Generalised linear models

Gianluca Campanella

Contents

Regression models

Regression models

Regression models explore associations between:

- A response variable $\vec{y}$
- Explanatory variables (or predictors) $\vec{x}_1, \ldots, \vec{x}_p$

Regression models

Regression models explore associations between:

- A response variable $\vec{y}$
- Explanatory variables (or predictors) $\vec{x}_1, \ldots, \vec{x}_p$

Question

Do the $\vec{x}_1, \ldots, \vec{x}_p$ capture the variability of $\vec{y}$?

Regression modelling steps

- Formulation
  1. Error distribution for the response $\vec{y}$
  2. Combination of predictors
  3. Link function
- Estimation of regression coefficients
- 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 $\vec{y}$

- $\vec{y}$ is continuous $\rightarrow$ normal distribution
- $\vec{y}$ is dichotomous $\rightarrow$ binomial distribution

Components of regression models

(1) A model for the variability of the response $\vec{y}$

- $\vec{y}$ is continuous $\rightarrow$ normal distribution
- $\vec{y}$ is dichotomous $\rightarrow$ binomial distribution

(2) A combination of predictors $\vec{x}_1, \ldots, \vec{x}_p$

- Often linear, e.g. $2\vec{x}_1 + 3\vec{x}_2$
- $\beta_1 = 2$ and $\beta_2 = 3$ are regression coefficients

Components of regression models

(1) A model for the variability of the response $\vec{y}$

- $\vec{y}$ is continuous $\to$ normal distribution
- $\vec{y}$ is dichotomous $\to$ binomial distribution

(2) A combination of predictors $\vec{x}_1, \ldots, \vec{x}_p$

- Often linear, e.g. $2\vec{x}_1 + 3\vec{x}_2$
- $\beta_1 = 2$ and $\beta_2 = 3$ are regression coefficients

(3) A link between the two

- Often depends on the model for the response
- Linear regression: $\mathbb{E}[\vec{y}] = 2\vec{x}_1 + 3\vec{x}_2$

Predictors and response

Predictors

- Viewed as fixed variables
- Assumed not to be affected by measurement error

→ 'Independent' or 'exogenous'

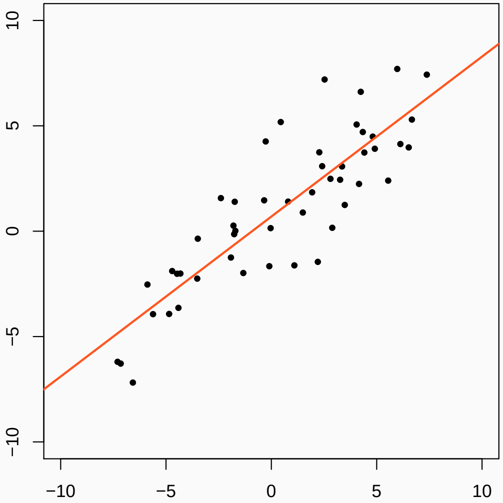
Response

- Variability is modelled
  (but could also be attributed to other factors)

→ 'Dependent' or 'endogenous'

Linear regression

Simple linear regression



For the i^{th} observation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

β_0	Intercept
β_1	Slope
ε_i	Individual error term

Regression coefficients

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Intercept Average y when $x = 0$

Slope Increase in y for a one-unit increase in x

The regression line passes through:

- The point $(0, \beta_0)$
- The 'centre' of the data $(\bar{\vec{x}}, \bar{\vec{y}})$

Error term

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- 'Sucks up' unaccounted variation in $\vec{y}$
- Model assumptions are mostly on ε

Multiple linear regression

For the i^{th} observation:

$$y_i = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon_i$$

β_0	Intercept
β_j	Slopes
ε_i	Individual error term

Intercept Average y when all $x_{\cdot j} = 0$

Slopes Increase in y for a one-unit increase in $x_{\cdot j}$
all else being equal

Multiple linear regression

In matrix form:

$$\vec{y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}$$

$\mathbf{X}$	Design matrix
$\vec{\beta}$	Regression coefficients
$\vec{\varepsilon}$	Error term

Gauss–Markov assumptions (plus one)

- The relationship between $\vec{y}$ and $\mathbf{X}$ is linear
- The $\vec{x}_1, \ldots, \vec{x}_p$ are not collinear
- Exogeneity
    - Given $\mathbf{X}$, errors have mean 0
    - Since $\mathbf{X}_i$ is deterministic, it is uncorrelated with ε_i
- Spherical errors
    - Errors have a fixed variance (homoscedasticity)
    - Errors are uncorrelated between observations (no autocorrelation)
- (Given $\mathbf{X}$, errors are normally distributed)

Model fitting by maximum likelihood

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \quad \text{where} \quad \mu_i = \beta_0 + \sum_j \beta_j x_{ij}$$

β_j ‘True’ values (fixed but unknown)

$\hat{\beta}_j$ Our estimates for the β_j (computed from the data)

Given some values for the $\hat{\beta}_j$...

- We can write down the probability of observing each Y_i alone
- Since the Y_i are independent by assumption, we can write down the joint probability of observing the Y_i together

$\rightarrow$ $f(\vec{y} \mid \hat{\beta}_j)$ is the probability of the data given the parameters

Model fitting by maximum likelihood

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \quad \text{where} \quad \mu_i = \beta_0 + \sum_j \beta_j x_{ij}$$

Maximum likelihood principle

- Consider instead the likelihood function $f(\hat{\beta}_j \mid \vec{y})$
- Same as $f(\vec{y} \mid \hat{\beta}_j)$, but interpreted as the probability of certain parameter values given the data

$\rightarrow$ Can optimise to estimate the $\hat{\beta}_j$

Hypothesis testing for parameters

How do we know the estimates $\hat{\beta}_j$ are not just random fluctuations?

Hypothesis testing for parameters

How do we know the estimates $\hat{\beta}_j$ are not just random fluctuations?

$$\text{Additional assumption: } \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

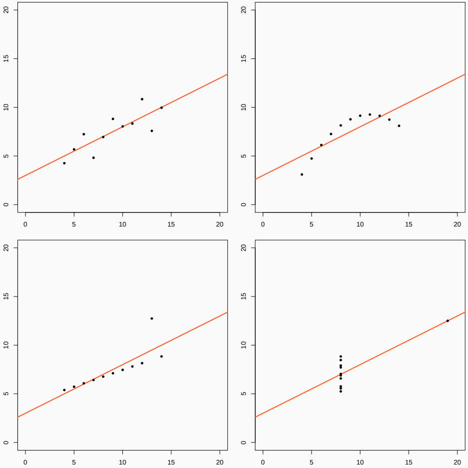
$$\downarrow$$

- Define confidence intervals for $\hat{\beta}_j$
- Test H_0 that $\hat{\beta}_j = 0$ (no effect)

Diagnostics for linear regression

Assumption violated	Severity	Causes
Linearity or additivity	++++	Model misspecification
Independence	+++	Autocorrelation (typical of time series)
Homoscedasticity	++	σ^2 changes over the range of $\vec{y}$
Normality	+	Outliers

Many datasets, one regression line



Logistic regression

Classification problems

What happens if the outcome $\vec{y}$ is dichotomous?

Classification problems

What happens if the outcome $\vec{y}$ is dichotomous?

We can model the probability

$$\Pr\left(y_i = 1 \;\middle|\; \vec{x}_i\right) = p_i,$$

i.e. the probability of belonging to some non-reference category, as a function of the predictors $\vec{x}_1, \ldots, \vec{x}_p$

…but how?

Logistic regression

Idea

Transform the linear predictor to lie on the unit interval

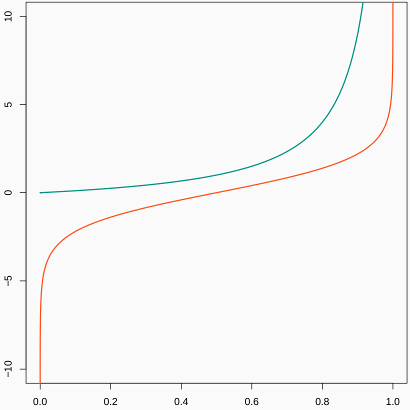
For the i^{th} observation:

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

$\beta_0, \ldots, \beta_p$ 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

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

Odds ratio

Example

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

Interpretation

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

Logistic regression recap

Model

- Outcome is the probability of being in some non-reference class
- Regression coefficients represent log odds ratios

Interpretation of coefficients

- $\exp(\beta)$ is the odds ratio between $y = 0$ and $y = 1$
- $\text{OR} = 1$ is the threshold corresponding to no effect