Nonparametric Statistics

Ch.5 Correlation analysis

Motivation

- Often, we are interested in assessing relationship between two random variables.

- (Pearson) sample correlation coefficient is a good candidate to assess the relationship and test its significance. However, this method is based on the normality assumption, which means that it fails when the true distribution of population is not normal.

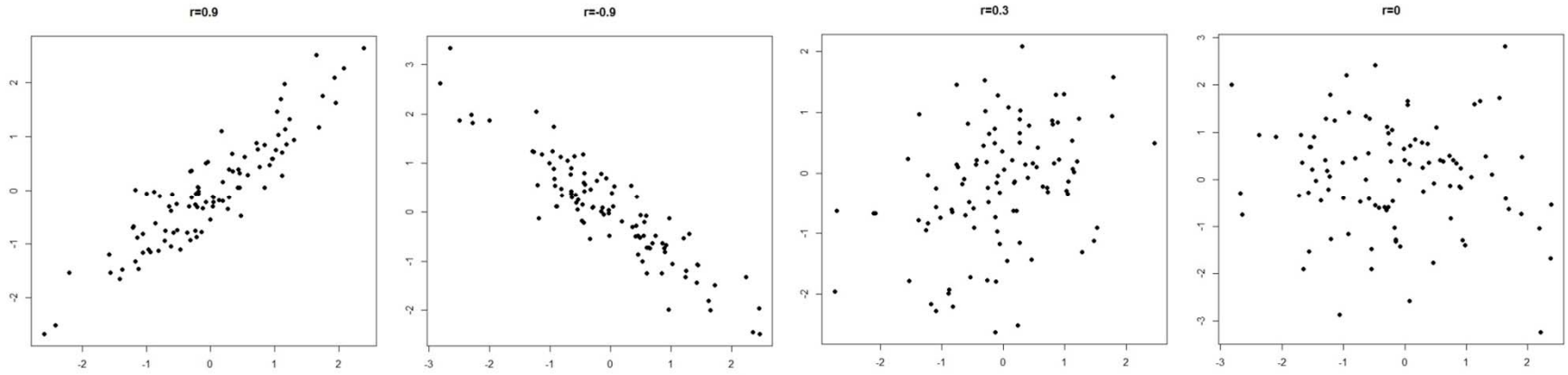
- Some nonparametric rank-based measures of association are introduced as alternatives to (Pearson) sample correlation coefficient which is a parametric one.

Review : Pearson correlation coefficient

- Pearson correlation coefficient measures how strong the linear relationship between two variables is.
- (Pearson) correlation coefficient

$$\rho = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

- $|\rho| \leq 1, \rho > 0$: positive correlation, $\rho < 0$: negative correlation
- If two variables are independent, then $\rho = 0$. But, the reverse is not true.
- Note that this only take "linear" relationship into account.



Review : Pearson correlation coefficient

- Pearson sample correlation coefficient : (X_i, Y_i) , $i = 1, \dots, n$.

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

- $|r| \leq 1$
- A test for the hypothesis

$$H_0: \rho = 0 \quad vs \quad H_1: \rho \neq 0$$

can be performed with the test statistic

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

- This follows $t(n-2)$ under H_0 when the true distribution is a bivariate normal distribution.

Spearman's rank correlation

Spearman's rank correlation coefficient : (X_i, Y_i) , $i = 1, \dots, n$.

$$r_s = \frac{\sum_{i=1}^{n}(r_{X,i} - \bar{r_X})(r_{Y,i} - \bar{r_Y})}{\sqrt{\sum_{i=1}^{n}(r_{X,i} - \bar{r_X})^2}\sqrt{\sum_{i=1}^{n}\left(r_{Y,i} - \bar{r_Y}\right)^2}}$$

where $r_{X,i}$ and $r_{Y,i}$ are the ranks of X_i's and Y_i's in each group.

- We assume that the true distribution is a continuous bivariate distribution.
- This is a measure of how well the relationship between two random variables can be described by a monotonic function.
- $|r_s| \leq 1$
- $r_s \approx 1$: strong positive monotonic relationship
- $r_s \approx -1$: strong negative monotonic relationship
- $r_s \approx 0$: no monotonic relationship

Spearman's rank correlation (large sample)

- A test for the hypothesis

$$H_0: there\ is\ a\ monotonic\ relationship \quad vs \quad H_1: not\ H_0$$

can be performed with the test statistic

$$Z_s = \frac{r_s}{\sqrt{1/(n-1)}} \approx N(0,1)$$

under H_0 for sufficiently large n.

Spearman's rank correlation

Note that r_s can be simplified as

$$r_s = 1 - \frac{6}{n(n-1)(n+1)} \sum_{i=1}^{n} (i - R_i)^2$$

where R_i is the corresponding ranks of $Y_1, \dots, Y_n$ when we suppose that $X_1 < X_2 < \cdots < X_n$ without loss of generality. That is, the ranks of n pairs are $(1, R_1), \dots, (n, R_n)$. Note that $i - R_i$ is the difference of ranks for i^{th} pair.

$(\because)$

$$r_s = \frac{\sum_{i=1}^{n} (i - \frac{n+1}{2})(R_i - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n} (i - \frac{n+1}{2})^2} \sqrt{\sum_{i=1}^{n} \left(R_i - \frac{n+1}{2}\right)^2}}$$

Kendall τ rank correlation

Definition] Concordance & discordance

We say that the pairs (X_i, Y_i) and (X_j, Y_j) are concordant if $X_i > X_j$ and $Y_i > Y_j$ or $X_i < X_j$ and $Y_i < Y_j$, that is,

$$(X_i - X_j)(Y_i - Y_j) > 0.$$

The pairs are discordant if

$$(X_i - X_j)(Y_i - Y_j) < 0.$$

Let C and D denote the numbers of concordant and discordant pairs, respectively. Then, Kendall τ rank correlation is given as

$$\tau = \frac{C - D}{C + D} = \frac{C - D}{n(n-1)/2} = 1 - \frac{4}{n(n-1)} D$$

- Note that $n(n-1)/2$ is the total number of pair combinations.

Kendall τ rank correlation

- We assume that the true distribution is a continuous bivariate distribution.
- This is a a measure of the portion of ranks that match between two data sets.
- $|\tau| \leq 1$
- Note that τ is a natural estimator for

$$2P\left[\left(X_i - X_j\right)\left(Y_i - Y_j\right) > 0\right] - 1.$$

  Here, $P\left[\left(X_i - X_j\right)\left(Y_i - Y_j\right) > 0\right]$ can be estimated by $\frac{C}{C+D}$.

- $\tau \approx 1$: strong positive monotonic relationship
- $\tau \approx -1$: strong negative monotonic relationship
- $\tau \approx 0$: no monotonic relationship

Kendall τ rank correlation (large sample)

- A test for the hypothesis

$$H_0: P\left[(X_i - X_j)(Y_i - Y_j) > 0\right] = \frac{1}{2} \qquad vs \qquad H_1: not\ H_0$$

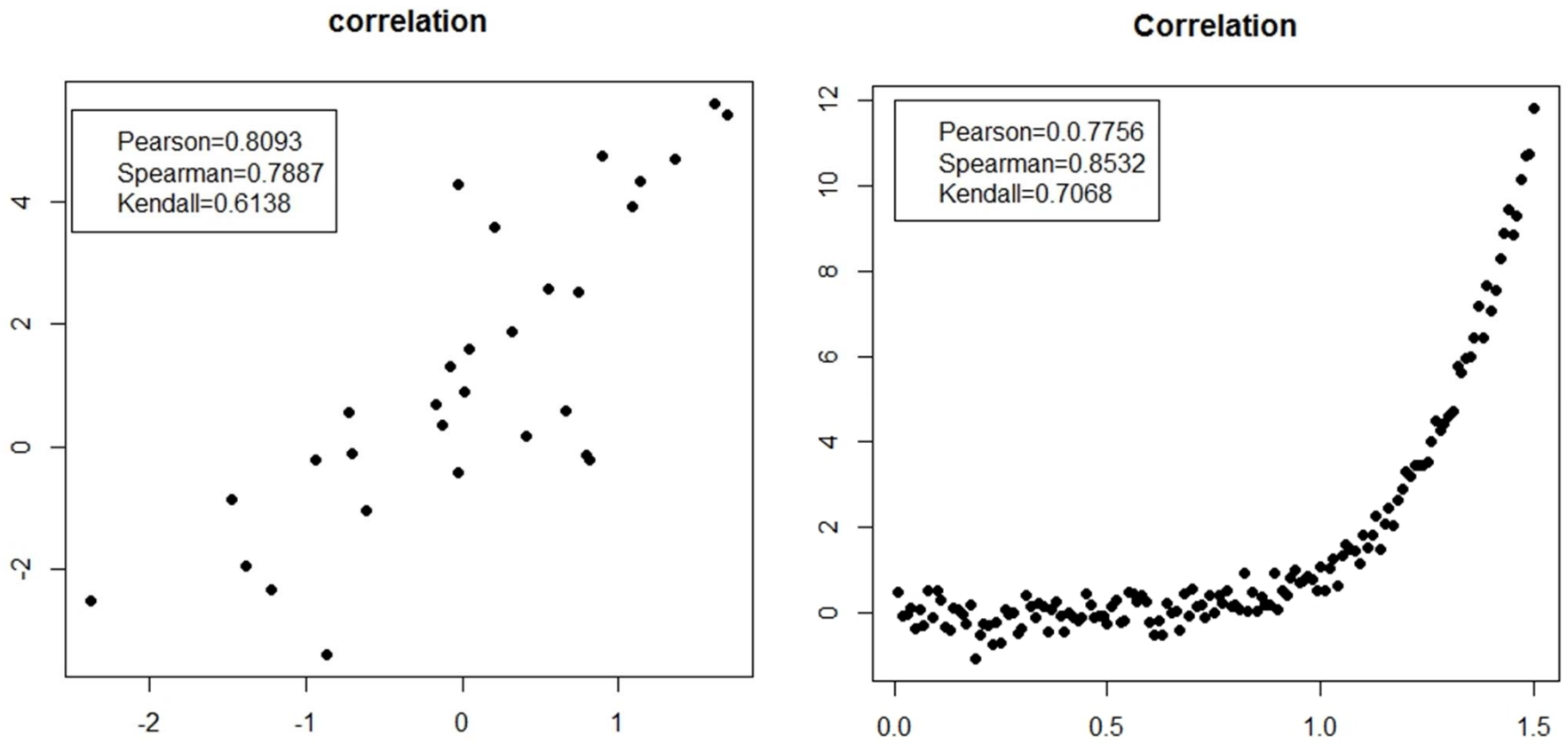
can be performed with the test statistic

$$Z_\tau = \frac{\tau}{\sqrt{(4n+10)/9n(n-1)}} = \frac{C-D}{\sqrt{n(n-1)(2n+5)/18}} \approx N(0,1)$$

under H_0 for sufficiently large n.

Example

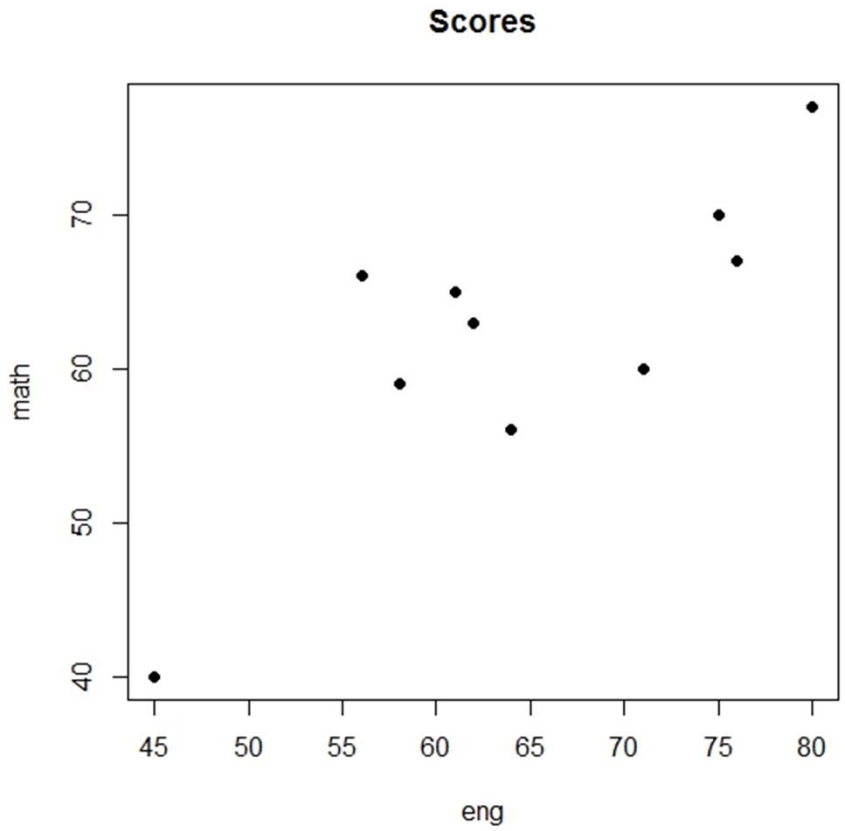
Ex1]


correlation
Pearson=0.8093
Spearman=0.7887
Kendall=0.6138
-2
0
2
4
-2
-1
0
1
Correlation
Pearson=0.0.7756
Spearman=0.8532
Kendall=0.7068
0
2
4
6
8
10
12
0.0
0.5
1.0
1.5

Example

Ex2] Calculate three correlation coefficients and test if they are significant.

English (X)	56	75	45	71	61	64	58	80	76	62
Math (Y)	66	70	40	60	65	56	59	77	67	63



Example

1) Pearson

From the table, $\sum x_i = 648, \sum y_i = 623, \sum x_i^2 = 43028, \sum y_i^2 = 39685, \sum x_i y_i = 41135.$
Therefore,

$$r = \frac{41135 - \frac{648 * 623}{10}}{\sqrt{(43028 - \frac{648^2}{10})(39685 - \frac{623^2}{10})}} = 0.8038.$$

The observed test statistic is

$$\sqrt{10-2}\frac{0.8038}{\sqrt{1-0.8038^2}} = 3.82.$$

P-value is 0.0051. We can conclude that there exists a significant correlation.

Example

2) Spearman

English (X)	56	75	45	71	61	64	58	80	76	62
Rank of X	2	8	1	7	4	6	3	10	9	5
Math (Y)	66	70	40	60	65	56	59	77	67	63
Rank of Y	7	9	1	4	6	2	3	10	8	5
Rank diff.	-5	-1	0	3	-2	4	0	0	1	0

From the table, $\sum(i - R_i)^2 = 56$. Therefore,

$$r_s = 1 - \frac{6}{10 * (10-1) * (10+1)} * 56 = 0.6606.$$

The observed test statistic is

$$\frac{0.6606}{\sqrt{1/(10-1)}} = 1.98.$$

P-value is 0.0475. We can conclude that there exists a significant correlation.

Example

3) Kendall

English (X)	56	75	45	71	61	64	58	80	76	62
Rank of X	2	8	1	7	4	6	3	10	9	5
Math (Y)	66	70	40	60	65	56	59	77	67	63
Rank of Y	7	9	1	4	6	2	3	10	8	5

From the table, $C = 33, D = 12$. Therefore,

$$\tau = \frac{33 - 12}{33 + 12} = 0.4667.$$

The observed test statistic is

$$\frac{0.4667}{\sqrt{(4 * 10 + 10)/(9 * 10 * (10 - 1))}} = 1.88.$$

P-value is 0.0603. No significant evidence.