Linear regression

Advanced Statistics and Data Analysis

Davide Risso

Table of contents

Linear Algebra

Linear Algebra

Linear algebra, also called matrix algebra, and its mathematical notation greatly facilitates the understanding of the concepts behind the linear models.

Here, we spend some time to introduce / review key concepts of linear algebra, with particular focus on matrix operations and matrix notation.

We will also see how to perform matrix operations in R.

Matrix Notation

Linear algebra was created to solve systems of linear equations, e.g.

$$\begin{aligned} a + b + c &= 6 \\ 3a - 2b + x &= 2 \\ 2a + b - c &= 1 \end{aligned}$$

which becomes, in matrix notation,

$$\begin{bmatrix} 1 & 1 & 1 \\ 3 & -2 & 1 \\ 2 & 1 & -1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \\ 1 \end{bmatrix}$$

The system is solved by

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 3 & -2 & 1 \\ 2 & 1 & -1 \end{bmatrix}^{-1} \begin{bmatrix} 6 \\ 2 \\ 1 \end{bmatrix}$$

where the -1 denotes the inverse of the matrix.

We can borrow this notation to solve linear models in statistics.

Vectors, matrices, and scalars

In most of the previous lectures we focused on *scalar quantities* (numbers).

We did use some vectors, e.g., when dealing with the n observations in a sample or with the parameter $\theta = (\mu, \sigma^2)$ of the normal model.

Here, we will use vectors as special cases of matrices, with one column, e.g.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

This means that all the operations defined for matrices will be applicable to vectors as well.

We can write matrices by either concatenating vectors or by explicitly writing their elements.

$$X = [X1\, X_2] = \begin{bmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ \vdots & \vdots \\ X_{n,1} & X_{n,2} \end{bmatrix}$$

Vectors, matrices, and scalars

We already know how to create these objects in R

```
x1 <- c(1, 4, 6)
x2 <- c(7, 7, 2)
cbind(x1, x2)
```

```
     x1 x2
[1,]  1  7
[2,]  4  7
[3,]  6  2
```

```
matrix(data = c(1, 4, 6, 7, 7, 2), nrow = 3, ncol = 2)
```

```
     [,1] [,2]
[1,]    1    7
[2,]    4    7
[3,]    6    2
```

Matrix operations

Here, we review some of the operations that can be performed on matrices, both in mathematical terms and in R.

- Multiplication by a scalar
- Transposition
- Matrix multiplication
- Inverse

Multiplication by a scalar

This is the simplest operation: each element of the matrix is multiplied by the scalar.

If

$$X = \begin{bmatrix} X_{1,1} & \dots & X_{1,p} \\ X_{2,1} & \dots & X_{2,p} \\ \dots & \dots & \dots \\ X_{n,1} & \dots & X_{n,p} \end{bmatrix}$$

then

$$a\,X = \begin{bmatrix} aX_{1,1} & \dots & aX_{1,p} \\ aX_{2,1} & \dots & aX_{2,p} \\ \vdots & \ddots & \vdots \\ aX_{n,1} & \dots & aX_{n,p} \end{bmatrix}$$

Multiplication by a scalar

R automatically recognize a scalar and a matrix and "does the right thing."

```
a <- 2
x <- matrix(data = c(1, 4, 6, 7, 7, 2), nrow = 3, ncol = 2)
x
```

```
     [,1] [,2]
[1,]    1    7
[2,]    4    7
[3,]    6    2
```

```
a * x
```

```
     [,1] [,2]
[1,]    2   14
[2,]    8   14
[3,]   12    4
```

The transpose

Transposition is a simple operation that simply changes columns to rows. We use $\top$ sign to denote the transposed of a matrix.

$$X^\top = \begin{bmatrix} X_{1,1} & \dots & X_{n,1} \\ X_{1,2} & \dots & X_{n,2} \\ \vdots & \ddots & \vdots \\ X_{1,p} & \dots & X_{n,p} \end{bmatrix}$$

The transpose

In R, we can use the `t()` operator to transpose a matrix.

```
t(x)
```

```
     [,1] [,2] [,3]
[1,]    1    4    6
[2,]    7    7    2
```

Matrix multiplication

If A is an $n \times m$ matrix and B is an $m \times p$ matrix, their matrix product $X = AB$ is an $n \times p$ matrix, in which each row of A is multiplied by each column of B and summed.

In more details, the element $x_{ij} = \sum_{k=1}^{m} a_{i,k} b_{k,j}$.

In order to multiply two matrices, the number of columns of the first matrix needs to be the same as the number of rows of the second matrix.

In R, we can use the %*% operator.

Matrix multiplication

```
y <- matrix(c(5, 6, 7, 2), ncol=2, nrow=2)
x %*% y
```

```
     [,1] [,2]
[1,]   47   21
[2,]   62   42
[3,]   42   46
```

Properties of matrix multiplication

In general, the matrix multiplication operator is

1. Not commutative: $A\,B \neq B\,A$.
2. Distributive over matrix addition: $A(B+C) = AB + AC$.
3. Compatible with scalar multiplication: $a(AB) = (aA)B$
4. Transposition: $(AB)^\top = B^\top A^\top$.

The identity matrix

For scalars we have the number 1, a number such that $1x = x$ for any x.

The analogous for matrices, is the *identity matrix*

$$I = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

By definition, for any matrix X, $X\,I = X$.

In R we can use the function `diag()`.

The identity matrix

```
x %*% diag(2)
```

```
     [,1] [,2]
[1,]    1    7
[2,]    4    7
[3,]    6    2
```

```
diag(3) %*% x
```

```
     [,1] [,2]
[1,]    1    7
[2,]    4    7
[3,]    6    2
```

The inverse

A matrix is called *square* if it has the same number of rows and columns.

The inverse of a square matrix X, denoted by X^{-1} is the matrix that when multiplied by X returns the identity matrix.

$$X\,X^{-1} = I.$$

Note that not all matrices have a defined inverse.

We will see later how this plays a role for linear models.

If the inverse exists, it can be computed using the `solve()` function in R. Note that the `solve()` function is numerically unstable and should be used with caution.

```
y
```

```
     [,1] [,2]
[1,]    5    7
[2,]    6    2
```

```
solve(y)
```

```
        [,1]     [,2]
[1,] -0.0625  0.21875
[2,]  0.1875 -0.15625
```

```
y %*% solve(y)
```

```
     [,1]          [,2]
[1,]    1 2.220446e-16
[2,]    0 1.000000e+00
```

Using matrix algebra in statistics

Why do we need linear algebra and matrix notation in statistics?

Linear algebra is a very convenient and compact mathematical way of formalizing many of the concepts of statistics, especially with regards to linear models.

Let's focus on a couple of simple examples.

Example: The mean

Assume that we have a vector X that contains the data for a sample of n observations. We can use matrix notation to denote the sample mean of that vector.

Define a $n \times 1$ matrix $\mathbf{1} = [1 \ ... \ 1]^\top$.

We can compute the mean simply as $\frac{1}{n}\mathbf{1}^\top X$.

$$\frac{1}{n}\mathbf{1}^\top X = \frac{1}{n}[1 \ ... \ 1] \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$

Example: The mean

```
n <- 10
x <- rnorm(n)
mean(x)
```

```
[1] 0.2947863
```

```
A <- rep(1, n)
1/n * t(A) %*% x
```

```
          [,1]
[1,] 0.2947863
```

Example: The variance

The same is true for the variance: We can simply multiply the centered matrix by its transpose to compute it.

$$R = X - \bar{X} = \begin{bmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{bmatrix}$$

Then,

$$\frac{1}{n-1} R^\top R = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Example: The variance

```
var(x)
```

```
[1] 1.27616
```

```
r <- x - mean(x)
1/(n-1) * t(r) %*% r
```

```
         [,1]
[1,] 1.27616
```

The crossprod function

The operation $X^\top Y$ is so important in statistics that R has a shortcut function for it, the `crossprod()` function.

```
A <- matrix(rnorm(4), nrow=2, ncol=2)
B <- matrix(rnorm(4), nrow=2, ncol=2)
t(A) %*% B
```

```
           [,1]        [,2]
[1,] -1.883999  0.91801387
[2,] -1.742544 -0.07409988
```

```
crossprod(A, B)
```

```
           [,1]        [,2]
[1,] -1.883999  0.91801387
[2,] -1.742544 -0.07409988
```

The crossprod function

The `crossprod` function can be used to compute the variance.

```
var(x)
```

```
[1] 1.27616
```

```
crossprod(r)/(n-1)
```

```
         [,1]
[1,] 1.27616
```

Note that `crossprod(r)` is a further shortcut for `crossprod(r, r)`.

Least-squares regression

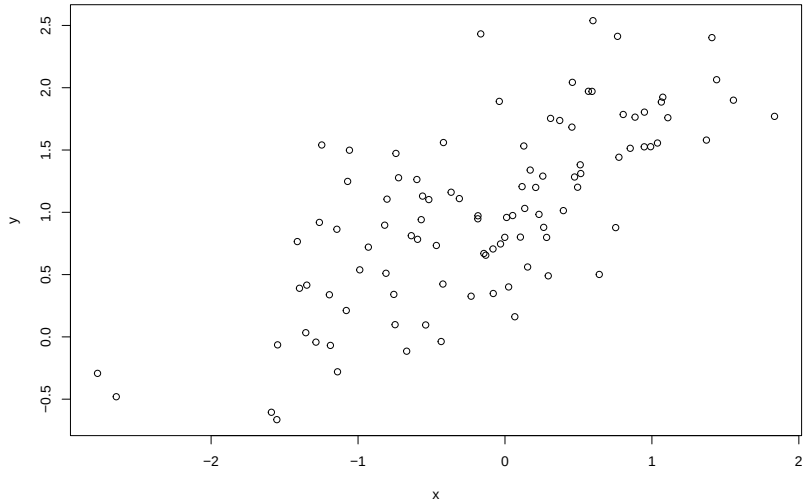
Least-Squares Regression

If we want to measure the linear association between two continuous variables, we can use the *correlation coefficient*.

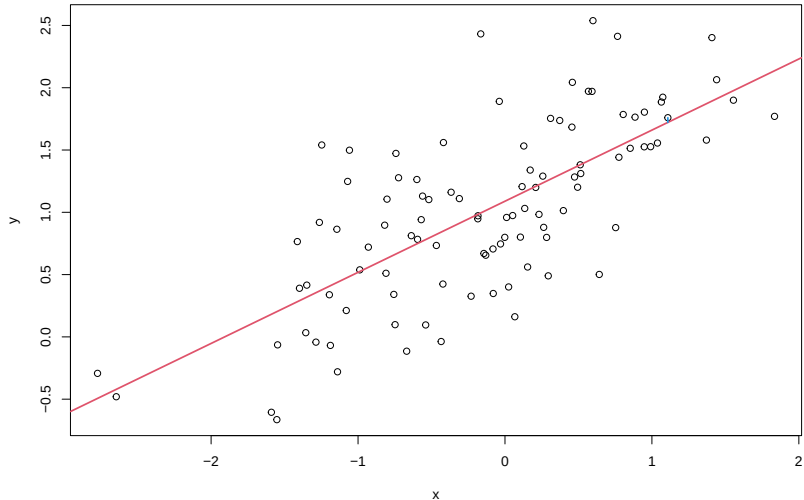
An alternative way to look at the association between two variables is to *determine the best line that describe their relationship*.

What do we mean by "best"? In the context of regression, "best" means the line that minimizes *the sum of squared distances*.

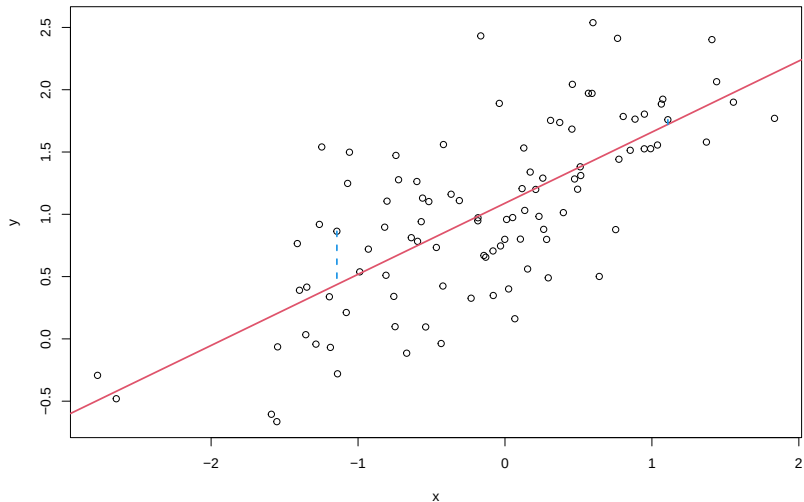
A simple example



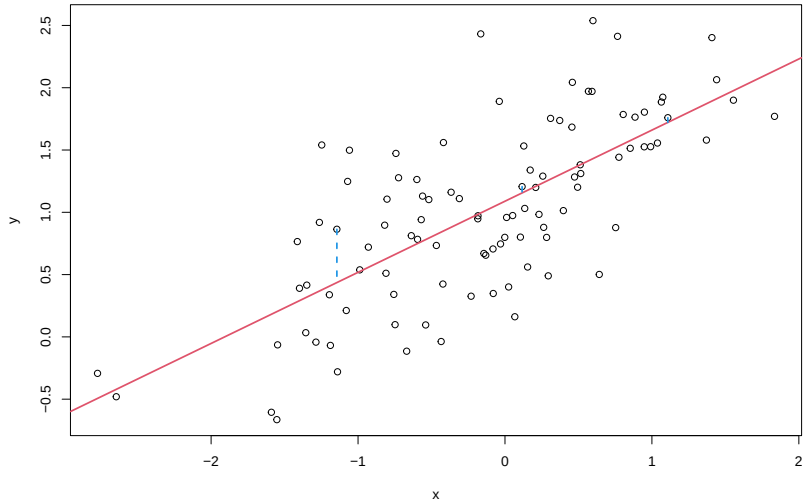
A simple example



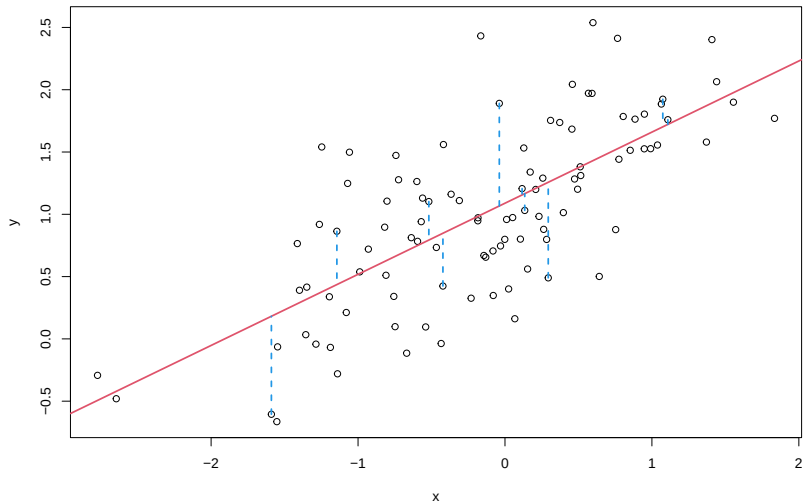
A simple example



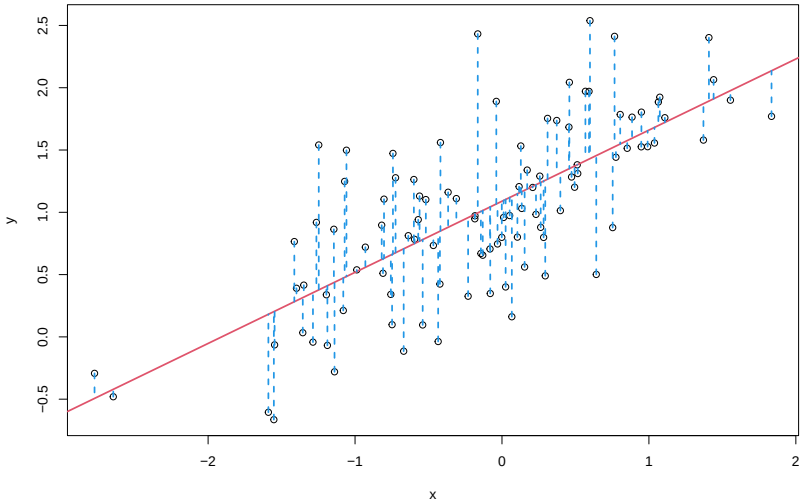
A simple example



A simple example



A simple example



Mathematically, we can describe the linear relation between the two quantities x and y as

$$\hat{y} = \alpha + \beta x.$$

Finding the best line corresponds to minimize the *sum of squared distances* between the values of y and the values of $\alpha + \beta x$, i.e.

$$\min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

To minimize this quantity, as usual, we compute the partial derivatives.

$$\frac{\partial}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0,$$
$$\frac{\partial}{\partial \beta} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)(x_i) = 0,$$

We can show that

$$\alpha = \bar{y} - \beta \bar{x},$$
$$\beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Least-Squares Regression

The line

$$\widehat{y} = \alpha + \beta x$$

is called the *least-squares regression* line.

For any pair (x_i, y_i), $\widehat{y}_i$ is called the *fitted value* (or predicted value).

α is called the *intercept* and β the *slope*.

The quantities $y_i - \widehat{y}_i$ are called the *residuals*.

By definition of least-square regression, *the sum of the residuals is equal to zero*.

As a consequence, *the mean of the fitted values is equal to the mean of the observed values*.

$$\begin{aligned}\bar{\hat{y}} &= \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i - \bar{y} + \bar{y} \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) + \bar{y} = \bar{y}.\end{aligned}$$

Relation between correlation and regression

Although correlation and regression can be both used to measure linear relationship between variables, there is a major difference between the two techniques.

In correlation, x and y have the same role, and correlation is a symmetric operator, i.e.

$$cor(x, y) = cor(y, x).$$

In linear regression, we use the values of x to "explain" the values of y.

y is called *dependent variable* or *response variable*.

x is called *independent variable*, *explanatory variable* or *covariate*.

Variance Explained

Our goal is to use the values of x to explain the values of y, so a natural question is to ask what proportion of variance is explained by the model.

The proportion of variance explained is called the R^2 or *R-squared coefficient*, and it is equal to the square of the Pearson correlation coefficient.

Proportion of Variance Explained

The R-squared coefficient can be used as a diagnostic for the regression model.

A large value of R-squared means that the linear model is doing a good job at explaining y using the values of x.

A small value of R-squared means that the linear model is failing to explain the values of y using the values of x. This may be because there is no strong relation between x and y or because the relation is not linear.

Why regression?

Compared to correlation, regression offers the following advantages.

- The values of x can be used to predict y.
- The analysis is not limited to linear association, but includes polynomial dependencies.
- The analysis is not limited to two variables and regression can be used to study the effects of many covariates on the response.

The ideas of linear regression can be applied in the case where we have more than one covariate.

Instead of fitting the line $\widehat{y} = \alpha + \beta x$, we can consider the more general form

$$\widehat{y} = \alpha + \beta_1 x_1 + ... + \beta_p x_p,$$

where p is the number of covariates and $x_1, ..., x_p$ are p variables.

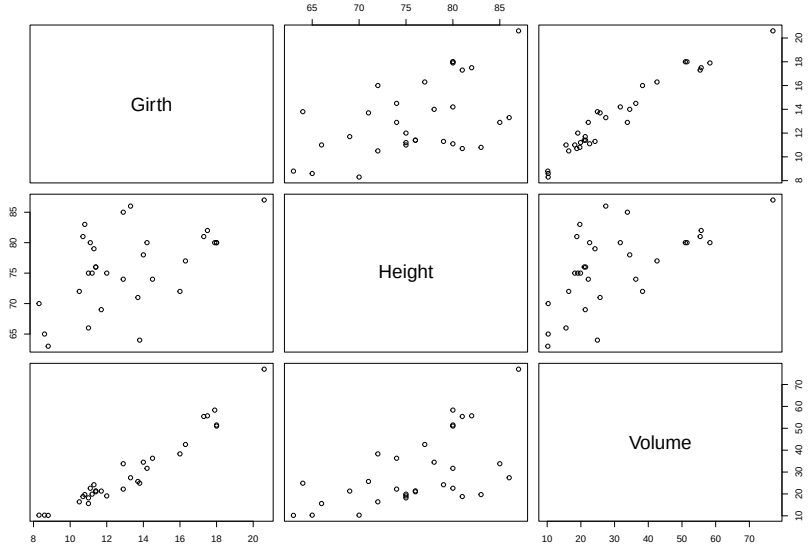
For symmetry, the intercept α is often denoted with β_0.

Example: Cherry trees

```
head(trees)
```

```
  Girth Height Volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
4  10.5     72   16.4
5  10.7     81   18.8
6  10.8     83   19.7
```

Example: Cherry trees



Regression as a statistical model

Up until now, we have only defined the regression line as a mathematical entity that we can compute starting from a vector y and a vector x.

However, in inferential statistics, both Y and X are random variables, and what we observe are the values of our random sample from the population.

This means that what we want to *infer* the true relationship between X and Y in the population starting from the relationship that we observe in the sample.

In other words, there is a *true parameter* β in the population, and we need to *estimate* it with the values of our sample.

In the context of inference, regression is often referred to as *the linear model*.

Linear Models: simple linear regression

Suppose that, on n cases, we observe a continuous response y_i and one explanatory variable x_i.

A simple linear model postulates that the observations are realizations of independent rv's Y_i such that:

$$E(Y_i|X_i = x_i) = \beta_0 + \beta_1 x_i, \;\; i = 1, \;\; \ldots, n.$$

Therefore, observations are realizations of independent rv's whose averages lie on the straight line $\beta_0 + \beta_1 x$.

Another way to look at this is:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \;\; i = 1, \dots, n,$$

where ε_i are random errors with mean 0. Often, it is assumed that the variance of the error terms, for all $i = 1, \dots, n$, is the same, i.e., errors are *homoschedastic*.

Interpretation of slope

Consider two subjects:

- A has covariate value x_A
- B has covariate value $x_B = x_A + 1$

The expected difference in the response of the two subjects is:

$$E(Y_B) - E(Y_A) = \beta_0 + \beta_1(x_A + 1) - \beta_0 - \beta_1 x_A = \beta_1,$$

i.e., β_1 is the effect of a one unit increase in x.

Multiple linear regression

In multiple linear regression, rather than one predictor, we have more than one, p, say.

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \;\; i = 1, ..., n$$

Multiple linear regression

With multiple linear regression, the matrix notation is particularly useful.

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ 1 & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{pmatrix}$$

and the model can be written as $Y = X\beta + \varepsilon$.

Interpretation of parameters

Consider two subjects:

- A has covariate values $x_{A1}, x_{A2}, \ldots, x_{Ap}$
- B has covariate values $x_{A1}, x_{A2} + 1, \ldots, x_{Ap}$

Expected difference in the response of the two subjects:

$$\begin{aligned} E(Y_B) - E(Y_A) = \; & \beta_0 + \beta_1 x_{A1} + \beta_2 (x_{A2} + 1) + \ldots + \beta_p x_{Ap} + \\ & -\beta_0 - \beta_1 x_{A1} - \beta_2 x_{A2} - \ldots - \beta_p x_{Ap} \\ = \; & \beta_2, \end{aligned}$$

i.e., β_2 is the effect of a one unit increase in x_2 for fixed level of the other predictors.

Coefficients $\beta_1, \ldots, \beta_p$ are called (partial) regression coefficients because they "allow" for the (partial) effect of other variables.

The linear model (in matrix notation)

Given a set of n independent observations, and p covariates, with $n > p$, the linear model is defined, in matrix notation, as

$$Y = X\beta + \varepsilon,$$

where

- Y is a $n \times 1$ vector containing the response variable.
- X is a $n \times p$ matrix of covariates, called the *design matrix*.
- β is a $p \times 1$ vector of regression parameters.
- ε is a $n \times 1$ vector of random errors.

As with all the statistical models, it is always a good idea to identify the observed / unobserved quantities, as well as the random / fixed quantities.

In this case,

- X and Y are *observed random variables*.
- ε is an *unobserved random variable*.
- β is a set of *unknown parameters* that we will need to estimate.

In some books, X is not considered a random variable, but a fixed quantity. In this course, we will consider it random, but we will *condition* our analysis to the observed values of X.

For practical purposes, we do not care about the random variation of X.

A note on linearity

The term *linear* in the linear model refers to the linearity of the model *with respect to the parameter* β.

It does not imply that we have linearity in X and X may contain non-linear transformation of the data. For instance the following models are valid linear models.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 \log X + \varepsilon$$

A note on the design matrix

We will see that in order to estimate β we need to compute the inverse of the design matrix X.

This means that X needs to be invertible. To insure that, we need the design matrix to be of *full rank* or of rank p.

The rank of a matrix is essentially the number of columns that are linearly independent.

Hence, to ensure that we can compute $\hat{\beta}$, we will need all the columns of X to be linearly independent.

This is often referred to as the *non collinearity* condition.

Assumptions

As with all the statistical models, we need some assumptions.

1. There is a linear relation between Y and X.
2. The error terms ε are i.i.d. with mean 0 and constant variance:
   - $E[\varepsilon] = 0$;
   - $Var(\varepsilon) = \sigma^2$ (homoscedasticity).
3. The error terms ε are independent of X: $\varepsilon \perp\!\!\!\perp X$.

As a consequence of these assumptions, we can see that the conditional expectation of Y is

$$E[Y|X] = X\beta.$$

Explanation

Here the idea is the that system under study really is (approximately) linear, and we are interested in the coefficients per se. For example, governments have linear models of the economy and use them to understand the effects of policy (e.g. tax, interest rate) changes. It is often of particular interest to find a minimal set of explanatory variables.

Prediction

Here a model is a convenient means to create predictions for new cases.The only interest is in the quality of the predictions.

To solve now

1. Fitting the model: how do we estimate (β, σ^2)?
   - Least squares
   - Other methods
2. Inference: what can we say about β (rarely, about σ^2) based on the n observations?

Ordinary Least Squares

Ordinary Least Squares

We do not observe the true parameter β and we need to estimate it from the data.

Luckily, we already know how to estimate β, as the value that minimizes the *sum of the squares of the differences between Y and $X\beta$.*

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2.$$

The value $\beta^\top = (\beta_0, \beta_1)$ such that:

$$\widehat{\beta} = \text{argmin}_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

is called the *ordinary least squares* (OLS) estimator.

Least squares: simple regression

In the simple linear model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \;\; i = 1, \ldots, n$, minimization of the residuals sum of squares

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

yields:

1. $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$;
2. $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{Cov(y, x)}{Var(x)}$.

Least squares: simple linear regression (cntd)

From the previous results, we get that estimated values $\hat{y}_i$ can be computed as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

The quantities $e_i = y_i - \hat{y}_i, i = 1, \ldots, n$, i.e., the differences between observed and estimated values, are called *residuals*.

To estimate the variance of the random terms, it seems natural to consider the variability of the residuals. An unbiased estimator for σ^2 is given by

$$s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}.$$

Example: cherry tree

For the Cherry tree data, assume to model the volume (y) of the trees as a function of girth (x) with the simple linear model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \, , i = 1, \dots , 31.$$

We have $\bar{x} = 13.25$, $\bar{y} = 24.20$, $Var(x) = 9.847914$, $Var(y) = 270.2028$, $Cov(y, x) = 49.89$.

From previous expressions, we have:

$$\hat{\beta}^\top = (\hat{\beta}_0, \hat{\beta}_1) = (-36.943, 5.066).$$

Simple algebra shows that:

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2 - \hat{\beta}_1 \sum (x_i - \bar{x})^2.$$

From previous result, we have $s^2 = 18.07950$.

Least squares: multiple regression

In the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i, \;\; i = 1, ..., n$$

minimization of the residuals sum of squares

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip}\right)^2 = (y - X\beta)^\top (y - X\beta)$$

yields:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

From this we get:

$$\hat{y} = X\hat{\beta},$$

$$e = y - \hat{y},$$

OLS estimate: derivation

The least-square equation is easier in matrix notation:

$$(Y - X\beta)^\top (Y - X\beta)$$

The derivative in matrix notation is

$$2X^\top (Y - X\beta) = 0$$

which leads to

$$X^\top X\beta = X^\top Y$$

and

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y$$

Example: cherry tree

For the Cherry tree data, assume to model the volume (y) of the trees as a function of girth (x_1) and height (x_2) with the linear model
$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i\,, \quad i = 1, \dots, 31.$

We have

$$y = \begin{pmatrix} 10.3 \\ 10.3 \\ \dots \\ 77.0 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 8.3 & 70 \\ 1 & 8.6 & 65 \\ \dots & \dots & \dots \\ 1 & 20.6 & 87 \end{pmatrix}.$$

Example: cherry tree

This yields

$$X^\top X = \begin{pmatrix} 31.0 & 410.70 & 2356.0 \\ 410.7 & 5736.55 & 31524.7 \\ 2356.0 & 31524.70 & 180274.0 \end{pmatrix}, \qquad X^\top y = \begin{pmatrix} 935.30 \\ 13887.86 \\ 72962.60 \end{pmatrix},$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = (X^\top X)^{-1} X^\top y = \begin{pmatrix} -57.9876589 \\ 4.7081605 \\ 0.3392512 \end{pmatrix}.$$

We can prove that $\hat{\beta}$ is conditionally unbiased, i.e.,

$$E[\hat{\beta}|X] = \beta.$$

The variance of the OLS estimator is

$$Var(\hat{\beta}|X) = \sigma^2(X^\top X)^{-1}.$$

The OLS estimator is conditionally unbiased

$$\begin{aligned}\hat{\beta} &= (X^\top X)^{-1} X^\top (X\beta + \varepsilon) \\ &= (X^\top X)^{-1} X^\top X\beta + (X^\top X)^{-1} X^\top \varepsilon \\ &= \beta + (X^\top X)^{-1} X^\top \varepsilon.\end{aligned}$$

Hence,

$$\hat{\beta} = \beta + \eta \quad \text{where} \quad \eta = (X^\top X)^{-1} X^\top \varepsilon.$$

The OLS estimator is conditionally unbiased

To show that $E[\hat{\beta}|X] = \beta$, we need to show that $E[\eta|X] = 0$.

$$E[\eta|X] = (X^\top X)^{-1} X^\top E[\varepsilon|X].$$

Since $\varepsilon \perp\!\!\!\perp X$, conditioning on X does not influence the distribution of ε, and we know by assumption that $E[\varepsilon] = 0$. Hence,

$$E[\hat{\beta}|X] = \beta.$$

Standard Error of the OLS estimator

As it is true in general for estimation, it is not enough to have a *point estimate* of the regression parameter, but we want to know what is the *distribution* of the estimator.

First, we need to compute the standard errors of $\widehat{\beta}$.

In order to do that we have to remember the assumptions of the linear model, especially those related to the error term ε.

We denote with Σ the variance/covariance matrix of the error term. In particular,

$$\Sigma_{i,j} = Cov(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}$$

Hence, we can write

$$\Sigma = \sigma^2 I.$$

Standard Error of the OLS estimator

As a consequence, the conditional variance of Y is

$$Var(Y|X) = Var(X\beta + \epsilon|X) = Var(\epsilon) = \sigma^2 I$$

Variance of a linear combinations

In matrix notation, the variance of a linear combination AY can be computed as

$$Var(AY) = AVar(Y)A^\top$$

Since $\widehat{\beta}$ is a linear combination of Y, we can use the same rule to compute its variance.

Standard Error of the OLS estimator

$$\begin{aligned} Var(\hat{\beta}|X) &= Var((X^\top X)^{-1}X^\top Y|X) \\ &= (X^\top X)^{-1}X^\top Var(Y|X)((X^\top X)^{-1}X^\top)^\top. \end{aligned}$$

Note that:

- $X^\top X$ is symmetric;
- if A is symmetric $A^\top = A$;
- $(AB)^\top = B^\top A^\top$.
- $(A^\top)^\top = A$.

Hence,

$$((X^\top X)^{-1}X^\top)^\top = X(X^\top X)^{-1}$$

Standard Error of the OLS estimator

Therefore,

$$\begin{aligned} Var(\hat{\beta}|X) &= (X^\top X)^{-1} X^\top Var(Y|X) X (X^\top X)^{-1} \\ &= (X^\top X)^{-1} X^\top \sigma^2 I X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1} X^\top X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1}. \end{aligned}$$

Hence, the diagonal of the square root of this matrix contains the standard errors of β.

How to estimate σ^2?

If we could observe ε, we could simply estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2.$$

Since we do not observe ε, we can use the residuals $e = Y - X\hat{\beta}$.

We also need to correct for the degrees of freedom (more on this later), to get our estimator:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2.$$

This estimator is *conditionally unbiased*, i.e.,

$$E[\hat{\sigma}^2|X] = \sigma^2.$$

A note on $n - p$

The fact that we divide by $n - p$ derives from the proof that $\hat{\sigma}^2$ is unbiased.

This proof is quite complicated mathematically, so we will skip it for now.

However, it derives from the geometric interpretation of regression, which we will see if we have time at the end of the course.

For now, notice that because we assumed that $n > p$, we do not have problems in estimating σ^2.

Also, note the similarity with the usual estimator of the variance: Here, instead of $n - 1$ we use $n - p$, because we have p parameters, and hence $n - p$ degrees of freedom.

A note on the assumptions

Note which assumptions we *did not need* to get to these results.

1. We did not assume normality!
2. We did not assume independence of the columns of X, just non-collinearity.

These results ensure that linear models can be applied in a lot of different settings, without requiring too stringent assumptions.

R: the `lm()` function

As usually is the case, R has a convenient function that we can use, which is called `lm()` for linear model.

We will see more about it later, but for now, you can see how we can use it to compute the values of β.

Note that we use the formula syntax to indicate that we have a response variable that is a function of the covariates.

R: the `lm()` function

```
lm(Volume ~ Girth + Height, data=trees)
```

```
Call:
lm(formula = Volume ~ Girth + Height, data = trees)

Coefficients:
(Intercept)        Girth       Height
   -57.9877       4.7082       0.3393
```

The Normal Linear Model

Normality assumption

Remember that for now, we did not make any assumptions about the distribution of the data.

Indeed, the assumptions of the linear model that we have made so far are:

1. There is a linear relation between Y and X; with X full-rank and $p < n$.
2. The error terms ε are i.i.d. with mean 0 and constant variance:
   - $E[\varepsilon] = 0$;
   - $Var(\varepsilon) = \sigma^2$ (homoscedasticity).
3. The error terms ε are independent of X: $\varepsilon \perp\!\!\!\perp X$.

For the next results to hold, we need to restrict our model to the following assumption:

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Normality assumption

We need the normality assumption in order to make inference on the coefficients of the model, i.e., to create confidence intervals and test hypotheses on the parameter β.

The main results that we will see are the t-test and the F-test.

In the simple linear model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \;\; i = 1, \ldots, n,$$

with independent $\varepsilon_i \sim N(0, \sigma^2)$, we have:

1. $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2})\right)$;
2. $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$;
3. $(n-2)s^2/\sigma^2 \sim \chi^2_{n-2}$;

From the previous results, it is easy to see that the rv

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{V(\hat{\beta}_j)}} \sim N(0,1), \quad j = 1, 2.$$

As $V(\hat{\beta}_j)$ involves σ^2, this quantity might need to be estimated. If the sample unbiased variance s^2 is used to estimate σ^2, then

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}(\hat{\beta}_j)}} \sim t_{n-2}, \quad j = 1, 2.$$

This allows to make inference (CI and tests) on the parameters β.

Example: cherry tree

For the Cherry tree data, we have:

$$\hat{\beta}_0 = -36.943, \quad \sqrt{\hat{V}(\hat{\beta}_0)} = 3.3651,$$

and

$$\hat{\beta}_1 = 5.066, \quad \sqrt{\hat{V}(\hat{\beta}_1)} = 0.2474.$$

Want to test $H_0 : \beta_1 = 0$. Under H_0, we have

$$T_1 = \frac{\hat{\beta}_1 - 0}{0.2474} \sim t_{n-2}.$$

The observed value for T_1 is $t_1^{obs} = 20.48$ and $p = P(|T_1| > t_1^{obs} | \beta_1 = 0) \approx 0$. Such small value of p testifies strong evidence against the null-hypothesis.

A 95% exact confidence interval for β_1 is given by:

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} \sqrt{\hat{V}(\hat{\beta}_1)} = (4.55991, 5.57189).$$

Testing $H_0 : \beta_1 = 0$

Consider again the hypothesis $H_0 : \beta_1 = 0$. Testing such hypothesis is equivalent to comparing two models, a "full" and a "reduced" model.

- Full: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Reduced: $Y_i = \beta_0 + \varepsilon_i$

Testing $H_0 : \beta_1 = 0$

Under H_0,

$$(n-2)s_F^2 = e_F^\top e_F \sim \sigma^2 \chi^2_{n-2}$$

$$(n-1)s_R^2 = e_R^\top e_R \sim \sigma^2 \chi^2_{n-1}$$

$$e_R^\top e_R - e_F^\top e_F \sim \sigma^2 \chi^2_1$$

$$F = \frac{(e_R^\top e_R - e_F^\top e_F)/1}{e_F^\top e_F/(n-2)} \sim F_{1,n-2}$$

Reject H_0 at level α if $F^{obs} > F_{1,n-2;1-\alpha}$. Note that, in this case, $F = T_1^2$.

In the general model

Previous results extend to the general model with p regressors. It can be proved that:

- $\hat{\beta} \sim N_p(\beta, \sigma^2(X^\top X)^{-1})$;
- $(n-p)s^2/\sigma^2 \sim \chi^2_{n-p}$;
- $\hat{\beta}$ and s^2 are independent rv.

We have, for $j = 1, \ldots, p$

1. $\hat{\beta}_j \sim N(\beta_j, \sigma^2[(X^\top X)^{-1}]_{jj})$;
2. $Z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{V(\hat{\beta}_j)}} \sim N(0,1)$;
3. $T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}(\hat{\beta}_j)}} \sim t_{n-p}$.

Testing $H_0 : \beta_j = 0$

Can be tested with a t-test:

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}(\hat{\beta}_j)}} \sim t_{n-p}.$$

Equivalently, using the comparison of models idea, we can use:

$$F = \frac{(e_R^\top e_R - e_F^\top e_F)/1}{e_F^\top e_F/(n-p)} \sim F_{1,n-p}$$

Reject H_0 at level α if $F^{obs} > F_{1,n-3;1-\alpha}$.

Note that, in this case, $F = T_j^2$.

Overall goodness of fit

Consider the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i.$$

Want to test the overall goodness of fit

$$H_0 : \beta_1 = \beta_2 = 0.$$

Two models

- Full: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$
- Reduced: $Y_i = \beta_0 + \varepsilon_i$

F-statistic, under H_0 (details similar as before):

$$F = \frac{(e_R^\top e_R - e_F^\top e_F)/2}{e_F^\top e_F/(n-3)} \sim F_{2,n-3}$$

Reject H_0 at level α if $F^{obs} > F_{2,n-3;1-\alpha}$.

Example: cherry trees

```
data(trees)
fit <- lm(Volume ~ Girth + Height, data=trees)
summary(fit)
```

```
Call:
lm(formula = Volume ~ Girth + Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948, Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Dropping subsets

Suppose we have the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \;\; i = 1, ..., n,$$

and we want to test whether we can simplify the model by dropping k variables, i.e. want to test

$$H_0 : \beta_{j_1} = \beta_{j_2} = ... = \beta_{j_k} = 0$$

Two models

- Full: above
- Reduced: model with the k columns $x_{j_1}, x_{j_2}, ..., x_{j_k}$ omitted from the design matrix.

Dropping subsets

F-statistic, under H_0 (details similar as before):

$$F = \frac{(e_R^\top e_R - e_F^\top e_F)/(p-k)}{e_F^\top e_F/(n-p)} \sim F_{p-k,n-p}.$$

Reject H_0 at level α if $F^{obs} > F_{p-k,n-p;1-\alpha}$.