# Lecture 18

## Correlation analysis and regression

Correlation coefficient r

t- test for significance of correlation

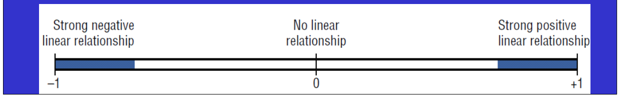
Reading materials: Chapter 10 (or 11 website learning center) of text book

## Pearson product moment correlation coefficient (PPMC)

The **correlation coefficient** computed from the sample data measures the strength and direction of a linear relationship between two variables. The symbol for the sample correlation coefficient is *r*. The symbol for the population correlation coefficient is $\rho$ (Greek letter rho).

$\rho$

$$r = \frac{\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$


Strong negative linear relationship
No linear relationship
Strong positive linear relationship
−1
0
+1

$H_0$: $\rho = 0$ This null hypothesis means that there is no correlation between *x* and *y* variables in the population.

$H_1$: $\rho \neq 0$ This alternative hypothesis means that there is a significant correlation between the variables in the population.

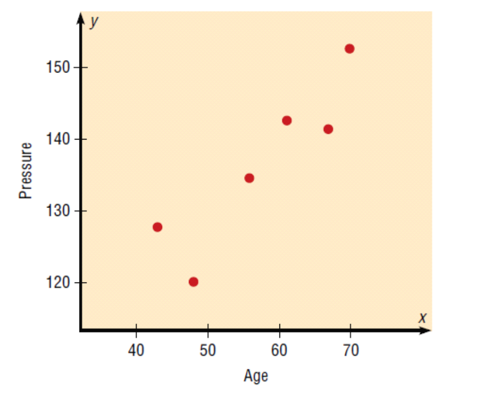
### Formula for the *t* test for the Correlation Coefficient

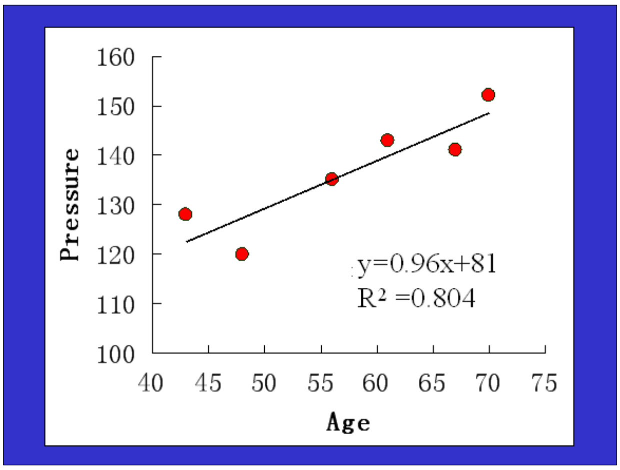
$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

with degrees of freedom equal to $n - 2$.

Construct a scatter plot for the data obtained in a study of age and systolic blood pressure of six randomly selected subjects. The data are shown in the following table.

| Subject | Age, *x* | Pressure, *y* |
|---|---|---|
| A | 43 | 128 |
| B | 48 | 120 |
| C | 56 | 135 |
| D | 61 | 143 |
| E | 67 | 141 |
| F | 70 | 152 |



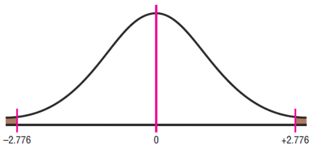


**Solution**

**STEP 1** State the hypotheses.

$$H_0\colon \rho = 0 \qquad \text{and} \qquad H_1\colon \rho \neq 0$$

**STEP 2** Find the critical values. Since $\alpha = 0.05$ and there are $6 - 2 = 4$ degrees of freedom, the critical values obtained from Table F are ±2.776, as shown in Figure 11–6.



**STEP 3** Compute the test value.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = (0.897)\sqrt{\frac{6-2}{1-(0897)^2}} = 4.059$$

**STEP 4** Make the decision. Reject the null hypothesis, since the test value falls in the critical region

**STEP 5** Summarize the results. There is a significant relationship between the variables of age and blood pressure.

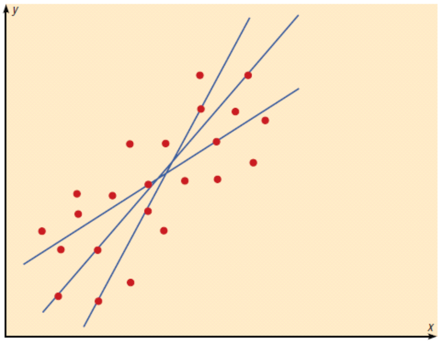
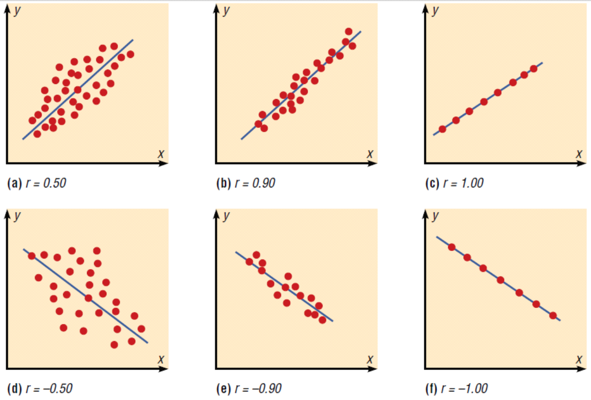


Figure 11–12 shows the relationship between the values of the correlation coefficient and the variability of the scores about the regression line. The closer the points fit the regression line, the higher the absolute value of $r$ is and the closer it will be to $+1$ or $-1$. When all the points fall exactly on the line, $r$ will equal $+1$ or $-1$, and this indicates



(a) $r = 0.50$ (b) $r = 0.90$ (c) $r = 1.00$

(d) $r = -0.50$ (e) $r = -0.90$ (f) $r = -1.00$

Last square fit

LS is a method which finds a linear regression line to fit the data with minimum standard error:

$$SE = \frac{1}{n-1}\sum(\hat{y}_i - y)^2 = \min$$

and the regression model involves two regression coefficient a and b

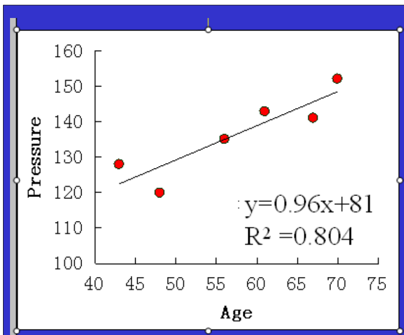
$$\hat{y} = a + bx$$

where the estimates of the two parameters are

$$\bar{y} = a + b\bar{x}$$

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

a is the intercept and b is the slop of the regression line



$$\bar{x} = 57.5$$
$$\bar{y} = 136.5$$
$$s_x = 10.6$$
$$s_y = 11.4$$
$$n = 6$$

$$b = 0.897\frac{11.4}{10.6} = 0.96$$

$$a = 136.5 - 0.96 * 57.5 = 81$$