

Chapter 7: Straight-Line Models

Straight-Lines

Least Squares

Analysis of Residuals

Outliers in Regression Analysis

With this chapter we begin a new phase of statistics—**statistical modeling**

Chap. 7 - page 1

Straight-Line Models (page 177)

$$y = b_0 + b_1x$$

b_0 is the **y-intercept**

(aka the **constant term**)

b_1 is the **slope**

Chap. 7 - page 3

Introduction (page 176)

A **regression model** has two parts:

- A mathematical curve that summarizes the general tendency of the relationship between the variables
- A measure of the amount of variation in the data around that mathematical curve

In regression analysis the variables x and y are treated differently.

y is called the **response variable** and x is called the **predictor variable**.

Chap. 7 - page 2

Example

x	y
4	10
5	8
5	10
7	14
9	18

Chap. 7 - page 4

Fitted Values and Residuals (Page 179)

For a particular x-value, the value of y on the line under consideration is called the **fitted value**.

It is written as \hat{y} .

Fitted value = $\hat{y} = b_0 + b_1x$

For a particular x-value, we next calculate the observed y-value minus the fitted value. This is called the **residual**.

residual = (observed y) – (fitted y)
 $= y - \hat{y} = y - (b_0 + b_1x)$

The smaller the residuals the better the line fits the data.

Typically, some residuals are positive and some are negative and there are many of them.

Example:

		Trial Line $\hat{y} = 2x$	
		Fitted \hat{y}	Residual $y - \hat{y}$
x	y		
4	10	8	+2
5	8	10	-2
5	10	10	0
7	14	14	0
9	18	18	0

The smaller the sum of squared residuals, the better the line fits the data.

Least Squares (page 182)

The **least squares regression line** is the line that makes the sum of squared residuals as small as possible.

Let's compare our trial line

$\hat{y} = 2x$

and a new line with equation

$\hat{y} = 0.75 + 1.875x$

		Trial Line $\hat{y} = 2x$		
		Fitted \hat{y}	Residual $y - \hat{y}$	Squared residual $(y - \hat{y})^2$
x	y			
4	10	8	+2	4
5	8	10	-2	4
5	10	10	0	0
7	14	14	0	0
9	18	18	0	0
		Sum		8

Least Squares Regression Line
 $\hat{y} = 0.75 + 1.875x$

		Fitted \hat{y}	Residual $y - \hat{y}$	Squared residual $(y - \hat{y})^2$
x	y			
4	10	8.250	1.750	3.06250
5	8	10.125	-2.125	4.51563
5	10	10.125	-0.125	0.01563
7	14	13.875	0.125	0.01563
9	18	17.625	0.375	0.14063
		Sum	0	7.75

The least squares regression line

Choose b_0 and b_1 to minimize

$$\sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2$$

The best way to describe the least squares regression line is in terms of standardized variables.

$$\text{Let } y_* = \frac{y - \bar{y}}{s_y}$$

$$\text{and } x_* = \frac{x - \bar{x}}{s_x}$$

Chap. 7 - page 9

In original terms we have (page 184)

$$\hat{y} = \left(\bar{y} - r \frac{s_y}{s_x} \bar{x} \right) + \left(r \frac{s_y}{s_x} \right) x$$

so that the slope is

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

and the y-intercept (the constant term) is

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

Chap. 7 - page 11

Then it may be shown that, in standardized terms, the least squares regression line has the equation

$$\hat{y}_* = r x_* \quad (\text{page 183})$$

where r is the correlation coefficient between x and y .

(math proofs in appendix, pages 204-206)

In standardized terms, the least squares regression line is the line through the origin (0,0) with slope r . (page 183)

Chap. 7 - page 10

Steps for computing the least squares regression line "by hand":

- Find the mean and standard deviation for both x and y
- Find the correlation coefficient between x and y
- Find the slope

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

- Find the y-intercept (the constant term)

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

- The equation of the least squares regression line is $\hat{y} = b_0 + b_1 x$

Chap. 7 - page 12

Example:

$$\bar{x} = 6, \bar{y} = 12, s_x = 2 \text{ and } s_y = 4.$$

Calculation shows that $r = 15/16 = 0.9375$

$$\text{so that the slope is } b_1 = r \frac{s_y}{s_x}$$

$$= (15/16)(4/2) = 15/8 = 1.875$$

The y-intercept is then $b_0 = \bar{y} - b_1\bar{x}$

$$= 12 - (15/8)6 = 12 - (45/4)$$

$$= (48 - 45)/4 = 3/4 = 0.75$$

The equation of the least squares regression line is as reported earlier:

$$\hat{y} = 0.75 + 1.875x$$

Chap. 7 - page 13

Extrapolation (page 186)

It is dangerous (foolish) to use a regression equation to predict by plugging in an x-value that is far from the set of x-values used in obtaining the equation.

We have no data to tell us about the relationship between x and y for such an x.

Chap. 7 - page 15

Using the Least Squares Regression Line

(page 186)

Prediction

A value of x is available for which we need to predict what y will be. To do the prediction just plug x into the least squares regression line:

$$\hat{y} = b_0 + b_1x$$

Example:

We need to predict y for $x = 5.5$.

$$\begin{aligned} \text{Prediction is } \hat{y} \\ &= 0.75 + 1.875(5.5) = 0.75 + 10.3125 \\ &= 11.0625. \end{aligned}$$

Chap. 7 - page 14

Analysis of Residuals (page 188)

Observed Response

$$= \text{Fitted Value} + \text{Residual}$$

$$\text{or } y = \hat{y} + (y - \hat{y})$$

When the model does a good job of describing the general pattern in the observed responses, the residuals will look as if they were generated by a random process. (page 188)

Residuals containing nonrandom patterns indicate that the model can be improved to capture those patterns. (page 188)

Chap. 7 - page 16

Residual Plots

Plot residuals versus x -values to look for patterns.

Example: Average Heights of Young Girls at Various Ages. (Exercise 7.7D, page 194)

Outliers in Regression

Regression analysis is sensitive to outliers.

Example: Battery Life in Portable Computers (pages 195-198)