εstimand GA

Linear regression for inference

Dr Gianluca Campanella

17th May 2016

Contents

Where we are

1. Define the **research question**
2. **Get** the data
3. **Explore** the data
    - (Re)format, clean, merge, stratify...
    - Identify trends and outliers
4. **Model** the data
    - Select and build model(s)
    - Evaluate and refine model(s)
5. **Summarise** the results
    - Summarise findings
    - Describe assumptions and limitations
    - Identify follow-up research questions

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 $\rightarrow$ normal distribution
- **y** is categorical $\rightarrow$ (more later in the course...)

Components of regression models

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

- y is continuous $\rightarrow$ normal distribution
- y is categorical $\rightarrow$ (more later in the course...)

(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 $\rightarrow$ (more later in the course...)

(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: $\mathbb{E}[y] = 2x_1 + 3x_2$

Predictors and response

Predictors

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

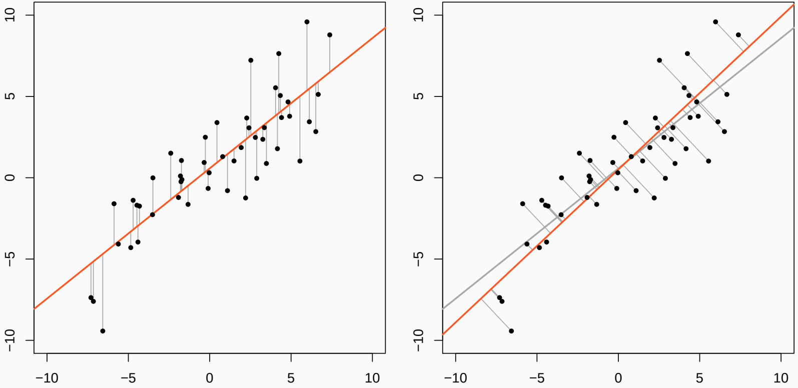
$\rightarrow$ 'Independent' or 'exogenous'

Response

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

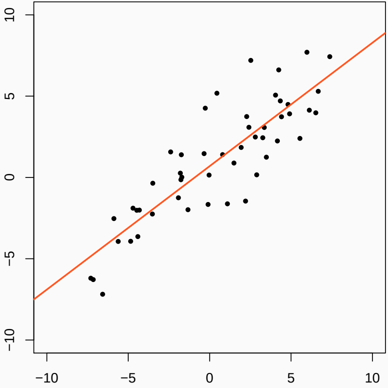
$\rightarrow$ 'Dependent' or 'endogenous'

Example: estimation of parameters



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 $(\bar{x}, \bar{y})$

Error term

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

- 'Sucks up' unaccounted variation in **y**
- Model assumptions are mostly on ε (more later...)

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 $x_{.j} = 0$

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

Multiple linear regression

In matrix form:

$$y = X\beta + \varepsilon$$

X	Design matrix
β	Regression coefficients
ε	Error term

Gauss—Markov assumptions

If the following holds...

- The relationship between y and X is linear
- X has full rank (no multicollinearity)
- Exogeneity: $\mathbb{E}[\varepsilon_i \mid \mathsf{X}_i] = 0$
- Homoscedasticity: $\mathbb{V}[\varepsilon_i \mid \mathsf{X}_i] = \sigma^2 < \infty$
- Uncorrelated error terms: $\mathbb{V}\left[\varepsilon_i, \varepsilon_j\right] = 0$, $i \neq j$
- X_i is deterministic and thus uncorrelated with ε_i: $\mathbb{V}[\mathsf{X}_i, \varepsilon_i] = \mathbb{E}[\mathsf{X}_i \varepsilon_i] - \mathbb{E}[\mathsf{X}_i]\,\mathbb{E}[\varepsilon_i] = \mathsf{X}_i \mathbb{E}[\varepsilon_i] - \mathsf{X}_i \mathbb{E}[\varepsilon_i] = 0$

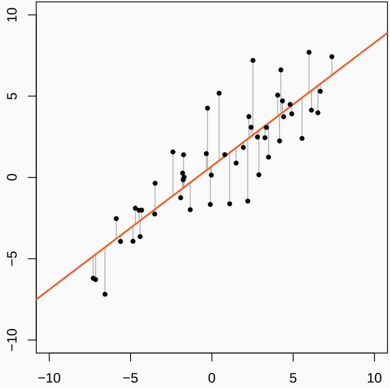
Gauss—Markov assumptions

…then the OLS estimator of β is **BLUE**:

Best	Minimal variance
Linear	Like the relationship between $\mathbf{y}$ and $\mathbf{X}$
Unbiased	$\mathbb{E}[\beta]$ equals the 'true' values for β
Estimator	

Outside of physics, these assumptions are often violated

Model fitting: ordinary least squares



For the i^{th} observation:

$$\hat{\varepsilon}_i = y_i - \left(\hat{\beta}_0 + \sum_j \hat{\beta}_j x_{ij}\right)$$

Ordinary least squares

Find $\hat{\beta}_j$ that minimise the **residual sum of squares**

$$\text{RSS} = \sum_i \hat{\varepsilon}_i^2$$

Model fitting: maximum likelihood

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

- Assume that there are **fixed**, 'true' values for the $\hat{\beta}_j$
- We can write down the **densities** of the Y_i
- Assuming **independence** of the Y_i, we can write down the **joint** density of the Y_i

$\rightarrow$ $f(y \mid \hat{\beta}_j)$ represents the probability of observing the data **given the parameters**

Model fitting: 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** $f(\hat{\beta}_j \mid y)$
- Same as before, but interpreted as the probability of certain parameter values **given the data**

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

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

Hypothesis testing for parameters

$\hat{\beta}_0, \hat{\beta}_1, \ldots$ are **estimated** from the data (they have a hat).
How do we know they are not just random fluctuations?

- Define confidence intervals for $\hat{\beta}_j$
- Test H_0 that $\hat{\beta}_j = 0$

$$\downarrow$$

Need sampling distribution of β_j

Hypothesis testing for parameters

- Standard deviation of the sampling distribution $\sqrt{\mathbb{V}[\beta_j]}$ (**standard error**) is known
- Test statistic is simply $T_j = {}^{\beta_j}/_{\sqrt{\mathbb{V}[\beta_j]}}$
- It can be shown that T_j follows a t-distribution

$\rightarrow$ Can compute confidence intervals and p-values

Diagnostics for linear regression

Violations of linearity or additivity

Extremely serious

- Model is misspecified
- Inference outside of observed range is misleading

Diagnostics

- Predicted $\hat{y}$ versus observed y values
- Residuals $\hat{\varepsilon}$ versus predicted $\hat{y}$ values
- Residuals $\hat{\varepsilon}$ versus each independent variable x_j

Violations of independence

Potentially very serious

- Especially if dealing with time series (autocorrelation)
- Can also result from model misspecification

Diagnostics

- Residuals $\hat{\varepsilon}$ versus time, row number...
- Residual autocorrelation
- Durbin—Watson test for autocorrelation at lag 1

Violations of homoscedasticity

Serious

- Confidence intervals are too wide or too narrow
- Data are weighted unequally

Diagnostics

- Residuals $\hat{\varepsilon}$ versus predicted $\hat{y}$ values
- Residuals $\hat{\varepsilon}$ versus time, row number...
- Residuals $\hat{\varepsilon}$ versus each independent variable x_j

Violations of normality

Somewhat serious

- Often due to a few outliers
- Confidence intervals and p-values unreliable

Diagnostics

- Normal quantile plot of the residuals $\hat{\boldsymbol{\varepsilon}}$
- Statistical tests for normality (e.g. Anderson—Darling)
- Studentised residuals:

$$\hat{r}_i = \frac{y_i - \hat{y}_i}{\sqrt{\mathbb{V}[\hat{y}_i]}}$$

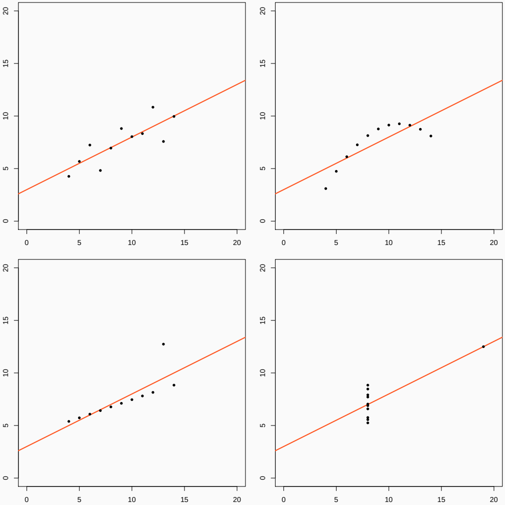
Transformations

- Can be applied to predictors and/or response
- Can improve model fit (e.g. when residuals are not normally distributed or homoscedastic)

Commonly used transformations

- $\log y$ and $\exp y$
- $\sqrt{y}$ and y^2
- $1/y$

Many datasets, one regression line



Model selection

Coefficient of determination

Total sum of squares

$$\mathrm{TSS} = \sum_i (y_i - \bar{y}_i)^2$$

Residual sum of squares

$$\mathrm{RSS} = \sum_i \hat{\varepsilon}_i^2 = \sum_i (y_i - \hat{y}_i)^2$$

Coefficient of determination

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

Coefficient of determination

Problem

R^2 increases with the number of predictors

Idea

Penalise larger models in the goodness-of-fit metric

- Adjusted R^2
- Mallows's C_p
- AIC and BIC