
Poisson regression	counts
Cox model	survival
Logistic regression	binomial
……	

- Choice of the tool according to study, objectives and the chosen variables
  - Control of confounding
  - Model building, prediction

Logistic regression

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

and

  - dichotomous variable Y

- Dichotomous (binary) outcome most common situation in biology and epidemiology
  -> Thus, logistic regression is the most common study design used in epidemiology

Logistic Regression (*ctnd*)…..

- logistic regression estimates for a randomly selected individual the probability that an event occurs (***p***) versus the probability that the event does not occur (***1-p***)

- needs a yes/no outcome variable for each individual in the data set (i.e. binary) → case-control study

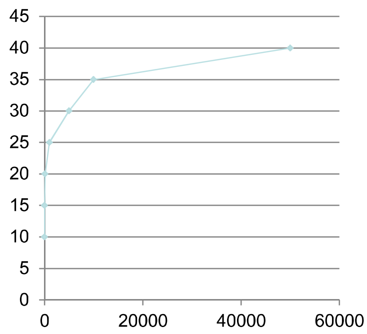
- yes/no data does not follow a normal distribution

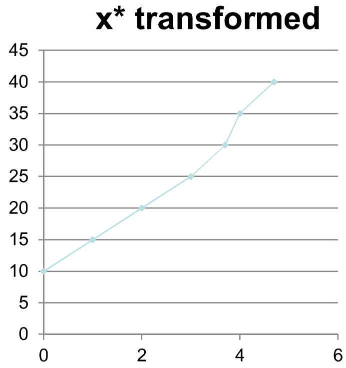
→ logistic regression

One way to model non-linear relations

- tranform ***x*** values to get a linear relation between ***x*** and ***y*** (e.g. log(***x***), $\boldsymbol{x^2}$, …)

X* transformed (log(*x*))	*x* value	*y* value
0	1	10
1	10	15
2	100	20
3	1000	25
3.7	5000	30
4	10000	35
4.7	50000	40





x* = log(x)

- for interpretation do not forget: $x = e^{x^*}$

Logistic regression (1)

Example:
Age and signs of coronary heart disease (CD) for 33 patients

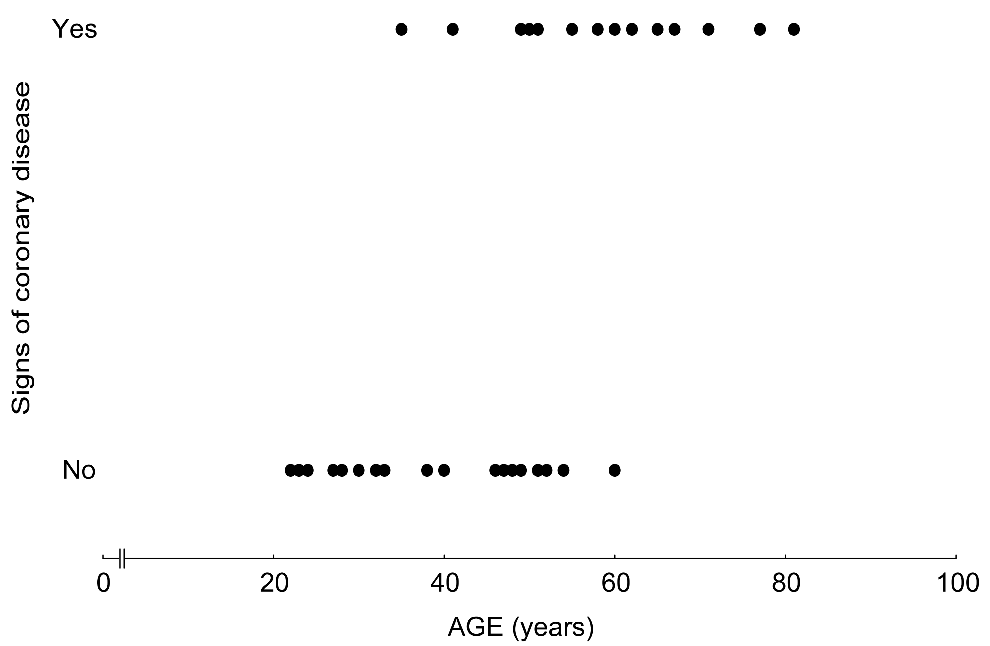
Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

What you see: age is continuous, signs of CD is binary (yes/no or 1/0)

How can we analyse these data?

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

Dot-plot of the data

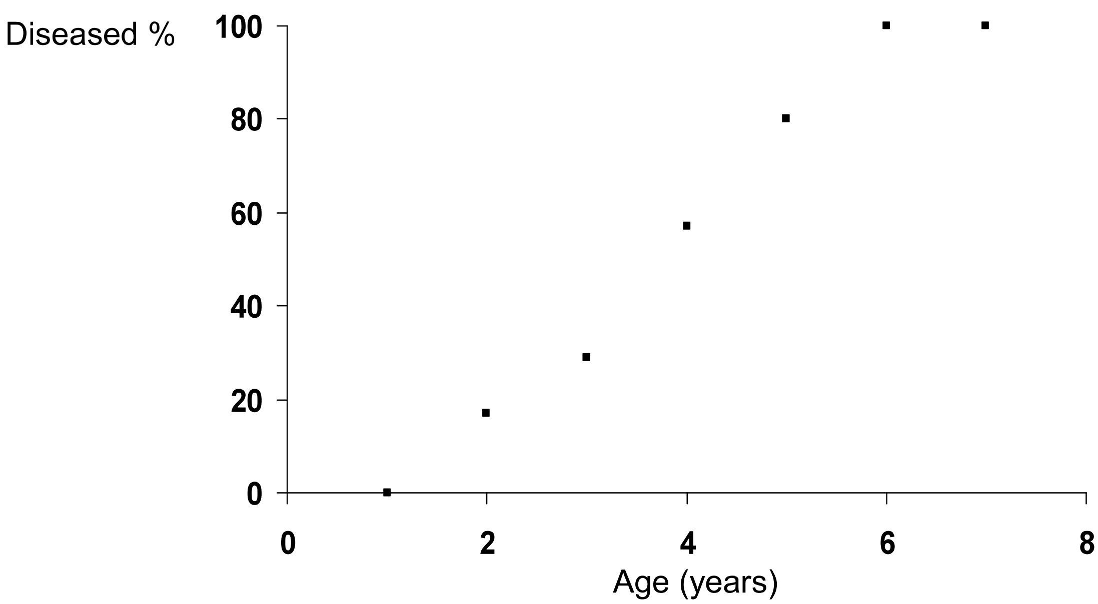


Logistic regression (2)

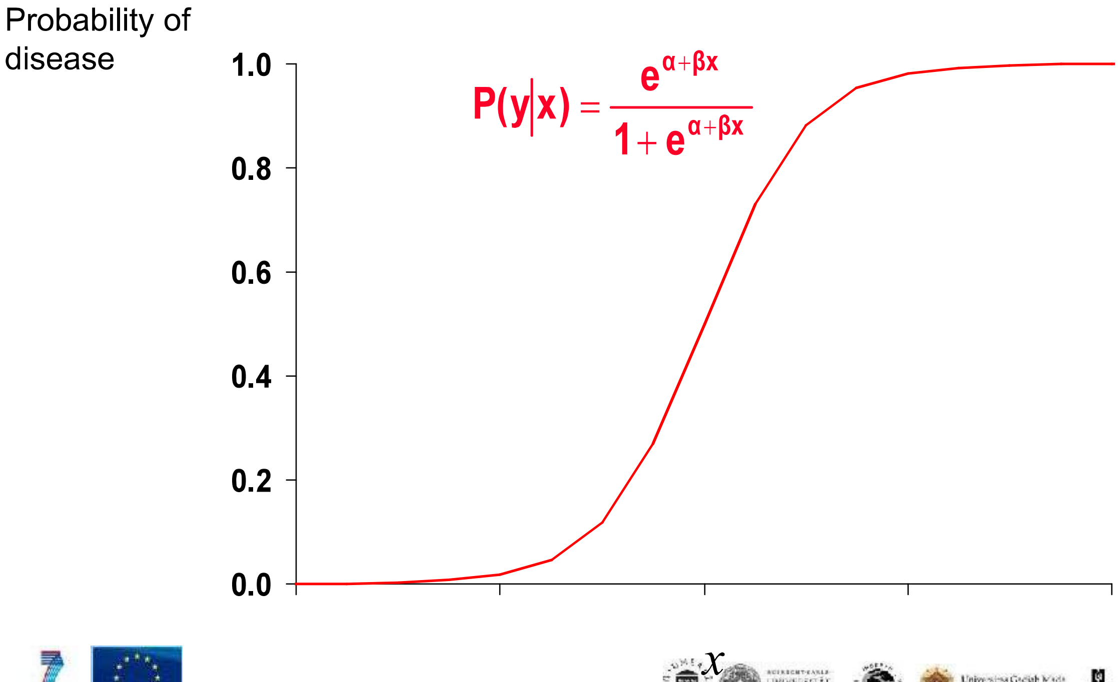
Accumulated data:
Prevalence (%) of signs of CD according to age group

Age group	# in group	Diseased #	Diseased %
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: accumulated data



The logistic function (1)



The logistic function (2)

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

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

The logistic function (3)

- Advantages of the logit
  - Simple transformation of P(y|x)
  - Linear relationship with x
  - Can be continuous (Logit between $-\infty$ to $+\infty$)
  - Known binomial distribution (P between 0 and 1)
  - Directly related to the notion of odds of disease

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

Coefficients

- In interpreting coefficients we're now thinking about a particular case's tendency toward some outcome
- The problem with probabilities is that they are non-linear
  - Going from .10 to .20 doubles the probability, but going from .80 to .90 only increases the probability somewhat

- With logistic regression we start to think about the odds
- Odds are just an alternative way of expressing the likelihood (probability) of an event.
  - Probability is the expected number of the event *divided by the total* number of possible outcomes
  - Odds are the expected number of the event *divided by the expected number of non-event occurrences*.
    - Expresses the likelihood of occurrence relative to likelihood of non-occurrence

Odds

- Let's begin with probability. Let's say that the probability of success is .8, thus
  - p = .8
- Then the probability of failure is
  - q = 1 - p = .2
- The odds of success are defined as
  - odds(success) = p/q = .8/.2 = 4,
  - that is, the odds of success are 4 to 1.
- We can also define the odds of failure
  - odds(failure) = q/p = .2/.8 = .25,
  - that is, the odds of failure are 1 to 4.

Odds Ratio

- Next, let's compute the odds ratio by
- OR = odds(success)/odds(failure) = 4/.25 = 16
- The interpretation of this odds ratio would be that the odds of success are 16 times greater than for failure.
- Now if we had formed the odds ratio the other way around with odds of failure in the numerator, we would have gotten
- OR = odds(failure)/odds(success) = .25/4 = .0625
- Here the interpretation is that the odds of failure are one-sixteenth the odds of success.

Logit

$$\text{logit} = \ln\left(\frac{P}{1-P}\right)$$

- Logit
  - Natural log (e) of an odds
  - Often called a *log odds*
    - *The logit scale is linear*

- Logits are continuous and are centered on zero (kind of like z-scores)
  - p = 0.50, odds = 1, then logit = 0
  - p = 0.70, odds = 2.33, then logit = 0.85
  - p = 0.30, odds = .43, then logit = -0.85

Logit

- So conceptually putting things in our standard regression form:
  - Log odds = $b_o + b_1X$

- Now a one unit change in X leads to a b_1 change in the log odds

- In terms of odds: $odds(Y=1) = e^{b_0 + b_1 X}$

- In terms of probability: $\Pr(Y=1) = \dfrac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$

- Thus the logit, odds and probability are different ways of expressing the same thing

Interpretation of β (1)

	Exposure (x)	
Disease (y)	Yes	No
Yes	$P(y \mid x=1)$	$P(y \mid x=0)$
No	$1-P(y \mid x=1)$	$1-P(y \mid x=0)$

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

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

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

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

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

Interpretation of β (2)

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

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

- Interval testing $\quad 95\%\ CI = e^{(\beta \pm 1.96 SE_{\beta})}$

Example

- Age (<55 and 55+ years) and risk of developing coronary heart disease (CD)

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

Odds of disease among exposed

Odds ratio =

Odds of disease among unexposed

- Results of fitting Logistic Regression Model

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

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

Log-odds = 2.094

OR = $e^{2.094}$ = 8.1

$$\text{Wald Test for effect of age} = 3.96^2 \text{ with 1 df, } p < 0.05$$

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

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

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

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 …

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

- 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

Multiple logistic regression

- Effect modification
  - Can be modelled by including interaction terms

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

Statistical testing

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

Likelihood ratio statistic

- Compares two nested models

  Log(odds) = $\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ (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

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

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

$$(\text{SE}\,0.2614) \quad (\text{SE}\,0.2664)$$

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

$$95\%\ \text{CI} = e^{(1.0047 \pm 1.96 \times 0.2614)} = 1.64 \text{ to } 4.56$$

Interaction

- Is there an interactive effect between smoking and exercise?

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

- Product term b_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

Model fit

- The Goodness-of-fit statistics helps you to determine whether the model adequately describes the data
- Calculating the deviance of a model

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 x 2 x 2 = 2^3 = 8

Continuous variable?

- Relationship between SBP>160 mmHg and body weight
- Introduce BW as a continuous variable?
  - Code weight as single variable, eg. 3 equal classes:
    40-60 kg = 0, 60-80 kg = 1, 80-100 kg = 2

BW	Cases	Controls	OR	
0	20	40	1.0	
1	22	30	1.5	$1.5^2 \approx 2.2$
2	12	11	2.2	

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 no 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 the same reference
- OR for each type of tobacco adjusted for the others in reference to non-smoking

Assumptions

Assumption	Issue	Recommendation
Sample Size	Sample should be large enough to populate categorical predictors. Limited cases in each category may result in failure to converge	Use crosstabs at variable selection stage to identify low populated cells, may result in recoding
Outliers	Cases that are strongly incorrectly predicted may have been poorly explained by the model and misclassified	Identify cases through classification table and residuals
Independence of Errors	Data observations should not be related i.e. one respondent per dataset, not repeated measures – overdispersion	Easy to avoid if the data collection has been conducted properly
Multicollinearity	Independent variables are highly inter-correlated (continuous) or strongly related to each other (categorical)	Use collinearity diagnostics in linear regression model and test high tolerance values using chi-square or correlation

Does not assume normal distribution of predictor variables – very useful!

Multicollinearity

- It occurs when one or more independent variables are highly correlated (i.e. not independent!)
- It tends to reduce or negate the influential effect of either predictor and can also have cumulating effects on the rest of the model
- It must be prevented at all costs and is more common than you might think: income, education, social class, age, house ownership, political party affiliation…

Reference

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