Chapter 5

Study of Correlation and Regression

Contents

Bivariate Data Set

Definition 5.0.1. *The simultaneous study of two statistical variables on the same population*

Scatter Plot

Definition 5.0.2. *The set of points M_i with coordinates (x_i, y_i).*

There are two cases

5.1 Case 1: Two-row table

x_i	x_1	x_2	...	x_n
y_i	y_1	y_2	...	y_n

Definition 5.1.1. *Marginal Means*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Marginal Variances

$$V(X) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{X}^2$$

$$V(Y) = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \bar{Y}^2$$

Standard Deviations

$$\delta_X = \sqrt{V(X)}$$

$$\delta_Y = \sqrt{V(Y)}$$

Covariance

$$cov(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \bar{X}\bar{Y}$$

5.1.1 Regression Line (Least Squares Method)

Theorem 5.1.1. *The regression line of Y on X, denoted $D_Y(X)$ has the equation $Y = aX + b$ where*

$$a = \frac{cov(X,Y)}{V(X)}$$

$$b = \bar{Y} - a\bar{X}$$

Property 5.1.1. ❶ *The regression line is unique*

❷ *It always passes through the point* $(\bar{X}, \bar{Y})$

5.1.2 Linear Correlation Coefficient

Definition 5.1.2. *The linear correlation coefficient of a bivariate data series is*

$$r = \frac{cov(X, Y)}{\delta_X \delta_Y}$$

Remark 5.1.1. ❶ $-1 \leq r \leq 1$

❷ *If* $r = 0$ *no correlation (X and Y are independent).*

❸ *If* $0 < r < 1$ *weak, moderate, or strong positive correlation between X and Y.*

❹ *If* $-1 < r < 0$ *weak, moderate, or strong negative correlation between X and Y.*

Example 5.1.1. *We have recorded the fuel consumption (in L/100km) for a car model at different speeds (in km/h). The following table was obtained:*

Speed x_i	60	70	90	110	130	150
Fuel consumption y_i	3	3.1	3.7	4.7	6	9

Marginal Means

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = 101.66 \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 4.91$$

Marginal Variances

$$V(X) = \Big(\frac{1}{n}\sum_{i=1}^{n} x_i^2\Big) - \bar{X}^2 = 1015.24 \qquad V(Y) = \Big(\frac{1}{n}\sum_{i=1}^{n} y_i^2\Big) - \bar{Y}^2 = 4.46$$

Standard Deviations

$$\delta_X = \sqrt{V(X)} = 31.86 \qquad \delta_Y = \sqrt{V(Y)} = 2.11$$

Covariance

$$cov(X,Y) = \Big(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\Big) - \bar{X}\bar{Y} = 63.68$$

Regression Line, $Y = aX + b$

$$a = \frac{cov(X,Y)}{V(X)} = 0.0627$$

$$b = \bar{Y} - a\bar{X} = -1.46$$

$$Y = 0.0627X - 1.46$$

Correlation Coefficient

$$r = \frac{cov(X,Y)}{\delta_X \delta_Y} = 0.947$$

so there is a strong positive linear correlation between X and Y.

5.2 Case 3: Contingency Table

XY	y_1	y_2	...	y_c
x_1	n_{11}	n_{12}	...	n_{1c}
x_2				
$\vdots$				
x_l	n_{l1}			n_{lc}

Definition 5.2.1. *Marginal Distributions*

x_i	x_1	x_2	...	x_l
n_i	n_1	n_2	...	n_l

y_j	y_1	y_2	...	y_c
n_j	n_1	n_2	...	n_c

Marginal Means

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{l} n_i x_i \qquad \bar{Y} = \frac{1}{n}\sum_{j=1}^{c} n_j y_j$$

Marginal Variances

$$V(X) = \Big(\frac{1}{n}\sum_{i=1}^{l} n_i x_i^2\Big) - \bar{X}^2 \qquad V(Y) = \Big(\frac{1}{n}\sum_{j=1}^{c} n_j y_j^2\Big) - \bar{Y}^2$$

Standard Deviations

$$\delta_X = \sqrt{V(X)} \qquad \delta_Y = \sqrt{V(Y)}$$

Covariance

$$cov(X, Y) = \Big(\frac{1}{n}\sum_{i=1}^{l}\sum_{j=1}^{c} n_{ij} x_i y_i\Big) - \bar{X}\bar{Y}$$

Example 5.2.1.

XY	1	2	4	n_i
3	2	0	3	5
5	4	6	1	11
6	5	1	7	13
n_j	11	7	11	n=29

Marginal Distributions

x_i	3	5	6
n_i	5	11	13

y_j	1	2	4
n_j	11	7	11

Marginal Means

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{l} n_i x_i = 5,10 \qquad \bar{Y} = \frac{1}{n}\sum_{j=1}^{c} n_j y_j = 2.38$$

Marginal Variances

$$V(X) = \Big(\frac{1}{n}\sum_{i=1}^{l} n_i x_i^2\Big) - \bar{X}^2 = 1.13 \qquad V(Y) = \Big(\frac{1}{n}\sum_{j=1}^{c} n_j y_j^2\Big) - \bar{Y}^2 = 1.75$$

Standard Deviations

$$\delta_X = \sqrt{V(X)} = 1.06 \qquad \delta_Y = \sqrt{V(Y)} = 1.32$$

Covariance

$$cov(X,Y) = \Big(\frac{1}{n}\sum_{i=1}^{l}\sum_{j=1}^{c} n_{ij} x_i y_j\Big) - \bar{X}\bar{Y} = 0$$

Example 5.2.2. *An experiment was conducted on 250 individuals to study the relationship between age X and sleep duration Y. The following table was obtained*

X/Y	[5,7[	[7,9[	[9,11[	[11,15[
[1,3[	0	0	2	36
[3,11[	0	3	12	26
[11,19[	2	8	35	16
[19,31[	0	26	22	10
[31,59[	26	15	6	5

1- Marginal Distributions

X	[1,3[	[3,11[	[11,19[	[19,31[	[31,59[
n_i	38	41	61	58	52
c_i	2	7	15	25	45

Y	[5,7[	[7,9[	[9,11[	[11,15[
n_j	28	52	77	93
c_j	6	8	10	13

2- Marginal Means

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{l} n_i c_i = 20.27 \qquad \bar{Y} = \frac{1}{n}\sum_{j=1}^{c} n_j c_j = 10.25$$

Marginal Variances

$$V(X) = \left(\frac{1}{n}\sum_{i=1}^{l} n_i c_i^2\right) - \bar{X}^2 = 218.87 \qquad V(Y) = \left(\frac{1}{n}\sum_{j=1}^{c} n_j c_j^2\right) - \bar{Y}^2 = 5.95$$

Standard Deviations

$$\delta_X = \sqrt{V(X)} = 14.79 \qquad \delta_Y = \sqrt{V(Y)} = 2.44$$

3- Covariance

$$cov(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{l}\sum_{j=1}^{c} n_{ij}c_i c_j\right) - \bar{X}\bar{Y} = -24.95$$

Correlation Coefficient

$$r = \frac{cov(X,Y)}{\delta_X \delta_Y} = -0.67$$

so there is a strong negative linear correlation between X and Y.

4- Regression Line, $Y = aX + b$

$$a = \frac{cov(X,Y)}{V(X)} = -0.11$$

$$b = \bar{Y} - a\bar{X} = 12.48$$

$$Y = -0.11X + 12.48$$

5- Estimate the sleep duration for a 66-year-old individual

$$Y = -0.11X + 12.48 = -0.11(66) + 12.48 = 5.22h$$