Correlation Theory

Still bi-variate statistics

X ~ random variable

Y ~ random variable

Covered Topics

- Pearson's Correlation
- Spearman's Rank Correlatiion

Population Covariance (1)

Definition

$$\sigma_{XY} = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

$$= \iint (x-\mu_X)(y-\mu_Y) f(x,y)\,dxdy$$

a constant

Population Covariance (2)

Sign of Covariance

Positive ==> if one RV is above or below its mean, the other RV tends to be also above or below its mean

Negative ==> if one RV is above or below its mean, the other RV tends to be below or above its mean

Population Covariance (3)

Magnitude of Covarinace

unbounded

depends on the units of both RV's

Unit of covariance

= unit of X times unit of Y

e.g., X is in Baht and Y is in Kilogram

σ_{XY} is in Baht-Kilogram

Population Correlation (1)

- Definition

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- Sign of Correlation

–same as that of Covariance

Population Correlation (2)

Magnitude of Correlation

always bounded between -1 and 1

$$-1 \leq \rho_{XY} \leq +1$$

Unit of Correlation

no unit

comparable between populations

Population Correlation (3)

Interpretation of Correlation

ρ_{XY} =+1 ==> If a variable is above or below its mean, the other will be above or below its own mean with certainty

ρ_{XY} =-1 ==> If a variable is above or below its mean, the other will be below or above its own mean with certainty

ρ_{XY} = 0 ==> If a variable is deviated from its mean, the other will be expected at its mean

Sample Covariance

s_{XY} is an estimator for σ_{XY}

Required paired sample

<u>Estimator</u>

$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Paired Sample of Size *n*

i	X_i	Y_i
1	X_1	Y_1
2	X_2	Y_2
:	:	:
:	:	:
n	X_n	Y_n

Sample Correlation (1)

r_{XY} is an estimator of ρ_{XY}

Definition $$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Sign of sample Correlation

same as that of sample Covariance

Sample Correlation (2)

Magnitude of Sample Correlation

same as population correlation

always bounded between -1 and 1

$$-1 \leq r_{XY} \leq +1$$

Unit of sample Correlation

no unit

comparable between data sets

Test for Zero Correlation

$$\mathrm{H}_0 : \rho_{XY} = 0$$

$$\mathrm{H}_1 : \rho_{XY} \neq 0$$

Theorem

$$t_{cal} = \frac{r_{XY}}{\sqrt{\frac{1-r_{XY}^2}{n-2}}} \sim t(n-2)$$

Perform a Two-sided test.

Test for Non-zero Correlation (1)

$$\mathrm{H}_0 : \rho_{XY} = a, \quad a \neq 0$$

$$\mathrm{H}_1 : \rho_{XY} \neq a$$

Theorem

$$\omega = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right),$$

$$\mu_\omega = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

Test for Non-zero Correlation (2)

$$\omega \sim N\left(\mu_\omega, \frac{1}{n-3}\right)$$

$$z_{cal} = \frac{\omega - \mu_\omega}{\sqrt{\frac{1}{n-3}}} \sim N(0,1)$$

Perform a Two-sided test.

Rank Correlation(1)

Two judges (A and B) are to rank n different objects (contestants)

Question: Are the two judges correlated?

How can similarity or dissimilarity be measured?

Rank Correlation(2)

Spearman's Rank Correlation (sample)

$$r' = 1 - \frac{6\sum_i D_i^2}{n(n^2-1)}$$

No definition for population rank correlation

Rank Correlation(3)

R_{ij} = rank given to object i by judge j

D_i = rank difference for object i

$= R_{iA} - R_{iB}$

Rank Correlation(4)

Paired Sample of Size *n*

i	RA_i	RB_i
1	RA_1	RB_1
2	RA_2	RB_2
:	:	:
:	:	:
n	RA_n	RB_n

Rank Correlation(5)

<u>Magnitude of Correlation</u>

always bounded between -1 and 1

$$-1 \le r'_{XY} \le +1$$

<u>Unit of Correlation</u>

no unit

comparable between populations

Rank Correlation(6)

Interpretation of Sample Rank Correlation

- r'_{XY} =+1 ==> If both judges totally agree on the rankings of all the n objects
- r'_{XY} =-1 ==> If both judges totally disagree on the rankings of all the n objects
- r'_{XY} = 0 ==> If the two judges are uncorrelated

Test for Zero Rank Correlation

$$H_0 : \rho'_{XY} = 0$$

$$H_1 : \rho'_{XY} \neq 0$$

Theorem

$$z_{cal} = \frac{r'_{XY}}{\sqrt{\frac{1}{n-1}}} \sim N(0,1)$$

Perform a Two-sided Z-test.

Test for Non-zero Rank Correlation

No such a thing??

Correlation Theory

Now tri-variate

X ~ random variable

Y ~ random variable

Z ~ random variable

Covered Topics

- Partial Correlation (Pearson)

Partial Correlation (1)

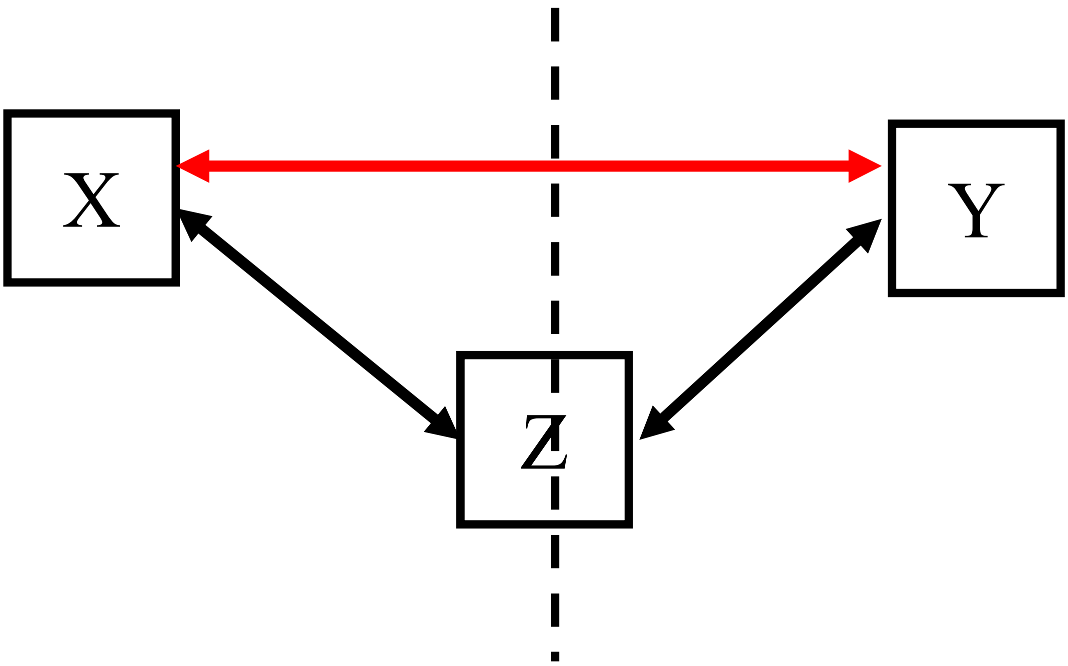
- X, Y and Z are three RV's
- They are assumed to be related.
- Ignoring Z, Corr (X,Y) = "direct" correlation (X,Y) + indirect effects from Z

Partial Correlation (2)

Partial Correlation

- Correlation when embedded effect of Z has been removed from both X and Y
- Parital Corr (X,Y) = “direct” corr(X,Y)

Partial Correlation (3)



Partial Correlation (4)

Assumptions

- X.Z are related as $X = \alpha_1 + \alpha_2 Z + \varepsilon$
- Y.Z are related as $Y = \beta_1 + \beta_2 Z + \xi$

Partial Correlation (5)

PCorr

- X.Z are related as $X = \alpha_1 + \alpha_2 Z + \varepsilon$
- Y.Z are related as $Y = \beta_1 + \beta_2 Z + \xi$

Partial Correlation (6)

By OLS

$$\hat{\alpha}_2 = \frac{\sum x_i z_i}{\sum z_i^2}, \hat{\alpha}_1 = \bar{X} - \hat{\alpha}_2 \bar{Z}$$

where $\quad x_i = X_i - \bar{X},$

$$z_i = Z_i - \bar{Z}$$

Partial Correlation (7)

By OLS

$$\hat{\beta}_2 = \frac{\sum y_i z_i}{\sum z_i^2}, \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{Z}$$

where $\quad y_i = y_i - \bar{Y},$

$$z_i = Z_i - \bar{Z}$$

Partial Correlation (8)

X without Z $$X - \alpha_2 Z = \alpha_1 + \varepsilon$$

Y without Z $$Y - \beta_2 Z = \beta_1 + \xi$$

Partial Corr between X and Y $(\rho_{XY.Z})$

$$= corr(\alpha_2 + \varepsilon, \beta_1 + \xi) = corr(\varepsilon, \xi)$$

Partial Correlation (9)

Sample partial correlation coefficient

$$r_{XY.Z} = \frac{\sum \left(X_i - \hat{X}_i\right)\left(Y_i - \hat{Y}_i\right)}{\sqrt{\sum \left(X_i - \hat{X}_i\right)^2}\sqrt{\sum \left(Y_i - \hat{Y}_i\right)^2}}$$

Partial Correlation (9)

Sample partial correlation coefficient

$$r_{XY.Z} = \frac{\sum\left(x_i - \hat{\alpha}_2 z_i\right)\left(y_i - \hat{\beta}_2 z_i\right)}{\sqrt{\sum\left(x_i - \hat{\alpha}_2 z_i\right)^2}\sqrt{\sum\left(y_i - \hat{\beta}_2 z_i\right)^2}}$$

Partial Correlation (9)

Sample partial correlation coefficient

$$\sum\left(x_i - \hat{\alpha}_2 z_i\right)^2 = \sum x_i^2 - 2\hat{\alpha}_2 \sum x_i z_i + \hat{\alpha}_2^2 \sum z_i^2$$

$$= \sum x_i^2 - 2\frac{\sum x_i z_i}{\sum z_i^2}\sum x_i z_i + \left(\frac{\sum x_i z_i}{\sum z_i^2}\right)^2 \sum z_i^2$$

$$= \sum x_i^2 - \frac{\left(\sum x_i z_i\right)^2}{\sum z_i^2}$$

Partial Correlation (10)

Sample partial correlation coefficient

$$= \sum x_i^2 - \frac{\left(\sum x_i z_i\right)^2}{\sum x_i^2 \sum z_i^2} \sum x_i^2$$

$$= \left(1 - r_{XZ}^2\right) \sum x_i^2$$

Partial Correlation (11)

Sample partial correlation coefficient

$$\sum \left(y_i - \hat{\alpha}_2 z_i\right)^2 = \left(1 - r_{YZ}^2\right) \sum y_i^2$$

Partial Correlation (12)

Sample partial correlation coefficient

$$\sum \left(x_i - \hat{\alpha}_2 z_i\right)\left(y_i - \hat{\beta}_2 z_i\right)$$

$$= \sum x_i y_i - \hat{\alpha}_2 \sum y_i z_i$$

$$- \hat{\beta}_2 \sum x_i z_i + \hat{\alpha}_2 \hat{\beta}_2 \sum z_i^2$$

Partial Correlation (13)

Sample partial correlation coefficient

$$= \sum x_i y_i - \frac{\sum x_i z_i}{\sum z_i^2} \sum y_i z_i$$

$$- \frac{\sum y_i z_i}{\sum z_i^2} \sum x_i z_i + \frac{\sum x_i z_i}{\sum z_i^2} \frac{\sum y_i z_i}{\sum z_i^2} \sum z_i^2$$

$$= \sum x_i y_i - \frac{\sum y_i z_i}{\sum z_i^2} \sum x_i z_i$$

Partial Correlation (14)

Sample partial correlation coefficient

$$= r_{YX}\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}$$

$$- \frac{r_{YZ}\sqrt{\sum y_i^2}\sqrt{\sum z_i^2}}{\sum z_i^2} r_{XZ}\sqrt{\sum x_i^2}\sqrt{\sum z_i^2}$$

$$= \left(r_{YX} - r_{YZ}r_{XZ}\right)\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}$$

Partial Correlation (15)

Sample partial correlation coefficient

$$r_{XY.Z} = \frac{\left(r_{YX} - r_{YZ}r_{XZ}\right)\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}{\sqrt{\left(1 - r_{XZ}^2\right)\sum x_i^2}\sqrt{\left(1 - r_{YZ}^2\right)\sum y_i^2}}$$

$$= \frac{r_{YX} - r_{YZ}r_{XZ}}{\sqrt{1 - r_{XZ}^2}\sqrt{1 - r_{YZ}^2}}$$