Logistic regression for inference

Dr Gianluca Campanella

31st May 2016

Contents

Regression models

Logistic regression

Regression models

Regression models

Regression models explore associations between:

- a **response** variable y
- **explanatory** variables (or **predictors**) $x_1, \ldots, x_p$

Regression models

Regression models explore associations between:

- a **response** variable y
- **explanatory** variables (or **predictors**) $x_1, \ldots, x_p$

Question

Do the $x_1, \ldots, x_p$ capture the **variability** of y?

Regression models

Regression models explore associations between:

- a **response** variable y
- **explanatory** variables (or **predictors**) $x_1, \dots, x_p$

Question

Do the $x_1, \dots, x_p$ capture the **variability** of y?

Aims

1. Predict the future (easy)
2. Understand the system being modelled (hard)

Regression modelling steps

- **Formulation**
  1. Error distribution for the response y
  2. Combination of predictors
  3. Link function
- **Estimation** of parameters
- **Diagnostics** (does the model fit the data well?)
- **Selection** (can we improve the fit?)

Components of regression models

(1) A model for the **variability** of the response **y**

- **y** is continuous → normal distribution
- **y** is categorical → **binomial distribution**

Components of regression models

(1) A model for the **variability** of the response **y**

- **y** is continuous $\rightarrow$ normal distribution
- **y** is categorical $\rightarrow$ **binomial distribution**

(2) A **combination of predictors** $x_1, \ldots, x_p$

- Often linear, e.g. $2x_1 + 3x_2$
- $\beta_1 = 2$ and $\beta_2 = 3$ are **regression coefficients**

Components of regression models

(1) A model for the **variability** of the response **y**

- **y** is continuous $\rightarrow$ normal distribution
- **y** is categorical $\rightarrow$ **binomial distribution**

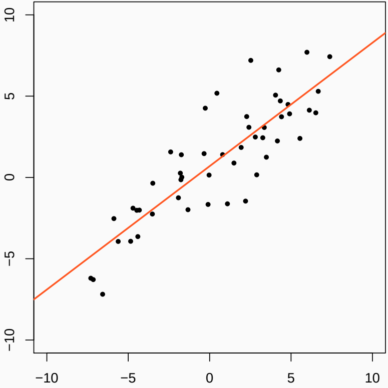
(2) A **combination of predictors** $x_1, \ldots, x_p$

- Often linear, e.g. $2x_1 + 3x_2$
- $\beta_1 = 2$ and $\beta_2 = 3$ are **regression coefficients**

(3) A **link** between the two

- Often depends on the model for the response
- Linear regression: $\mathbb{E}[y] = 2x_1 + 3x_2$

Linear regression



For the i^{th} observation:

$$\mathbb{E}[y_i] = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon$$

β_0	Intercept
β_j	Regression coefficients
ε	Error term

- Easy to estimate
- Easy to interpret

Classification problems

What happens if the outcome y is categorical?

Classification problems

What happens if the outcome $\mathbf{y}$ is **categorical**?

For **binary** outcomes, we can model the **probability**

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = p_i,$$

i.e. the probability of belonging to some category, as a function of the predictors $\mathbf{x_1}, \ldots, \mathbf{x_p}$

...but how?

Logistic regression

Logistic regression

Idea

Transform the linear predictor to lie on the unit interval

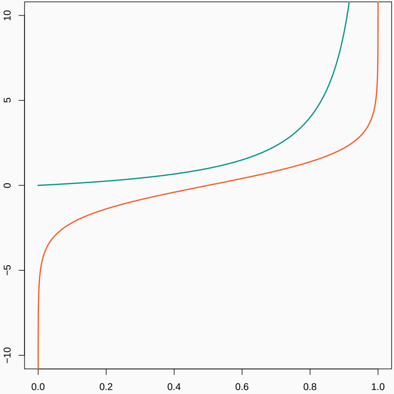
For the i^{th} observation:

$$\operatorname{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon$$

$\beta_0, \ldots, \beta_p$ represent the **log odds ratios** between classes

Probability and odds

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$



Throw a fair die. How often will you get a 1?

Probability

$$p = \frac{1}{6} \approx 16.67\% \text{ of the time}$$

Odds

$$\frac{p}{1-p} = \frac{1/6}{5/6} = \frac{1}{5} = 0.2$$

(once for every 5 times you don't)

Odds ratio

$$\mathrm{OR} = \frac{\text{odds in some group } (y = 1)}{\text{odds in a reference group } (y = 0)}$$

Odds ratio

Example

$$\text{OR} = \frac{\text{odds of smoking in lung cancer patients}}{\text{odds of smoking in cancer-free individuals}}$$

Interpretation

$$\text{OR} \begin{cases} < 1 & \text{smoking is } \textbf{less likely} \\ = 1 & \text{smoking is } \textbf{no more likely} \text{ in lung cancer patients} \\ > 1 & \text{smoking is } \textbf{more likely} \end{cases}$$

Example: Crime Survey for England and Wales

Outcome

'Did you experience any crime in the previous 12 months?'
(1,297 yes + 6,976 no = 8,273 respondents)

Predictors

- Sex
- Age
- 'How safe do you feel walking alone in after dark?' (variable `walkdark`, four categories)

Logistic regression coefficients

	β	$\exp(\beta)$
(Intercept)	−0.56	0.57
Sex		
Male	—	—
Female	−0.29	0.75
Age	−0.03	0.97
`walkdark`		
Very safe	—	—
Fairly safe	0.17	1.19
A bit unsafe	0.50	1.64
Very unsafe	0.81	2.24

Interpretation of logistic regression coefficients

Age

- $\exp(\beta) = 0.97 < 1$
- **Reduction** in risk of $1 - 0.97 = 3\%$ per year of age *(other things being equal)*

$\rightarrow$ 'Elderly less subject to crime'

Interpretation of logistic regression coefficients

Age

- $\exp(\beta) = 0.97 < 1$
- **Reduction** in risk of $1 - 0.97 = 3\%$ per year of age *(other things being equal)*

$\rightarrow$ 'Elderly less subject to crime'

`walkdark` = 'Very unsafe'

- $\exp(\beta) = 2.24 > 1$
- **Increase** in risk of $2.24 - 1 = 124\%$ if participant declares feeling 'very unsafe'

$\rightarrow$ 'Trust your gut feeling'

RECAP: logistic regression

Model

- Outcome is the **probability** of being in some class (dichotomised for prediction)
- Regression coefficients represent **log odds ratios**

Interpretation

- $\exp(\beta)$ is the **odds ratio** between $y = 0$ and $y = 1$
- OR $= 1$ is threshold ($=$ no effect)