Imperial College London

Lecture 11: Introduction to Linear Models

Statistical Modelling I

Dr. Riccardo Passeggeri

Last time

Lectures 1-10: focus on methods for inference in samples that are iid

Outline

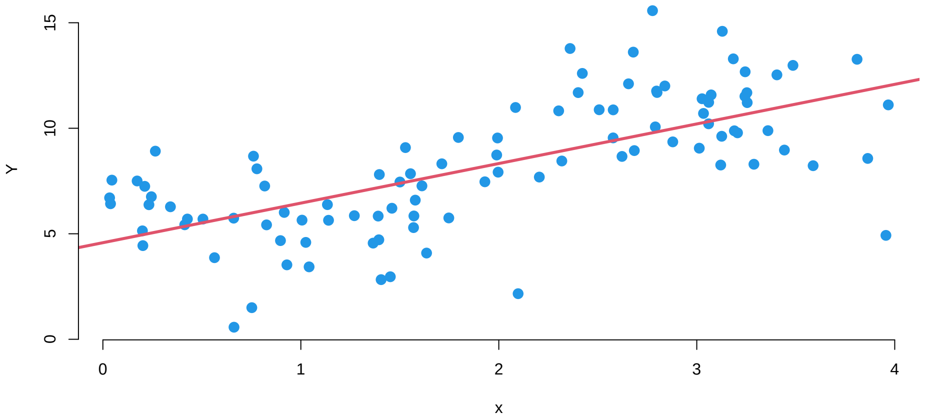
1. Introduction

2. Matrix Algebra

3. Expectations of Random Vectors

Introduction

Why linear models?



Definition: Simple Linear Model

$$Y_i = \beta_1 + x_i\beta_2 + \epsilon_i, \quad i = 1, \ldots, n$$

- Y_i "outcome", "response"; observable random variable.
- x_i "covariate"; observable constant.
- β_1, β_2 unknown parameters.
- Error $\epsilon_1, \ldots, \epsilon_n$ iid, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ for $i = 1, \ldots, n$.
- $\sigma^2 > 0$ is another unknown parameter.
- The errors $\epsilon_1, \ldots, \epsilon_n$ are not observable.

Least squares estimators

The *least squares estimators* $\hat{\beta}_1$, $\hat{\beta}_2$ of β_1 and β_2 are defined as the minimisers of

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - x_i \beta_2)^2.$$

Note that:

- $e_i = y_i - \hat{\beta}_1 - x_i \hat{\beta}_2$, the so-called residuals, are observable. They are not iid, as dependence is introduced via $\hat{\beta}_1$, $\hat{\beta}_2$.
- The unknown parameters are β_1, β_2 and σ^2.
- In linear regression models $Y_1, \ldots, Y_n$ are generally not iid observations. Independence will still hold if the errors $\epsilon_1, \ldots, \epsilon_n$ are independent. However, the Y_i do not have the same distribution; the distribution of Y_i depends on the covariate x_i.

Matrix Algebra

A toolkit for linear algebra

Linear regression naturally leads to a connection between statistics and linear algebra

This lecture, we highlight some useful results about matrices.

A^T denotes the transpose of a matrix. I will use the terms "invertible" and "non-singular" synonymously.

Matrix transposition, multiplication and inversion:

- Let $A \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{m \times n}$. Then $(AB)^T = B^T A^T$
- Let $A \in \mathbb{R}^{n \times n}$ be non-singular. Then $(A^{-1})^T = (A^T)^{-1}$.

Transpose and trace

(Trace) Let $A = (A_{ij}) \in \mathbb{R}^{n \times n}$. Then

$$\operatorname{trace}(A) = \sum_{i=1}^{n} A_{ii}$$

Lemma. $\operatorname{trace}(AB) = \operatorname{trace}(BA)$.

Proof. Recall that $AB = (\sum_j A_{ij} B_{jk})_{i,k}$. Thus, we have that

$$\operatorname{trace}(AB) = \sum_i \sum_j A_{ij} B_{ji} = \sum_j \sum_i B_{ji} A_{ij} = \operatorname{trace}(BA).$$

Example. Let $A = (1, 1)$, $B = (1, 1)^T$. Then $AB = 2 \neq \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = BA$, but $\operatorname{trace}(AB) = 2 = \operatorname{trace}(BA)$.

Rank of $X^T X$

Let X be an $n \times p$ matrix. Then $\text{rank}(X^T X) = \text{rank}(X)$.

Proof. Let $\text{kern}(X) = \{\boldsymbol{x} \in \mathbb{R}^p : X\boldsymbol{x} = \boldsymbol{0}\}$. Then $p = \text{rank}\, X + \dim \text{kern}(X)$. Similarly, $p = \text{rank}\, X^T X + \dim \text{kern}(X^T X)$

It suffices to show: $\text{kern}(X) = \text{kern}(X^T X)$.

If $\boldsymbol{x} \in \text{kern}(X)$ then $\boldsymbol{0} = X\boldsymbol{x}$ and hence $\boldsymbol{0} = X^T X\boldsymbol{x}$ which shows $\boldsymbol{x} \in \text{kern}(X^T X) = \{\boldsymbol{y} : X^T X\boldsymbol{y} = \boldsymbol{0}\}$.

If $\boldsymbol{x} \in \text{kern}(X^T X)$ then $\boldsymbol{0} = X^T X\boldsymbol{x}$ and thus

$$\boldsymbol{0} = \boldsymbol{x}^T X^T X\boldsymbol{x} = (X\boldsymbol{x})^T X\boldsymbol{x} = \|X\boldsymbol{x}\|^2$$

which shows $X\boldsymbol{x} = \boldsymbol{0}$, i.e. $\boldsymbol{x} \in \text{kern}(X)$.

A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is **positive definite** if

$$\forall \boldsymbol{x} \in \mathbb{R}^n \setminus \{0\} : \boldsymbol{x}^T A \boldsymbol{x} > 0.$$

Lemma. $A \in \mathbb{R}^{n \times n}$ is symmetric $\implies \exists$ orthogonal matrix P (i.e. $P^T P = I$) s.t. $P^T A P$ is diagonal (with diagonal entries equal to the eigenvalues of A).

A an $n \times n$ positive definite symmetric matrix $\implies \exists$ non-singular matrix Q s.t. $Q^T A Q = I_n$.

Proof.

First part is a standard linear algebra result.
The second result can be derived from it: A p.d. $\implies$ its eigenvalues are > 0.
Hence, $P^T A P = D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ where $\lambda_i > 0 \, \forall i$.
Let $E = D^{\frac{1}{2}} = \operatorname{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$ and define $Q = PE^{-1}$. Then

$$Q^T A Q = (PE^{-1})^T APE^{-1} = (E^{-1})^T P^T APE^{-1} = (E^{-1})^T EEE^{-1} = I.$$

$\square$

Expectations of Random Vectors

Why do we need expectations of random vectors?

Linear regression models describe the relationship between Y and x based on $E(Y \mid x)$.

The parameter *vector* (β_0, β_1) suggests there may be correlation between least squares estimators.

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be a random vector.

Then

$$\mathrm{E}(\boldsymbol{X}) = (\mathrm{E}\, X_1, \ldots, \mathrm{E}\, X_n)^T,$$

i.e. the expectation is defined componentwise. For random matrices the expectation is also defined componentwise.

Lemma: Linearity of expectations

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be n-variate random vectors. Then the following hold:

- $\mathrm{E}(\boldsymbol{X} + \boldsymbol{Y}) = \mathrm{E}\,\boldsymbol{X} + \mathrm{E}\,\boldsymbol{Y}$.
- Let $a \in \mathbb{R}$ then $\mathrm{E}(a\boldsymbol{X}) = a\,\mathrm{E}(\boldsymbol{X})$
- Let A, B be deterministic matrices of "suitable dimensions" (deterministic means that they are not random). Then $\mathrm{E}(A\boldsymbol{X}) = A\,\mathrm{E}(\boldsymbol{X})$ and $\mathrm{E}(\boldsymbol{X}^T B) = \mathrm{E}(\boldsymbol{X})^T B$.

Proof. Use properties of one-dimensional random variables, for example

$$\mathrm{E}(A\boldsymbol{X}) = (\mathrm{E}(\sum_j A_{ij} X_j))_i = (\sum_j A_{ij}\,\mathrm{E}(X_j))_i = A\,\mathrm{E}(\boldsymbol{X}).$$

Covariance of random vectors

If $\boldsymbol{X}$, $\boldsymbol{Y}$ are random vectors then

$$\begin{aligned}\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y}) &:= (\operatorname{cov}(X_i, Y_j))_{i,j} \\ &= \mathrm{E}[(\boldsymbol{X}-\mathrm{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathrm{E}(\boldsymbol{Y}))^T] = \mathrm{E}[\boldsymbol{X}\boldsymbol{Y}^T] - \mathrm{E}(\boldsymbol{X})\,\mathrm{E}(\boldsymbol{Y})^T.\end{aligned}$$

Furthermore $\operatorname{cov}(\boldsymbol{X}) := \operatorname{cov}(\boldsymbol{X},\boldsymbol{X})$.

Lemma: Covariance properties

If $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are random vectors, A, B are deterministic matrices and $a, b \in \mathbb{R}$ are constants then (assuming appropriate dimensions)

- $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})^T$
- $\operatorname{cov}(a\boldsymbol{X} + b\boldsymbol{Y}, \boldsymbol{Z}) = a \operatorname{cov}(\boldsymbol{X}, \boldsymbol{Z}) + b \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{Z})$
- $\operatorname{cov}(A\boldsymbol{X}, B\boldsymbol{Y}) = A \operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) B^T$
- $\operatorname{cov}(A\boldsymbol{X}) = A \operatorname{cov}(\boldsymbol{X}) A^T$
- $\operatorname{cov}(\boldsymbol{X})$ is positive semidefinite and symmetric, i.e. $\boldsymbol{c}^T \operatorname{cov}(\boldsymbol{X}) \boldsymbol{c} \geq 0$ for all vectors $\boldsymbol{c}$, or, equivalently, all eigenvalues of $\operatorname{cov}(\boldsymbol{X})$ are nonnegative.
- If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent then $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) = 0$.

Proof. Work from properties of one-dimensional covariance or work with one of the vector definitions of the covariance.

Examples 1 and 2

Let $X \sim \mathit{Binomial}(17, 0.4)$. Then

$\text{cov}(X) =$

If $Y_1, \ldots, Y_n$ are independent then

$$\text{cov}\left(\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}\right) =$$

Example 3

Let X, Y be independent r.v. with $X \sim N(5, 2)$ and $Y \sim \text{Binomial}(10, 0.5)$. Then

$$\text{cov}\left(\begin{pmatrix} X \\ -X \end{pmatrix}\right) =$$

$$\text{cov}\left(\begin{pmatrix} X \\ X + Y \end{pmatrix}\right) =$$

$$\text{cov}\left(X, \begin{pmatrix} 2X \\ X - Y \end{pmatrix}\right) =$$

Looking ahead

In the next lecture we discuss how to use these concepts to specify and work with general linear models (with multiple predictors)