

Lecture 31

(1)

Today we begin the topic of simple linear regression.

Basic Principles

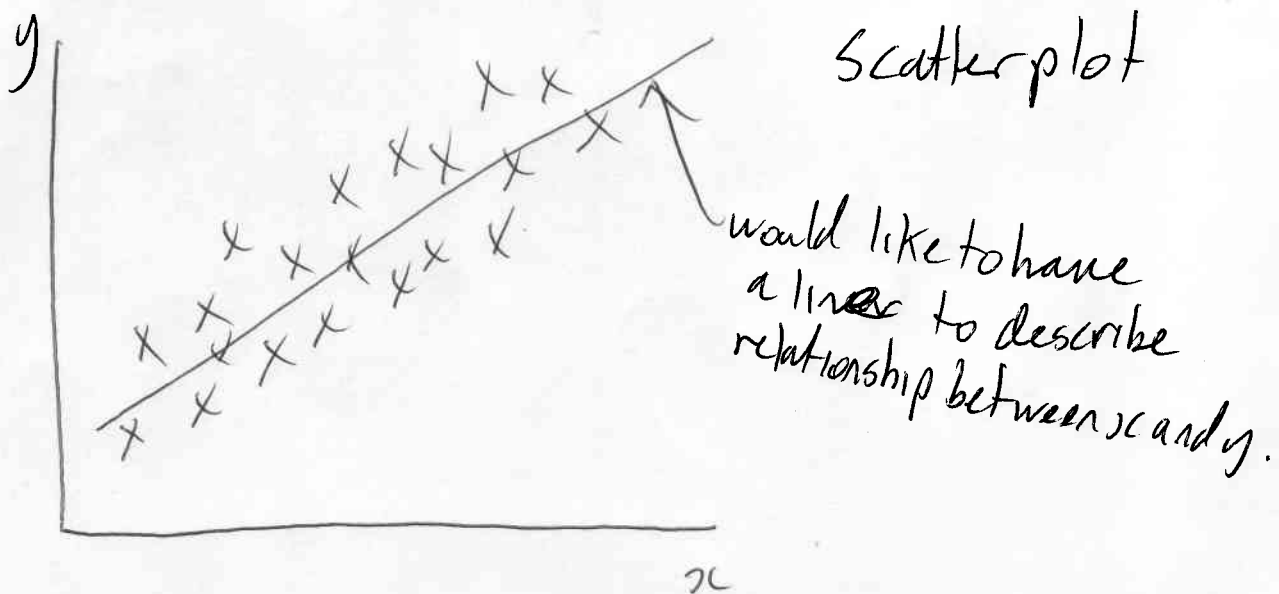
n observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

on x, y paired variables

x is an explanatory variable

y is a response variable

A visual way to examine the data



Of course we may visually interpret the scatterplot as we have discussed in earlier lectures. However,

It would be nice to have a mathematical ⁽²⁾ description of the relationship also. Often, the relationship observed in the scatterplot appears to be linear so it makes sense to describe the relationship in the form

$$\mu_y = \beta_0 + \beta_1 x$$

Annotations for the equation above:
- An arrow points from μ_y to "the mean of y for a given x ".
- An arrow points from β_0 to "intercept".
- An arrow points from β_1 to "slope".
- An arrow points from x to " x variable".

Note β_0, β_1 are called parameters.

However we do not observe the regression line when we draw a scatterplot because the observed y values vary about their means (μ_y).

So a stricter statistical description of the regression relationship should ^{also} include a description of the variability. We do this by including

a term in the model which we call the residual. It is the difference between the regression line and the observed y . We represent it with an ϵ (greek symbol epsilon). ③

In addition we describe the distribution of ϵ by assuming that it is $N(0, \sigma)$. This leads to the following formal description known as the

"Simple linear regression model"

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

Annotations for the equation above:

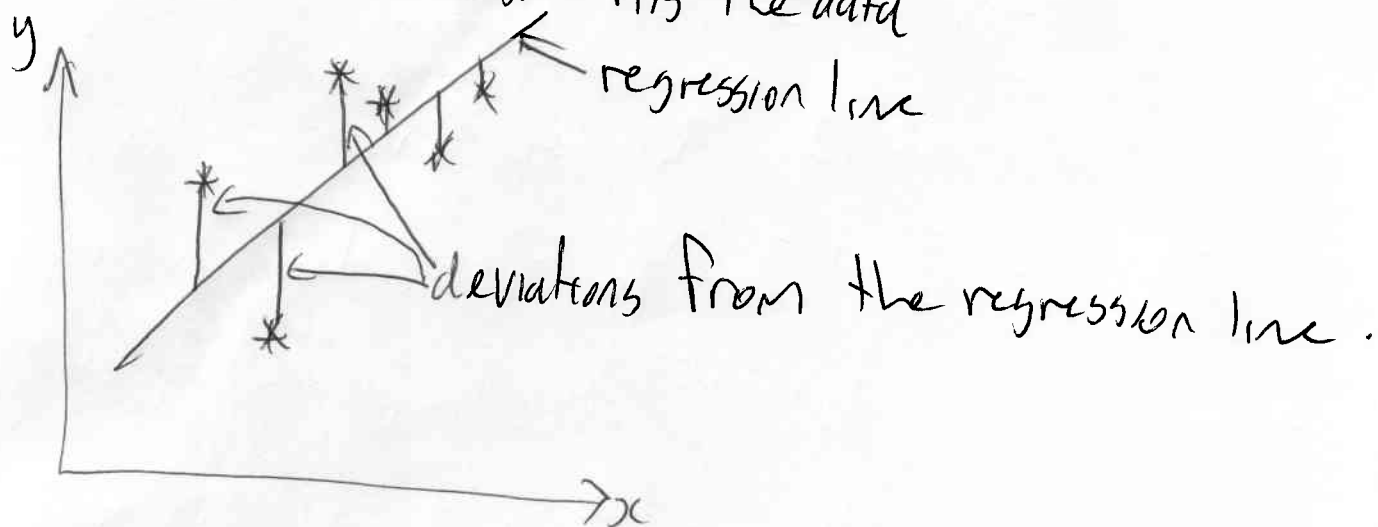
- Intercept parameter β_0
- slope parameter β_1
- observed explanatory x_i
- the residual ϵ_i
- observed response y_i

with the assumption that the ϵ_i are independent and distributed $N(0, \sigma)$ also a parameter.

(9)

Estimating β_0, β_1, σ (aka fitting the regression line)

Basic idea Find the equation of the line which best fits the data



The line which best fits the data will be the one which minimizes the deviations. What is conventionally done is to minimize the sum of the deviations squared. This is called the method of least squares. With lots of mathematics it can be shown that the estimates of β_0 and β_1 which we will call b_0 and b_1 respectively given by this method

are

$$b_1 = r \frac{s_x}{s_y}$$

standard deviation of x (pointing to s_x)
 standard deviation of y (pointing to s_y)

Correlation between x and y



this value is from above

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

mean of x (pointing to \bar{x})

↑
this is the mean of y

We call

$$\hat{y} = b_0 + b_1 x$$

the predicted/fitted regression line. And given a value of x say x^* we can predict a value of y (ie $\hat{y} = b_0 + b_1 x^*$)

We estimate the residuals using e_i (in place of e_i)

$$\begin{aligned} e_i &= \text{observed response variable} - \text{predicted response variable.} \\ &= y_i - \hat{y}_i \\ &= y_i - (b_0 + b_1 x_i) \end{aligned}$$

and use these to estimate σ .

(6)

In particular,

$$s = \hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}} \quad \left(= \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \right)$$