



BITS Pilani
Pilani Campus

MATH F113


Probability and Statistics

Dr. Shivi Agarwal
Department of Mathematics



Sections 5.3 and Chapter- 11

SIMPLE REGRESSION AND CORRELATION



Let Y be a random variable dependent on a mathematical (i.e. not a random) variable X .

Regression curve is a relationship between $\mu_{Y|x}$ and x ,
Linear curve of regression of Y on X

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

Where β_0 and β_1 denote real numbers.


11.1 MODEL AND PARAMETER ESTIMATION



In simple linear regression model,

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

β_0 denotes the intercept and β_1 the slope of the regression line.



Let x_1, x_2, \dots, x_n be values of X for which observations are made. These points are assumed to be measured without error.

controlled study : x_1, \dots, x_n are preselected by the experimenter,

observational study : they are selected at random.

To study the n random variables $Y|x_1, Y|x_2, \dots, Y|x_n$. Recall that the random variable varies about its mean value. Let

$$E_i = Y | x_i - \mu_Y | x_i$$

Solving that equation , we get


$$Y | x_i = E_i + \mu_{Y | x_i}$$

In this expression, we have assumed that the random difference E_i has a mean zero. Since we are assuming a linear regression, we have

$$Y | x_i = \beta_0 + \beta_1 x_i + E_i$$

It is customary to drop the conditional notation and denote $Y|x_i$ by Y . Hence

$$Y = \beta_0 + \beta_1 x_i + E_i$$




Our data : (x_i, y_i) $i = 1, 2, \dots, n$, where x_i denotes the observed value of the variable X and y_i is the observation for the random variable Y . This above idea is mathematically expressed as

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

In this equation , ε_i denotes the realization of the random variable E_i when Y_i takes the value y_i .

In a regression study, it is useful to plot the xy data. Such a plot is called the *scattergram*. We do not expect the points to lie on a straight line. However if linear regression is applicable, then they should exhibit a linear trend.



Note that since we do not know the true values of β_0 and β_1 we shall not know the true value of ε_i .

Letting b_0 and b_1 denote the estimates of β_0 and β_1 respectively, the estimated line of regression takes the form,

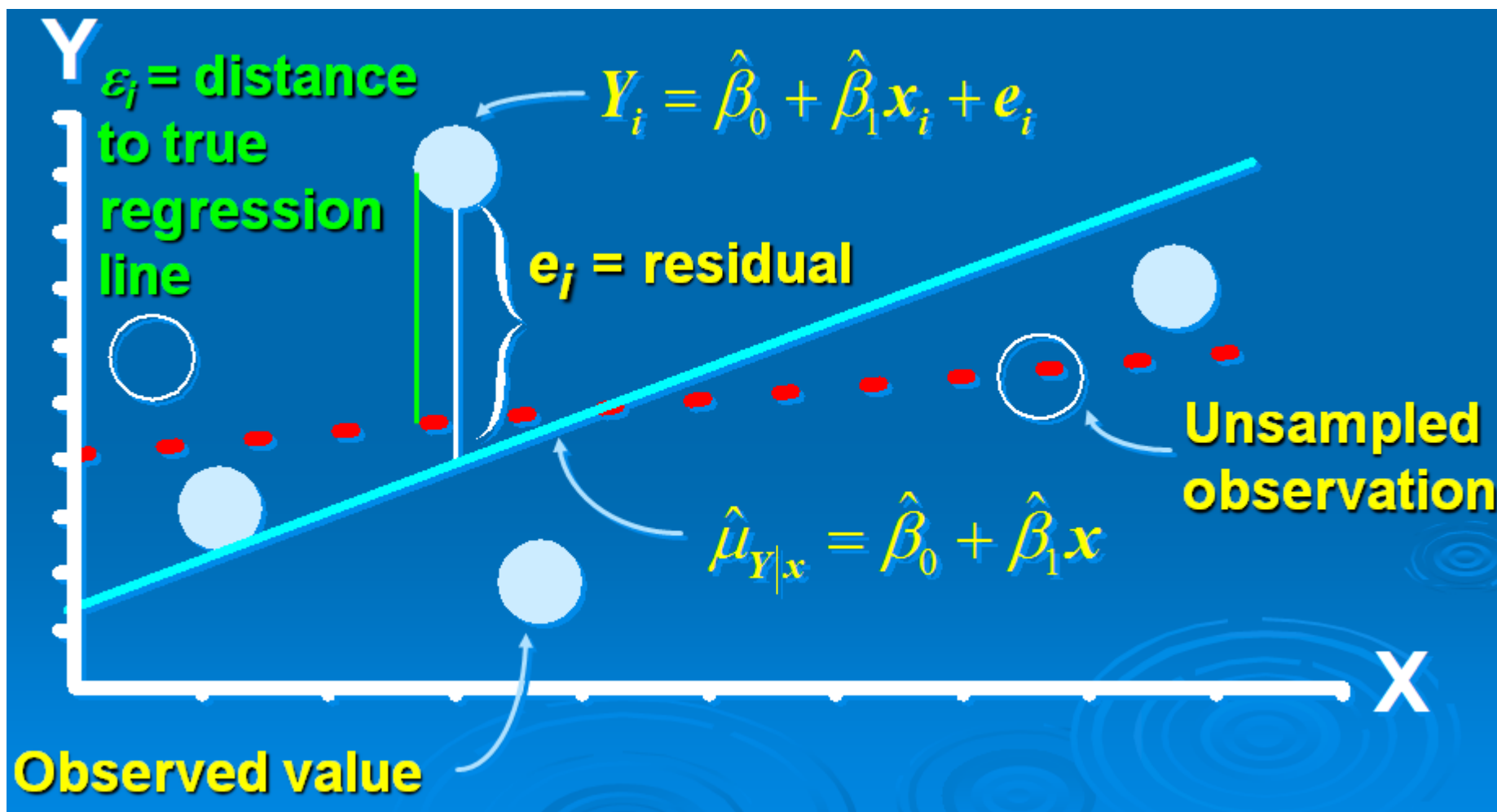
$$\hat{\mu}_{Y|x} = b_0 + b_1 x$$

Just as the data points do not all lie on the theoretical line of regression, they also do not all lie on this estimated regression line. If we let e_i denote the vertical distance from a point (x_i, y_i) to the estimated regression line,

Then each data point satisfies the equation

$$y_i = b_0 + b_1 x_i + e_i$$

This term e_i is termed as residual.

ε_i and e_i 

Least –squares estimation:



The parameters β_0 and β_1 are determined by method of least squares. This works best because we wish to pick the one that that best fits the data. We choose b_0 and b_1 such that we minimize the sum of the squares of the residuals.

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

The sum of the squares of the errors about the estimated regression line is given by

Differentiating this with respect to b_0 and b_1 we obtain,

$$\frac{\partial SSE}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i)$$

$$\frac{\partial SSE}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i$$

We set these partial derivatives as zero and use the rules of summation to obtain the equations

$$nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

These equations are called the **normal equations**. They can be solved easily to obtain these estimates for β_0 and β_1

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

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

Let

Alternate method

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n \sum x_i^2 - \left(\sum x_i \right)^2}{n};$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \frac{n \sum x_i Y_i - \left(\sum x_i \right) \left(\sum Y_i \right)}{n};$$

$$S_{yy} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{n \sum Y_i^2 - \left(\sum Y_i \right)^2}{n}.$$

Then

$$SSE = \sum_{i=1}^n (Y_i - b_0 - b_1 x_i)^2 = \sum \left((Y_i - \bar{Y}) - b_1 (x_i - \bar{x}) \right)^2$$

$$= S_{yy} - 2b_1 S_{xy} + b_1^2 S_{xx}$$

To minimize this, equate to 0 derivative wrt b_1 ,

$$-2S_{xy} + 2b_1S_{xx} = 0 \Rightarrow b_1 = \frac{S_{xy}}{S_{xx}}.$$

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

EX.

In the following table, x is the tensile force applied to a steel specimen in thousand of pounds and y is the resulting elongation in thousands of an inch:

:



x

y

1

14

2

33

3

40

4

63

5

76

6

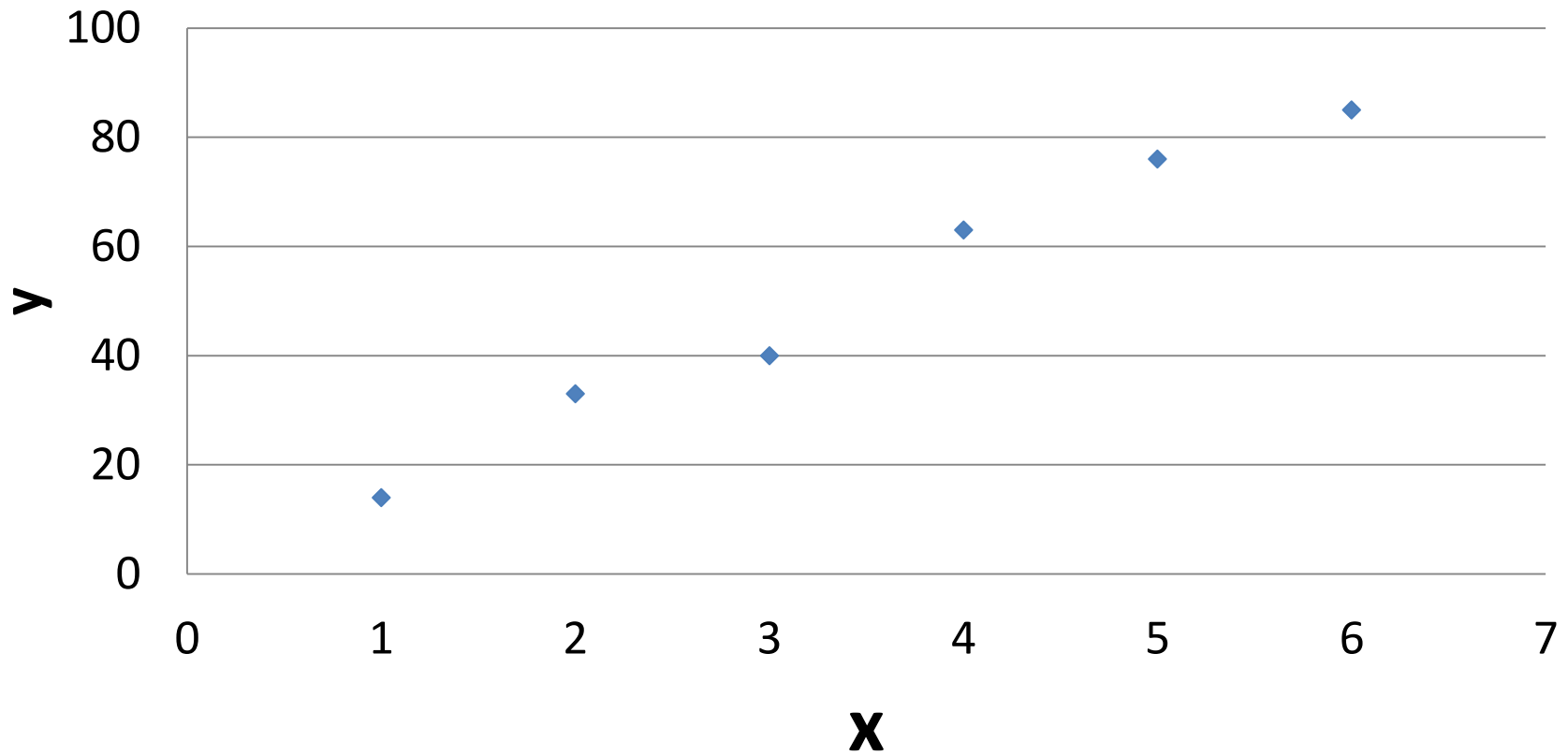
85



(a) Graph the data to verify that it is reasonable to assume that the regression of Y on X is linear

(b) Find the equation of the least square line and use it to predict the elongation when the tensile force is 3.5 thousand pounds.

scattergram



Solution: (a).

(b). Since $n=6$, we need to have:

$$\sum_{i=1}^n x_i = 21, \quad \sum_{i=1}^n x_i^2 = 91, \quad \sum_{i=1}^n y_i = 311,$$

$$\sum_{i=1}^n x_i y_i = 1342, \quad \sum_{i=1}^n y_i^2 = 19855,$$

$$s_{xx} = \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 / n = 91 - (21)^2 / 6 = 17.5$$

$$s_{xy} = \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) / n = 253.5$$

$$b = \frac{s_{xy}}{s_{xx}} = 14.486, \quad a = \bar{y} - b \bar{x} = 1.133$$

Therefore, when the tensile force
is 3.5, we have:

$$y = 1.133 + (14.486)(3.5) = 51.83$$

Exercises

Section 11.1

2. For each of the three following data sets, plot a scatter gram and subjectively state whether it appears that a linear regression will (i) fit the data well (ii) give only a fair fit (iii) fit the data poorly.

x	5	15	25	35	45	50
Y	10	18	20	25	32	45

6. For each of the data sets of problem 2, estimate β_0 and β_1 . Find the residuals in each case and verify that apart from round-off error, the residuals sum to 0.

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

$$= \frac{6(5385) - 175(150)}{6(6625) - (175)^2} = 0.66$$

$$b_0 = 25 - 0.66(29.2) = 5.7$$

$$y = 5.7 + 0.66x$$

true y	10	18	20	25	32	45
estimated y	8.95	15.59	22.23	28.87	35.52	38.84
e	1.05	2.41	-2.23	-3.87	-3.52	6.16

$$\sum e_i = 0$$

7. The relationship between energy consumption and household income was studied, yielding the following data on household income X (in units of \$1000/year) and energy consumption Y (in units of 10^8 Btu/Year)

Energy Consumption (y)	Household Income (x)
1.8	20.0
3.0	30.5
4.8	40.0
5.0	55.1
6.5	60.3
7.0	74.9
9.0	88.4
9.1	95.2

Contd..

- Plot a scatter gram of these data.
- Estimate the linear regression equation $\mu_{Y|x} = \beta_0 + \beta_1 x$.
- If $x = 50$ (household income of \$50,000), estimate the average energy consumed for households of this income. What would your estimate be for a single household?

8. Consider the data in exercise 7.

a) Write the normal equations for these data.

b) Solve the normal equations for b_0 and b_1 . Verify that your results are the same as those you obtained in part (b) of exercise 7.

normal equations are:

$$nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

$$8b_0 + 464.4b_1 = 46.2$$

$$464.4b_0 + 32089.96b_1 = 3173.17$$

11.6 Correlation



In this section, we are trying to measure the linear relation between two random variables X and Y .

The theoretical parameter used to measure the linear relationship between X and Y is the Pearson coefficient ρ . This parameter is defined by

$$\rho = \frac{Cov(X, Y)}{\sqrt{(VarX)(VarY)}}$$

Correlation



Correlation is a normalized covariance. It provides a linear relationship between X and Y . If X and Y show similar behaviour, then the correlation is +ve, else -ve. Let X, Y be RVs with means μ_X, μ_Y , and variance σ_X^2, σ_Y^2 , respectively. Then

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Remarks:

- (i) $-1 \leq \rho_{XY} \leq 1$
- (ii) If two RVs X and Y are independent, then they are uncorrelated ($\rho_{XY} = 0$).
- (iii) However, $\rho_{XY} = 0$ does not imply that X and Y are independent (WHY?).
- (iv) $\rho_{XY} = \rho_{YX}$ and $\rho_{XX} = 1$

The parameter ρ assumes values between **-1 and 1 inclusive**. Values of -1 and 1 indicate perfect positive and negative relationships respectively. A value of 0 indicates **no linear** relationship. When this occurs, we say that X and Y are uncorrelated.


Previously, we found that the theoretical value of ρ based on knowledge of the joint density function for X and Y . But these are seldom known in practice. So our job is to estimate ρ based on set of observations $\{(x_i, y_i): i=1, 2, \dots, n\}$ on the random variable (X, Y) . We must use $\text{Var } X$, $\text{Var } Y$ and $\text{Cov } (X, Y)$ for this purpose.

$$\hat{Var} X = \sum_{i=1}^n (X_i - \bar{X})^2 / n = S_{xx} / n$$

$$\hat{Var} Y = \sum_{i=1}^n (Y_i - \bar{Y})^2 / n = S_{yy} / n$$

To estimate $Cov(X, Y)$, note that

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$



We estimate $\text{Cov}(X, Y)$ by averaging products analogous to that on the right hand side of the previous equation.

$$\widehat{\text{Cov}}(X, Y) = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) / n = S_{xy} / n$$

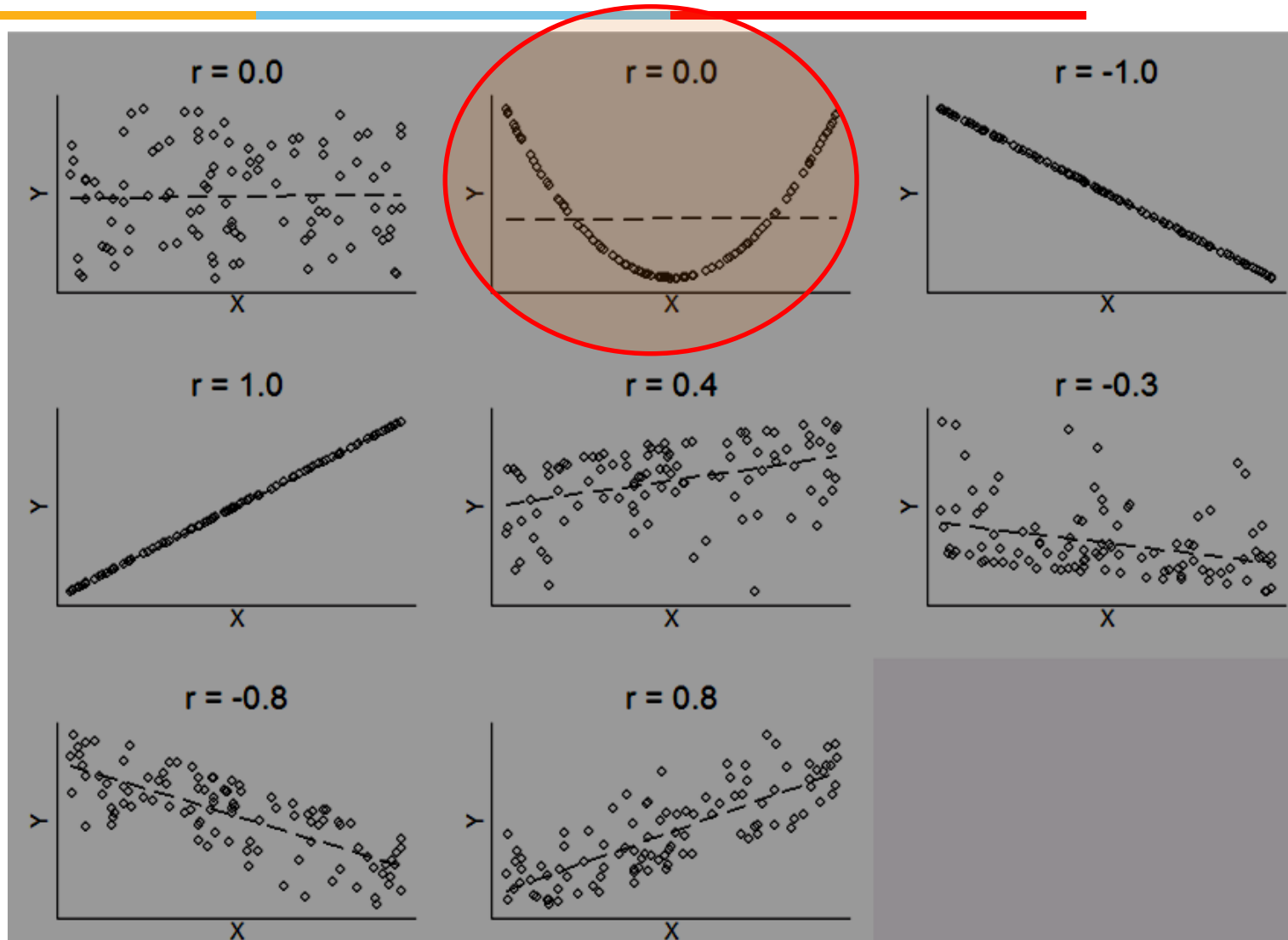
When we combine these estimators, the estimator for ρ is given by

$$\widehat{\rho} = R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

The computational formula for the estimator for ρ is

$$\hat{\rho} = r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

Correlation Coefficient



Section 11.6

Pesticides used in food production can be found in food consumed by humans. A study focusing on chickens exposed to malaoxon was conducted. The chickens were also exposed to a liver enzyme inducer to determine whether liver detoxification of the pesticide is affected. The following data were reported as a percentage of normal pesticide detoxification (y) and percentage of normal liver enzyme levels (x)

Enzyme level (x)	Detoxification level (y)
95	108
110	126
118	102
124	121
145	118
140	155
185	158
190	178
205	159
222	184

47.a) Plot a scattergram of the data.

b) Estimate ρ , the correlation between X and Y.

7. The relationship between energy consumption and household income was studied, yielding the following data on household income X (in units of \$1000/year) and energy consumption Y (in units of 10^8 Btu/Year)

Energy Consumption (y)	Household Income (x)
1.8	20.0
3.0	30.5
4.8	40.0
5.0	55.1
6.5	60.3
7.0	74.9
9.0	88.4
9.1	95.2

Example

An engineer found that by including small amount of compound rechargeable batteries for portable computers, she could extend their lifetimes. She experimented with different amounts of the additive and the data are :

Obtain the least square fit of a straight line to the amount of additive and correlation.

Amount of Additive(x)	Life (y)
5	10
15	18
25	20
35	25
45	32
50	45

Example

- a) Plot a scattergram of the data.
- b) Estimate r , the correlation between X and Y

x	y
4	8
2	12
10	4
5	10
8	2

x	y	x^2	y^2	xy
4	8	16	64	32
2	12	4	144	24
10	4	100	16	40
5	10	25	100	50
8	2	64	4	16
$\Sigma 29$	$\Sigma 36$	$\Sigma 209$	$\Sigma 328$	$\Sigma 162$

$$\begin{aligned}
 \hat{\rho} = r &= \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \\
 &= \frac{5 \times 162 - 29 \times 36}{\sqrt{[5 \times 209 - (29)^2][5 \times 328 - (36)^2]}} \\
 &= -0.883.
 \end{aligned}$$