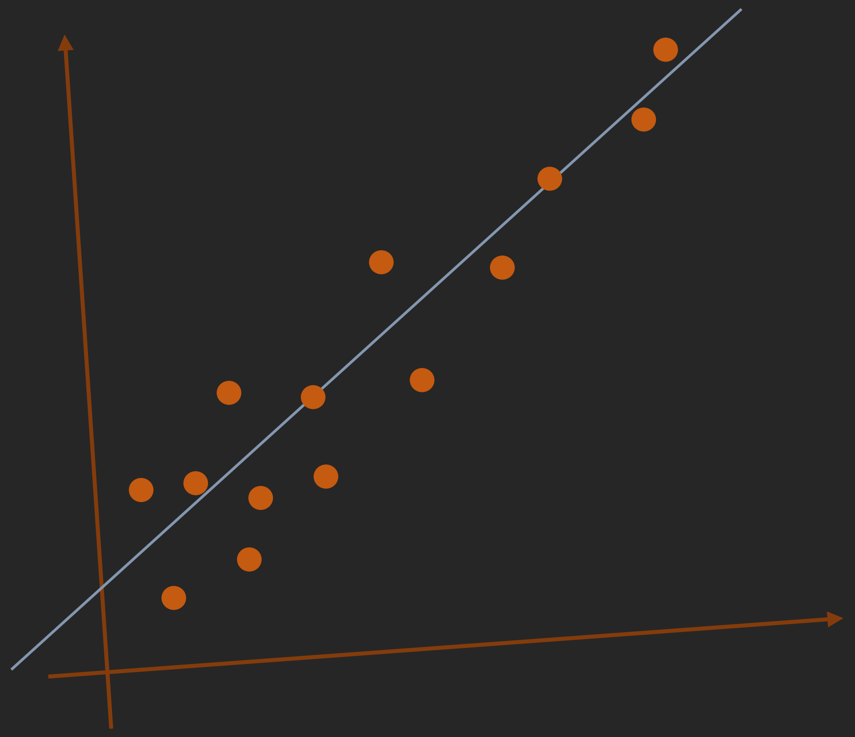
Applied Regression Analysis

STAT 4043/ STAT 5543



Introduction to simple linear regression – The model

Pratyaydipta Rudra

Simple linear regression

How good is a linear model (equation for a straight line) to explain the relationship of two variables?

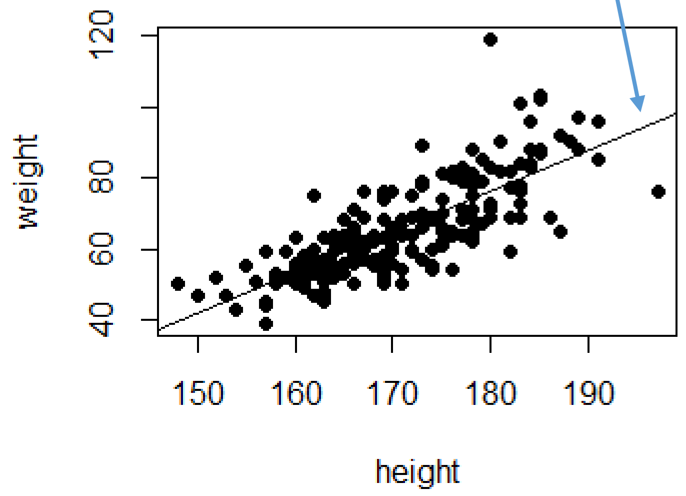
Simple linear regression model – an example

- $y = \beta_0 + \beta_1 x + \epsilon$
- Here, y is weight and x is height.

- y: Dependent variable/**Response**/Outcome
- x: Independent variable/**predictor**/explanatory variable
- ϵ: statistical error
- β_0: intercept,

β_1: slope/ regression coefficient.

True regression line, not known in practice.
Distance of the points from the line are errors.



Simple Linear Regression (SLR) model

- Because the model holds for all data points (x_i, y_i), we can write the model also as:

Predictor/

Independent variable

Response/Dependent variable

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

Intercept

Slope

Error

$$i = 1,2, \dots, n.$$

The model

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1,2, \dots, n.$
- The unknown parameters are β_0 and β_1 which need to be estimated.
- Estimating them involves finding the line that is 'best fit' through the points. What is best fit? We will see.
- Thus the estimation is also called "fitting the line". These estimates are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$.

True model vs fitted model?

Boardwork

The SLR model revisited

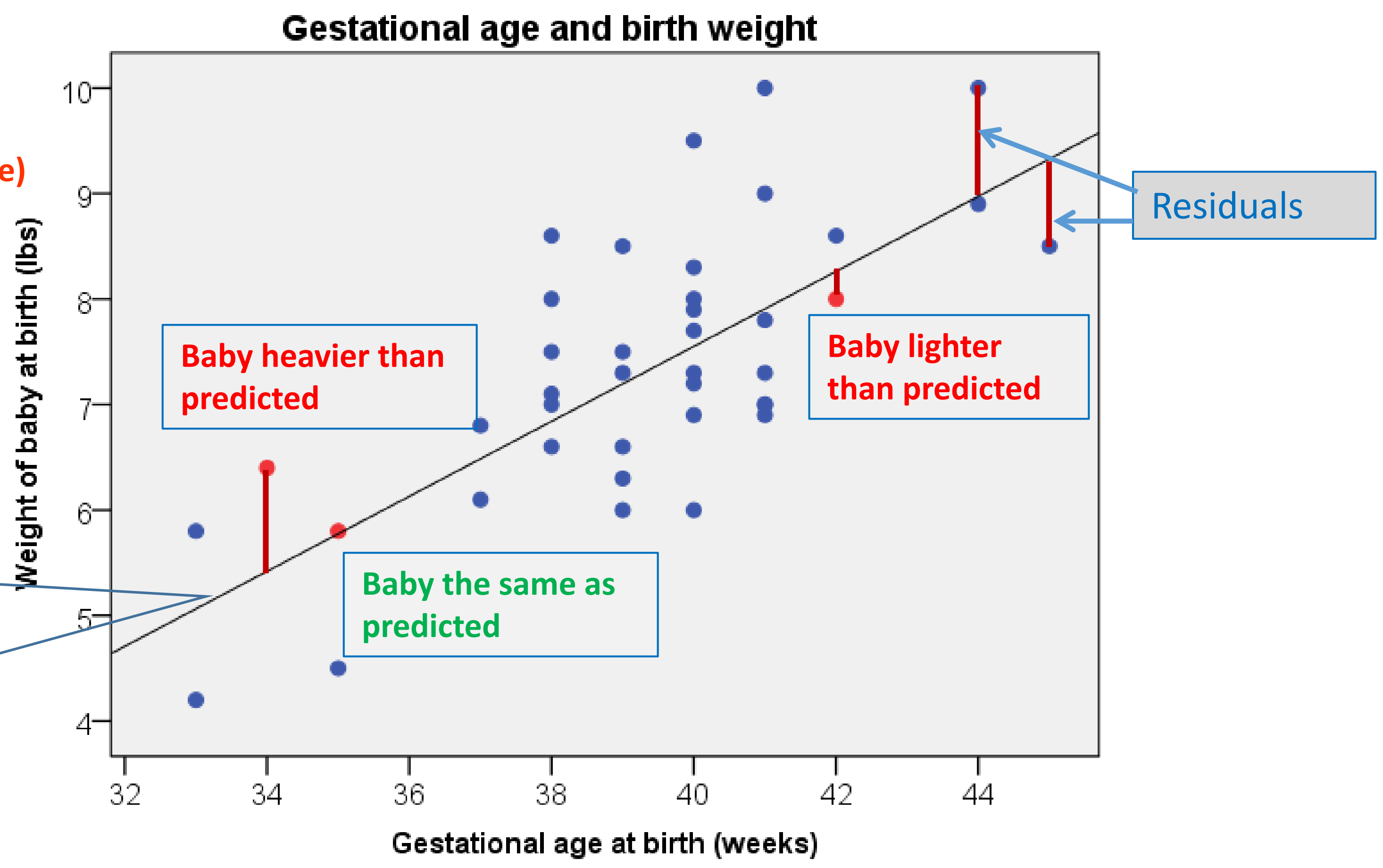
- True model

$$\underbrace{y_i}_{\text{Response/Dependent variable}} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} \underbrace{x_i}_{\text{Predictor/ Independent variable}} + \underbrace{\epsilon_i}_{\text{Error}}, \qquad i = 1,2,\ldots,n.$$

- Fitted model

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \underbrace{e_i}_{\text{Residuals}}, \qquad i = 1,2,\ldots,n.$$

Residuals (deviation from the fitted line)



y: Response
(dependent variable)

Equation of the Fitted Regression line
$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

x: Predictor (independent variable)

Source: www.statstutor.ac.uk

The concept of statistical error (ϵ_i)

- Functional vs statistical relationships.
  - Tax calculation – functional.
  - Calculating your final grades – functional.
  - Relation between height and weight – statistical.
- Random component, uncertainty.
- ‘Error’ does not mean mistake, it is an inherent part of statistical model.
- The unsystematic or unexplained part of the variability belongs to the random error component.

How much is random and how much can be explained?

- Why is the relationship between height and weight not exact?
- What are the other factors that result in variation in weight among people who have the same height?
- Can we explain some of this variability using other variables?
- Sure, but in most cases, we cannot explain all of it.

Assumptions about the two variables

- We assume that the response (y) is a **continuous random variable**.
- The predictor (x) can be continuous, discrete or even categorical.
- But the predictor is assumed to be **non-random/fixed**.

More examples

- An experiment was conducted to study the effectiveness of different dosage of a drug to reduce blood pressure. A random sample of 500 hypertension patients was chosen and everyone received one of the 10 different dosages. The dosage (x) and reduction in blood pressure (y) after taking the drug for a month was measured for each patient.
- An instructor wanted to see if there is any relationship of the final exam score (y) of students in his statistics class with whether the student had taken a linear algebra class ever (x).
- A study was conducted to find the effect of body mass index (x) of a person on whether the person ends up having a heart attack or not (y).

Many more examples in chapter 2 and 3 of your textbook

The examples also come with data sets. I have posted the data sets on Canvas for your convenience.

Assumptions about the error term

- It is reasonable to think that the error will be 0, on average.
- In fact, the idea is to model the average value of y as a function of x.

$$E(y) = \beta_0 + \beta_1 x$$

- It is assumed that the error term ϵ_i follows a normal distribution with mean 0 and variance σ^2.
- It is assumed that the ϵ_i terms are independent, implying that observations (y_i) are independent of each other.

Multiple Linear Regression

- If we want to include other variables that might have effect on the response, we will end up with a model like

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- This is called a multiple linear regression (MLR).

- Simple linear regression (SLR) is a special case. We will develop the techniques for the SLR first and study MLR later.

Fitting SLR model using R