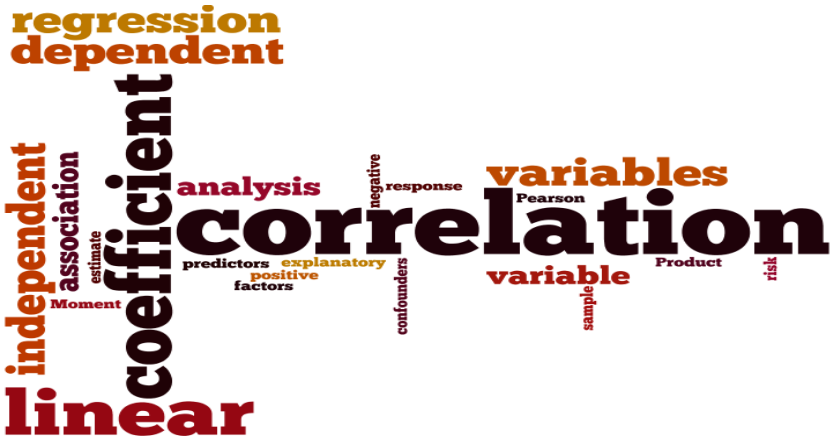
MAT 3103: Computational Statistics and Probability

Chapter 11: Correlation and Regression



Correlation:

A measure of linear relationship between two quantitative variables (for example, age and weight). Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

Types of correlation:

a) Positive correlation, b) Negative correlation, c) No (Zero) correlation

Positive correlation:

If the values of a variable increase, the values of the other variable also increase and as the values of a variable decrease, the values of the other variable also decrease the positive correlation is raised. The points lie close to a straight line, which has a positive gradient.

Example:

Relation between training and performance of employees in a company

Relation between price and supply of a product

Negative correlation:

If the values of a variable increase, the values of the other variable decrease and as the values of a variable decrease, the values of the other variable increase the negative correlation is raised. The points lie close to a straight line, which has a negative gradient.

Example:

Relation between television viewing and exam grades

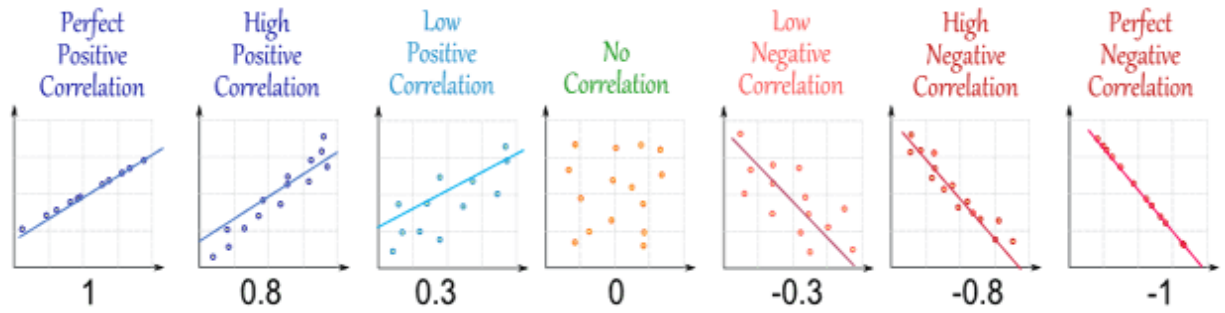
Relation between price and demand of a product

No (Zero) correlation:

If change in one variable has no effect on the other variable. There is no pattern to the points.

Example:

Relation between height and exam grades



Correlation coefficient:

Let (x_1, y_1), (x_2, y_2), …, (x_n, y_n) be n pairs of observations of the variables X and Y observed from n sample points. The linear relationship of X and Y is called simple correlation. The degree of linear relationship of X and Y is estimated by a quantity r, where

$$r = \frac{\mathrm{Cov}(xy)}{\sqrt{\mathrm{V}(x)\mathrm{V}(y)}} = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2 \frac{1}{n}\Sigma(y-\bar{y})^2}} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{[\Sigma x^2 - \frac{(\Sigma x)^2}{n}][\Sigma y^2 - \frac{(\Sigma y)^2}{n}]}} = \frac{\mathrm{SP}(xy)}{\sqrt{\mathrm{SS}(x)\mathrm{SS}(y)}}$$

The values of r range between -1 to 1. r = -1 means a perfect negative correlation, r = 1 means a perfect positive correlation, and r = 0 means no linear relationship between the variables.

Real-life applications of correlation analysis:

A similar negative correlation can be seen with education and the number of children a woman has: **more education**, **fewer children** on average; the **less educated**, the **more children**, on average. Both of these are negative correlations are easy to see on a global map, as well as among populations of a given country.

Correlation of pseudorandom binary codes makes GPS working; lots of radar systems, and lots **CDMA** (code division multiple access) systems. That's why **GPS** is power hungry, it takes a lot of processing to find the correct correlation, with correct satellite (correct code) and the correct delay.

Correlation is one of the promising methods in **analyzing network-based intrusion alerts** to find significant relationships among alerts that have been triggered by multiple intrusion detection sensors. Security Admin (SA) needs to understand and study these alerts. They are meaningless if being analyzed individually. Somehow, they must be 'connected' with previous alerts or future alerts. So, SA can figure out the sequences of attacks that have been launched on the network. This is important to identify preventive measure in the future.

Once correlation is known, we can use it to make **predictions**. If we know a score on one measure, we can make a more accurate prediction of another measure that is highly related to it. The stronger the relationship between/among variables the more accurate the prediction.

Show by example, r = 1	
Let $x = 1, 2, 3$ and $y = 1, 2, 3$. $r = \frac{\mathrm{SP}(xy)}{\sqrt{\mathrm{SS}(x)\,\mathrm{SS}(y)}} = \frac{2}{\sqrt{2 \times 2}} = 1$	$\mathrm{SS}(x) = \sum x^2 - \frac{(\sum x)^2}{n} = 14 - \frac{(6)^2}{3} = 2$ $\mathrm{SS}(y) = \sum y^2 - \frac{(\sum y)^2}{n} = 14 - \frac{(6)^2}{3} = 2$ $\mathrm{SP}(xy) = \sum xy - \frac{\sum x \sum y}{n} = 14 - \frac{6 \times 6}{3} = 2$
r = 1 indicates that *X* and *Y* are perfectly and positively correlated. It happens, if both *X* and *Y* change uniformly in the same direction. Here *X* increases by 1 unit and *Y* increases by 1 unit.	
Show by example, r = −1	
Let $x = 1, 2, 3$ and $y = 9, 6, 3$. $r = \frac{\mathrm{SP}(xy)}{\sqrt{\mathrm{SS}(x)\,\mathrm{SS}(y)}} = \frac{-6}{\sqrt{2 \times 18}} = -1$	$\mathrm{SS}(x) = \sum x^2 - \frac{(\sum x)^2}{n} = 14 - \frac{(6)^2}{3} = 2$ $\mathrm{SS}(y) = \sum y^2 - \frac{(\sum y)^2}{n} = 126 - \frac{(18)^2}{3} = 18$ $\mathrm{SP}(xy) = \sum xy - \frac{\sum x \sum y}{n} = 30 - \frac{6 \times 18}{3} = -6$
r = − 1 indicates that *X* and *Y* are perfectly negatively correlated. It happens when both *X* and *Y* change uniformly but in opposite direction. Here *X* increases by 1 unit and *Y* decreases by 3 units.	

Problem 11.1: The following are the data representing age and BP of some selected persons.

Age $(x, in\ year)$	25	30	35	30	32	40	45	40	36	35
BP $(y, in\ mm\ Hg)$	75	80	85	90	95	85	100	90	85	80

a) Compute correlation coefficient.

b) Do you think that BP increases significantly with the increase in age?

Solution:

a)

$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{xy}$	$\boldsymbol{x^2}$	$\boldsymbol{y^2}$
25	75	1875	625	5625
30	80	2400	900	6400
35	85	2975	1225	7225
30	90	2700	900	8100
32	95	3040	1024	9025
40	85	3400	1600	7225
45	100	4500	2025	10000
40	90	3600	1600	8100
36	85	3060	1296	7225
35	80	2800	1225	6400
$\boldsymbol{\Sigma x} = 348$	$\boldsymbol{\Sigma y} = 865$	$\boldsymbol{\Sigma xy} = 30350$	$\boldsymbol{\Sigma x^2} = 12420$	$\boldsymbol{\Sigma y^2} = 75325$

$$\mathrm{SS}(x) = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$= 12420 - \frac{348^2}{10} = 309.6$$

$$\mathrm{SS}(y) = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$= 75325 - \frac{865^2}{10} = 502.5$$

$$\mathrm{SP}(xy) = \sum xy - \frac{\sum x \sum y}{n}$$

$$= 30350 - \frac{348 \times 865}{10} = 248$$

$$r = \frac{\mathrm{SP}(xy)}{\sqrt{\mathrm{SS}(x)\mathrm{SS}(y)}} = \frac{248}{\sqrt{309.6 \times 502.5}} = 0.63.$$

The variables *X* (age) and *Y* (BP) are positively correlated.

Test of the significance of the correlation coefficient:

We perform a hypothesis test of the **significance of the correlation coefficient** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

Performing the hypothesis test:

$$H_0: \rho = 0 \text{ against } H_A: \rho \neq 0$$

Test Statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Decision rule: With $\alpha = .05$ and df = n – 2, then the critical value of *t* is found from *t* table.

We reject H_0 if $|t| > t_{0.05,(n-2)}$.

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is significant. There is a significant linear relationship between *x* and *y*.

If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is not significant. There is not a significant linear relationship between *x* and *y*.

b) We need to test,

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho \neq 0.$$

Test Statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.63\sqrt{10-2}}{\sqrt{1-0.63^2}} = 2.29$$

Since $|t| < t_{n-2} = t_8 = 2.306$. So H_0 is accepted.

We can conclude that BP of the investigated persons is not significantly correlated with their age.

MATLAB code

To compute the correlation coefficient matrix between two normally distributed, random vectors of 10 observations each.

```
A = randn(10,1);
B = randn(10,1);
R = corrcoef(A,B)
```

Regression:

It is a method of setting a function of dependent variable y based on independent variable x so that for any value of x, value of y can be estimated. Mathematically, the linear regression model is given by,

$$\boldsymbol{Y = \alpha + \beta x + \epsilon},$$

where

α = the value of y when $x = 0$

β = regression coefficient of y on x. It measures the rate of change of y for unit change in x.

ϵ = random error. It is used in the model to measure the influences of other variables which are not included in the model.

The problem is to fit the regression equation in such a way that the sum of squares due to error is minimum. Let the fitted model be

$$\boldsymbol{\hat{y} = a + bx},$$

where, a is the estimate of α and b is the estimate of β. Here,

$$a = \bar{y} - b\bar{x}\text{ , and }b = \frac{\text{SP}(xy)}{\text{SS}(x)}.$$

Show, by example, b = 1	
$Let\ x = 1, 2, 3\ and\ y = 1, 2, 3.$	$\text{SS}(x) = \sum x^2 - \frac{(\sum x)^2}{n} = 14 - \frac{(6)^2}{3} = 2$
$b = \frac{\text{SP}(xy)}{\text{SS}(x)} = \frac{2}{2} = 1$	$\text{SP}(xy) = \sum xy - \frac{\sum x \sum y}{n} = 14 - \frac{6 \times 6}{3} = 2$
Show, by example, b = −2	
$Let\ x = 1, 2, 3\ and\ y = 8, 6, 4.$	$\text{SS}(x) = \sum x^2 - \frac{(\sum x)^2}{n} = 14 - \frac{(6)^2}{3} = 2$
$b = \frac{\text{SP}(xy)}{\text{SS}(x)} = \frac{-4}{2} = -2$	$\text{SP}(xy) = \sum xy - \frac{\sum x \sum y}{n} = 32 - \frac{6 \times 18}{3} = -4$

Real-life applications of regression analysis:

To estimate the impact of CGPA on university admissions.

To estimate the impact of rainfall amount on number fruits yielded.

To predict the sale of products in the future based on past buying behavior, etc.

Problem 11.2: The following are the data representing the number of ever born children (y) to different mothers having different levels of education (x in completed years of schooling):

$$x: \quad 8, 4, 5, 10, 12, 8, 5, 10, 0, 6, 8, 5, 0$$

$$y: \quad 2, 6, 5, 3, 1, 2, 5, 2, 7, 3, 4, 2, 5$$

a) Fit a regression line of y on x.

b) Estimate the number of children of a mother who complete 14 years of schooling.

c) Test the significance of regression.

Solution:

a) $SS(x) = \sum x^2 - \frac{(\sum x)^2}{n} = 663 - \frac{(81)^2}{13}$ $\quad \mathrm{SP}(xy) = \sum xy - \frac{\sum x \sum y}{n} = 228 - \frac{81 \times 47}{13}$

$= 158.31$ $\qquad = -64.85$

$$b = \frac{\mathrm{SP}(xy)}{\mathrm{SS}(x)} = \frac{-64.85}{158.31} = -0.41 \qquad a = \bar{y} - b\bar{x} = \frac{\sum y}{n} - b\frac{\sum x}{n}$$

$$= \frac{47}{13} - (-0.41)\frac{81}{13} = 6.17$$

Fitted line: $\hat{y} = a + bx = 6.17 - 0.41x$

b) If $x = 14$, then $\hat{y} = 6.17 - 0.41(14) = 0.49$

c) We need to test $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$.

Test Statistic: $t = \frac{b}{\sqrt{\frac{s^2}{ss\,(x)}}} = \frac{-0.41}{\sqrt{\frac{1.317}{158.31}}} = -4.5$

$$s^2 = \frac{\mathrm{ss}(y) - b\ \mathrm{sp}(xy)}{n-2} = \frac{41.08 - (-0.41)(-64.85)}{13-2} = 1.317$$

Since $|t| > t_{13-2} = t_{11} = 2.201$, so H_0 is rejected. Hence, the regression is significant.

Exercise 11

11.1 The following data are given for the inflation rate(x) and the corresponding lending rate(y)

x	y			
11.8	10.4			
12.5	16.5			
15.7	22.9			
19.2	26.6			
21.9	33.8			
23.3	42.8			

a. Compute correlation coefficient.

b. Do you think that lending rate increases significantly with the increase of inflation rate?

c. Fit a regression line of y on x.

d. Estimate the lending rate when the inflation rate will be 25.5.

e. Test the significance of regression.

11.2 The following data are given for the educational qualification (year of schooling)(x) of a person and the corresponding yearly income (in lac) (y)

x	y			
5	13.6			
8	15.6			
10	18.7			
12	20.8			
16	25.2			
18	29.5			

a. Compute correlation coefficient.
b. Do you think that yearly income increases significantly with the increase of year of schooling?
c. Fit a regression line of y on x.
d. Estimate the yearly income of an illiterate person.
e. Test the significance of regression.

11.3 The following data are given for the day temperature (in ℃)(x) of Dhaka and the corresponding humidity (in %) (y)

x	y			
30	90			
32	78			
34	84			
36	73			
38	88			
40	72			

a. Compute correlation coefficient.

b. Do you think that humidity increases significantly with the increase of temperature?

c. Fit a regression line of y on x.

d. Estimate the humidity of a day with the temperature 37℃.

e. Test the significance of regression.

11.4 The following data are given for the day temperature (in ℃)(x) of Dhaka in December and the corresponding sales of ice cream (in thousand) (y)

x	y			
22	83.6			
16	61.4			
19	72.0			
21	78.2			
24	87.6			
26	98.2			

a. Compute correlation coefficient.

b. Do you think that sales of ice cream significantly change with the temperature?

c. Fit a regression line of y on x.

d. Estimate the sales of ice cream of a day with the temperature 14℃and 30℃.

e. Test the significance of regression.

Sample MCQs

1. The following data are given for the day temperature (in ℃) (x) of Chittagong and the corresponding humidity (in %) (y). Compute correlation coefficient r.

x	35	32	30	40	38
y	85	75	90	70	85

a) 0.73 b) -0.52 c) -0.73 d) 0.52

2. Test the significance of correlation coefficient (r), where r = 0.85 and sample size is 15.

a) Significant b) Not significant c) Inconclusive d) None of the above

3. The following data are given for the inflation rate(x) and the corresponding lending rate(y). Fit a regression line of y on x.

x	15.5	12.5	11.5	21.5	23.5
y	22.5	17.0	10.5	33.5	42.8

a) $y = -2.37 + 14.79x$
b) $y = 14.79 - 2.37x$
c) $y = -14.79 + 2.37x$
d) $y = 2.37 - 14.79x$

4. Regression coefficient $b = 0.53, \mathrm{SS}(x) = 117, s^2 = 2.5, n = 10,$ comment regarding hypothesis, where $H_0: \beta = 0$.

a) Null hypothesis is accepted
b) Null hypothesis is rejected
c) Both a and b
d) None of them