STAT241/251 Lecture Notes

Yew-Wei Lim

Ch11- Simple linear regression model

Introduction

Two variables may be related in a number of ways. If the relationship between two variables X and Y is linearly related, then we can consider using simple linear regression (Ch 11) as a model.

What does it mean for X and Y to be linearly related?

1 If large values of X are associated with large values of Y, or as X increases, the corresponding value of Y tends to increase, then this is a positive linear relationship between X and Y. Another way to think of this is: as the values of X increase, then so do the values of Y.

2 If small values of X are associated with large values of Y, or as X increases, the corresponding value of Y tends to decrease, then this is a negative linear relationship between X and Y. Another way to think of this is: as the values of X decrease, then the values of Y increase.

Why study simple linear regression?

- If X and Y are linearly related, you might want to use the value of x to predict the value of Y
- You might want to check if there is a significant linear relationship (to do that, we test if $\beta_1 = 0$... you will understand what this means when you complete this chapter)

Independent variables and dependent variable

For example, you are the owner of an ice cream stand. You wish to investigate the relationship between total weekly revenue during the summer months and the amount of money spent on advertising.

The independent variable, fixed by the experimenter, is usually denoted by x. In the previous example, x= amount of money spent on advertising. Independent variables are also called explanatory variables.

For a fixed value of x, the second variable (Y) is randomly distributed. In our example, Y= total weekly revenue during summer months. This variable is the dependent variable and is usually denoted Y. A dependent variable is also called the response variable.

Distinguish between deterministic relationship model, versus a probabilistic model

A deterministic relationship between two variables x and y is one in which the value of y is completely determined by the value of x. Example, $y = \beta_0 + \beta_1 x$ is a deterministic relationship between x and y. (Tell students in linear regression, we usually use the notation β_0 to refer to the y-intercept, and β_1 as the slope - make sure you know how to sketch straight lines)

On the other hand, a probabilistic model is different: for a fixed value of x, the value of the second variable is randomly distributed. It has a deterministic part, and a random part. The model can be written:

$$\begin{aligned} Y =& \text{(deterministic function of X)} + \text{(random deviation)} \\ =& f(x) + \epsilon \qquad \text{where } \epsilon \text{ is called the random error} \end{aligned}$$

$\Big($ If we have n observations, we then use $Y_i = f(x_i) + \epsilon_i$ where $i = 1, 2, 3, \cdots, n.\Big)$

It is important to know that for the usual simple linear regression model, we have four assumptions about the random errors:

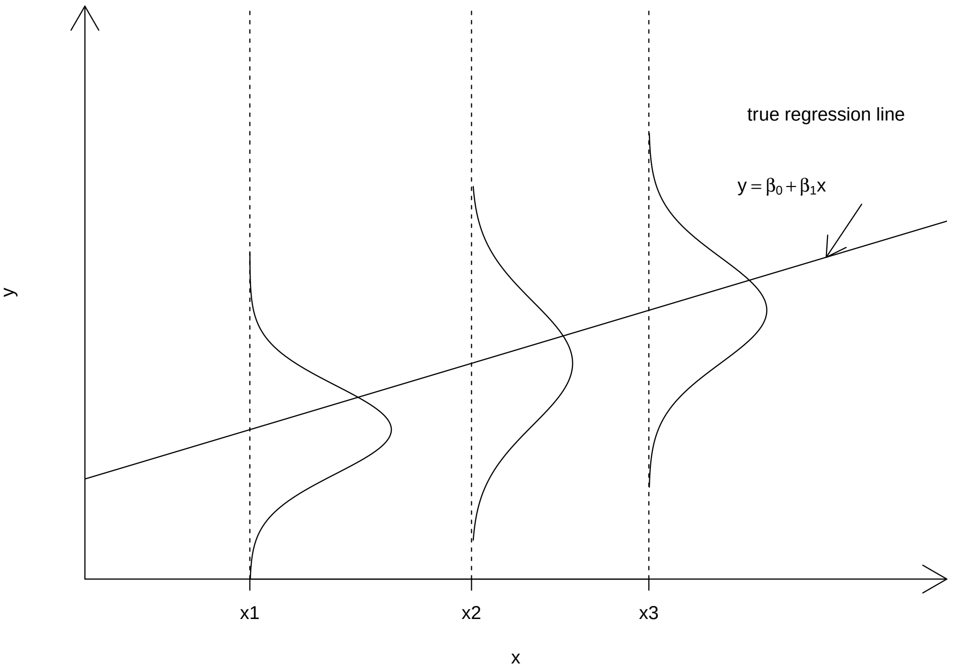
(1) The random variables ϵ_i are independent.

(2) The expected value of ϵ_i is 0 for all i.

(3) $Var(\epsilon_i) = \sigma^2$ for all i.

(4) The ϵ_i's are normally distributed.

(explain to class how these assumptions about ϵ_i actually lead to assumptions about Y). It is best to show these assumptions through a drawing (see next page).



Y_i has a normal distribution centered at $\beta_0 + \beta_1 x_i$. All Y_i distributions have the same variance σ^2

Example:

The owner of a small ice cream stand believes that total weekly revenue (Y, in dollars) during the summer months is related to money spent on advertising (x, dollars per week). A random sample of summer weeks was selected, and the resulting data are given in the following table:

x	30	300	380	275	350	190	85
Y	957	1125	1202	1028	1134	1124	1062

Idea: your aim as an ice-cream vendor is to see if you can predict your total weekly revenue during the summer months if you invest $ 200 per week in advertising.

Thought process: This ice-cream vendor remembers learning simple linear regression in his STAT 251 class and decides to use it to predict total weekly revenue.

First step: Because simple linear regression requires that the relationship between x and Y is linear, we first check by drawing a scatterplot if this linear assumption is reasonable.

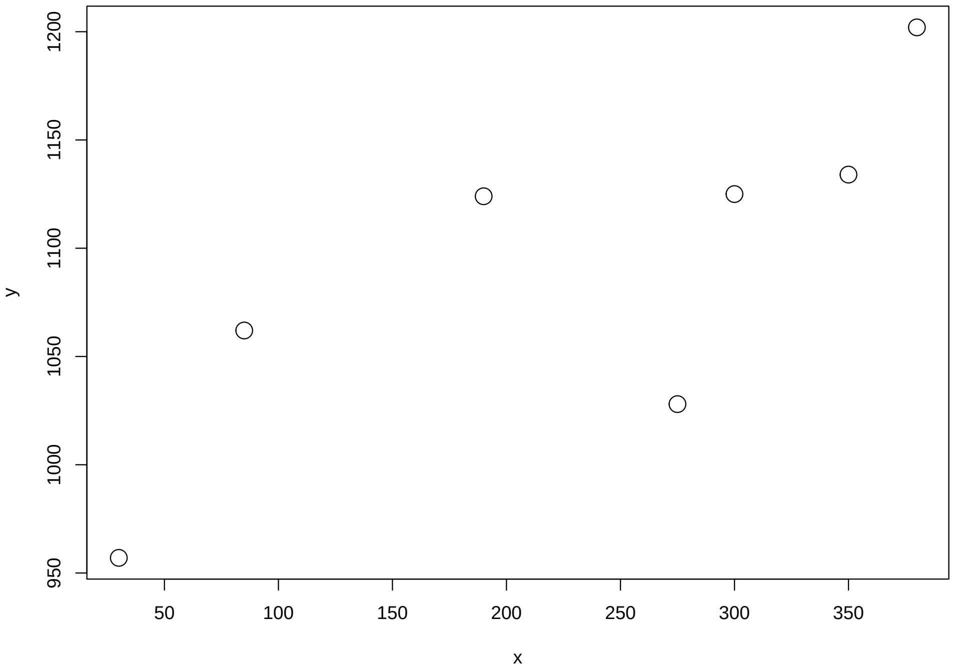
(Using R, I did a scatterplot. Code is:

$> x < -c(30, 300, 380, 275, 350, 190, 85)$

$> y < -c(957, 1125, 1202, 1028, 1134, 1124, 1062)$

$> plot(x, y)$

Scatterplot:

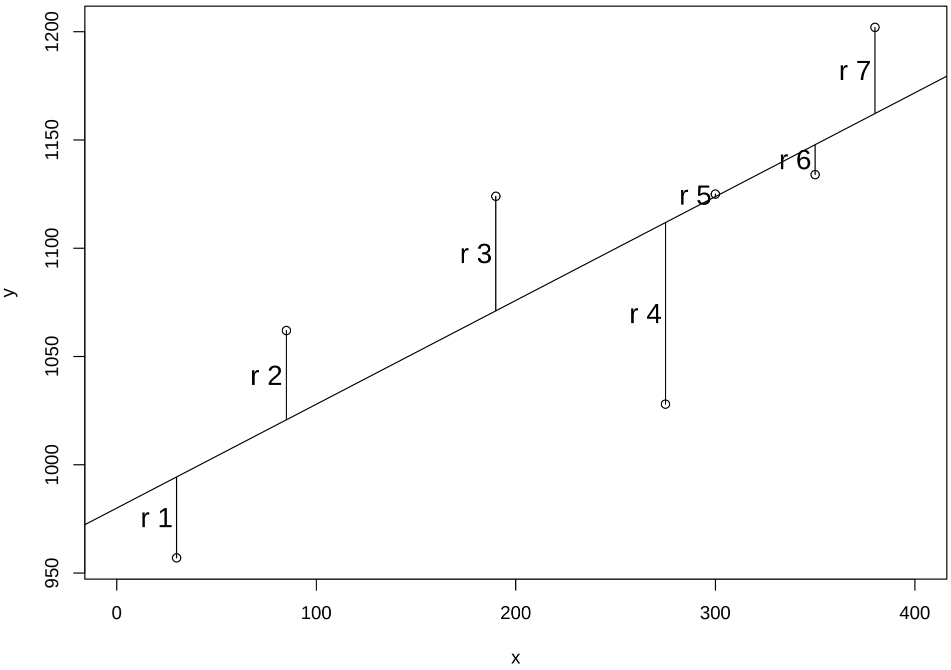


Thought process: envision drawing the 'best' line through these points. I can envision a straight line so using simple linear regression is not unreasonable (large values of x are associated with large values of Y, small values of x are associated with small values of Y).

But what is the 'best' line, and how do we calculate it?

Least squares regression line

The line of best fit is obtained by using the principle of least squares. Principle: minimize the sum of the squared deviations (that is, the vertical distances from the observed points to the line).



Aim: Minimize $\sum_{i=1}^{n} r_i^2$

For students who wish to know, the solution to finding the 'least square' line is proved in the course notes (page 171-172)

Solution:

The least-squares estimates of the regression line is $y = \hat{\beta}_0 + \hat{\beta}_1 x$ where

$$\text{call this } b_1 \Rightarrow \hat{\beta}_1 = \frac{n\sum x_i y_i - (\sum x_1)(\sum y_i)}{n\sum(x_i^2) - (\sum x_i)^2}$$

$$\text{call this } b_0 \Rightarrow \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \overline{y} - \hat{\beta}_1 \overline{x}$$

Using the ice cream vendor example:

X	30	300	380	275	350	190	85
Y	957	1125	1202	1028	1134	1124	1062

Can you work out $\hat{\beta}_1$ and $\hat{\beta}_0$ yourself?

Answer:

$$\hat{\beta}_1 = 0.4795, \hat{\beta}_0 = 980.0067$$

In addition to knowing how to calculate $\hat{\beta}_1$ and $\hat{\beta}_0$ by hand, you need to be able to read the answer from an R output (this may be tested on exam).

The R code to make R perform the calculations for you is given on the next page, along with the output (make sure you know how to use the output).

```
> x <- c(30, 300, 380, 275, 350, 190, 85)
> y <- c(957, 1125, 1202, 1028, 1134, 1124, 1062)
> fit1 <- lm(y~x)
> summary(fit1)
```

we will learn more about this output in the next few lectures
but for now, note the output circled in red

```
Call:
lm(formula = y ~ x)

Residuals:
      1      2      3       4       5      6       7
-37.391 1.151 39.793 -83.862 -13.823 52.893 41.238
```

$b_0 (= \hat{\beta}_0)$

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 980.0067 43.3783 22.592 3.16e-06 ***
x            0.4795 0.1662 2.885 0.0344 *
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 54.23 on 5 degrees of freedom
Multiple R-squared: 0.6247, Adjusted R-squared: 0.5496
F-statistic: 8.321 on 1 and 5 DF, p-value: 0.0344
```

$b_1 \Rightarrow$

Hence the regression line is

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{Y} = 980.0067 + 0.4795x$

To answer the ice-cream vendor's question,
if he invests \$ 200 in advertising (that is $x = 200$),
the <u>expected</u> total weekly revenue is

$\hat{Y} = 980.0067 + 0.4795(200) = \$1075.91$

page 2 of the lecture

recall in earlier notes,
we mention σ^2
(look under 4 assumptions of ϵ_i)
output says $\hat{\sigma} = 54.23$