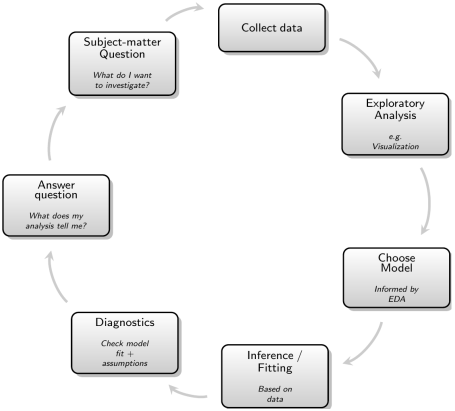
# MA50260 Statistical Modelling

## Lecture 2: Linear Regression Definition

Ilaria Bussoli

February 9, 2024

# Philosophy of Statistical Modelling



# Motivating Example: Birth Weights

Recorded weight and gestational age for 24 newborn babies.

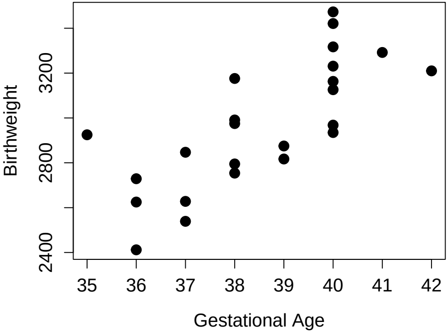



Figure 1:Birthweight (grams) vs gestational age (weeks) for 24 children.

# Motivating Example: Birth Weights

A wide range of subject-matter questions may arise, for instance

1. Is there evidence of a positive relationship between birth weight and gestational age? If so, what is this relationship?
2. Can we predict the birth weight for a child born at 34 weeks? What is a 95% confidence interval for this prediction?

**Linear regression helps us to address these questions.**

# Motivating Example: Birth Weights

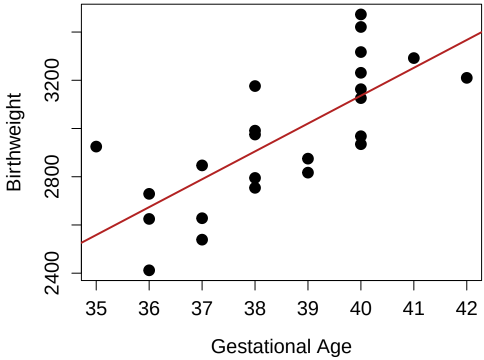



Figure 2:Birthweight (grams) vs gestational age (weeks) for 24 children. The red line shows the estimated linear regression model.

# Simple Linear Regression I

**Definition 2.1 (Response variable).** A response variable $Y$ is a random variable whose distribution depends on the value of another variable.

**Definition 2.2 (Explanatory variable).** An explanatory variable $x$ is considered to be **non-random** and to influence the outcome of the response variable.

We assume that some of the variability in the response variable $Y$ can be explained by a linear relationship between $Y$ and $x$.

# Simple Linear Regression II

Let $Y_i$ and $x_i$ denote the response variable and explanatory variable for observation $i$.

We then model

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

where $n$ is the number of individuals and $\epsilon_1, \ldots, \epsilon_n$ are termed **regression residuals**.

# Simple Linear Regression II

Let $Y_i$ and $x_i$ denote the response variable and explanatory variable for observation $i$.

We then model

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

where $n$ is the number of individuals and $\epsilon_1, \ldots, \epsilon_n$ are termed **regression residuals**.

We assume that

1. The residuals are mutually independent;
2. The residuals have zero mean and common variance $\sigma^2$;
3. The residuals each follow a normal distribution,

$$\epsilon_i \sim \mathrm{Normal}(0; \sigma^2), \qquad i = 1, \ldots, n.$$

# Simple Linear Regression III

We could also have said

1. $Y_i \sim \text{Normal}\left(\beta_1 + \beta_2 x_i\,,\,\sigma^2\right), \qquad i = 1, \ldots, n.$
2. $Y_1, \ldots, Y_n$ are mutually independent.

# Simple Linear Regression III

We could also have said

1. $Y_i \sim \mathrm{Normal}\left(\beta_1 + \beta_2 x_i\,,\, \sigma^2\right), \qquad i = 1, \ldots, n.$
2. $Y_1, \ldots, Y_n$ are mutually independent.

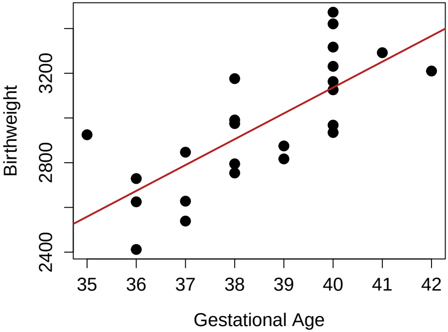
The unknown parameters are

- The **regression parameters**, or **coefficients**, $\beta_1$ and $\beta_2$;
- The **residual variance** $\sigma^2$;
- The **residuals** $\epsilon_1, \ldots, \epsilon_n$.

# Example: Birth Weights

The estimated regression line for birth weight $Y_i$, conditional on gestational age $x_i$, is

$$\mathbb{E}(Y_i) = -1485 + 115.5x_i.$$



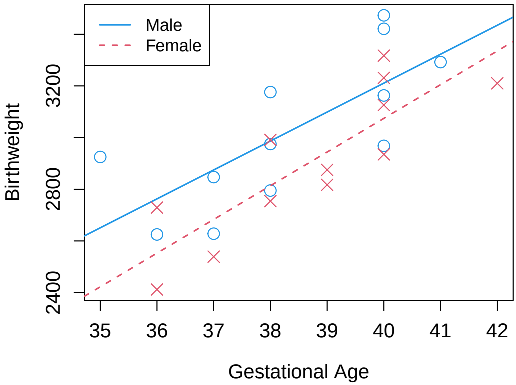
# Multiple Linear Regression: Motivation

We are further given the sex-at-birth of each newborn.

This motivates new subject-matter questions:

1. Do males and females gain weight at different rates?
2. Do we need both gestational age and sex-at-birth to explain variability in birth weights, or is one of these sufficient?

# Multiple Linear Regression: Motivation



# Multiple Linear Regression: Definition

The setup is very similar to simple linear regression, but

1. Each individual has a single response variable $Y_i$ and a vector of explanatory variables $(x_{i,1}, \ldots, x_{i,p})$.
2. There are $p$ regression coefficients $\beta_1, \beta_2, \ldots, \beta_p$.

**Definition 2.3 (Multiple linear regression model).** For $i = 1, \ldots, n$, we model

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} + \epsilon_i,$$

where

$$\epsilon_i \sim \text{Normal}(0, \sigma^2)$$

and $\epsilon_1 \ldots, \epsilon_n$ are mutually independent.

# Multiple Linear Regression: Matrix Notation

We often write a linear regression model as

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \underline{\epsilon},$$

where $\underline{\epsilon} \sim \mathrm{MVN}_n(0, \sigma^2 I_n)$ and $I_n$ is the $n \times n$ identity matrix.

# Multiple Linear Regression: Matrix Notation

We often write a linear regression model as

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \underline{\epsilon},$$

where $\underline{\epsilon} \sim \mathrm{MVN}_n(0, \sigma^2 I_n)$ and $I_n$ is the $n \times n$ identity matrix.

The different terms are

- The **response vector** $\mathbf{Y} = (Y_1, \ldots, Y_n)$;
- The **design matrix** $\mathbf{X}$ whose columns correspond to explanatory variables and whose rows correspond to subjects;
- The **residual vector** $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$;
- The **vector of coefficients** $\underline{\beta} = (\beta_1, \ldots, \beta_p)$.

# Factors

Explanatory variables may be continuous or discrete, qualitative or quantitative.

We discuss two types of explanatory variable.

- A **covariate** is a *quantitative* explanatory variable.
- A **factor** is a *qualitative* explanatory variable. The possible values for the factor are called **levels**.

# Factors

Explanatory variables may be continuous or discrete, qualitative or quantitative.

We discuss two types of explanatory variable.

- A **covariate** is a *quantitative* explanatory variable.
- A **factor** is a *qualitative* explanatory variable. The possible values for the factor are called **levels**.

Factors are represented by indicator variables in a linear regression model.

Eg., to include sex-at-birth, we create the indicator variables

$$x_{i,1} = \begin{cases} 1 & \text{if child } i \text{ is male} \\ 0 & \text{if child } i \text{ is female} \end{cases} \qquad x_{i,2} = \begin{cases} 1 & \text{if child } i \text{ is female} \\ 0 & \text{if child } i \text{ is male} \end{cases}$$

# Example: Birth Weights

Let $x_{i,3}$ refer to the gestational age of child $i$.

Three possible models which account for sex-at-birth of a child are

$$
\begin{aligned}
\mathbb{E}(Y_i) &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} \\
\mathbb{E}(Y_i) &= \beta_1 + \beta_2 x_{i,1} x_{i,3} + \beta_3 x_{i,2} x_{i,3} \\
\mathbb{E}(Y_i) &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} x_{i,3} + \beta_4 x_{i,2} x_{i,3}
\end{aligned}
$$

What is the interpretation of the different coefficients?