Linear Association

Sean Hellingman

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2023

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- We are continuing to make statistical inferences about target populations.

- We are interested in the linear association between two or more variables.
    - Covariance
    - Correlation
    - Linear Regression

Covariance

- The **covariance** is a measure of *linear association* between two random variables.
  - Measures how much and to what extent two variables change together.
  - Measures the direction of the linear association between two random variables.

- Covariance between two (population) random variables X and Y:

$$\sigma_{XY} = Cov(X, Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)]. \tag{1}$$

  - If $\sigma_{XY} > 0$ then X and Y have a positive *linear* association.
  - If $\sigma_{XY} < 0$ then X and Y have a negative *linear* association.
  - If $\sigma_{XY} = 0$ then X and Y have no *linear* association.

Correlation

- The **correlation** is a measure of the direction and strength of the *linear association* between two random variables.
  - The correlation coefficient takes on values between -1 and 1.

- Correlation between two (population) random variables X and Y:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}. \tag{2}$$

- Where σ_X and σ_Y are the standard deviations of X and Y.
  - If $\rho = -1$ then there is a perfect (negative) linear relationship between X and Y.
  - If $\rho = 1$ then there is a perfect (positive) linear relationship between X and Y.
  - If ρ is close to 0 then there is a weak linear relationship between X and Y.
- The strength of the linear association increases as ρ moves away from 0.

Example 1

- Determine the nature of the linear association between the two variables in each of the 5 example plots in the *Linear Associations Examples* R Markdown file.

Sample Covariance

- Often we do not have all of the observations in populations and we need to estimate parameters.

- **Sample covariance** estimates the actual covariance between two random variables:

$$S_{XY} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}). \tag{3}$$

- Where $\bar{x}$ and $\bar{y}$ are the sample means.

Sample Covariance in R

- R function: `cov(x, y, method = "pearson")`
    - `x`: a numeric vector, matrix, or data frame.
    - If `x` is a vector we must give a vector `y`.
    - `method = "pearson"` is the default method.

Covariance Matrix

Variable	X	Y	Z
X	$V[X]$	$Cov(X, Y)$	$Cov(X, Z)$
Y	$Cov(X, Y)$	$V[Y]$	$Cov(Y, Z)$
Z	$Cov(X, Z)$	$Cov(Y, Z)$	$V[Z]$

Example 2

- Using R, calculate the sample covariance between each of the pairs of variables used in Example 1
- What do these values imply?

Sample Correlation

- Often we do not have all of the observations in populations and we need to estimate parameters.

- **Sample correlation** estimates the actual correlation between two random variables:

$$\rho = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{S_{XY}}{\sqrt{S_X^2 \cdot S_Y^2}}. \tag{4}$$

- Where S_{XY}, S_X, and S_Y are the sample covariance and standard deviations.

Sample Correlation in R

- R function: `cor(x, y, method = "pearson")`
  - `x`: a numeric vector, matrix, or data frame.
  - If `x` is a vector we must give a vector `y`.
  - `method = "pearson"` is the default method.
- `method = "kendall"` to measure the ordinal association between two measured quantities.
- `method = "spearman"` (rank correlation) to assess monotonic relationships between two measured quantities.

Correlation Matrix

Variable	X	Y	Z
X	$\rho_{XX} = 1$	ρ_{XY}	ρ_{XZ}
Y	ρ_{XY}	$\rho_{YY} = 1$	ρ_{YZ}
Z	ρ_{XZ}	ρ_{YZ}	$\rho_{ZZ} = 1$

Example 3

- Using R, calculate the sample correlation between each of the pairs of variables used in Example 1
- What do these values imply?

Properties of Sample Correlation

- Values near 0 indicate a weak linear relationship.
- Linear relationship strength increases as ρ moves towards 1 or -1.
- If ρ is close to 1 or -1 then the scatterplot will be close to a straight line.

Effect of Outliers

- An *outlier* is a data point that differs significantly from other observations.
- As the measures are based on differences from mean values, **outliers can have a large impact on the estimates of linear association.**
- We have covered methods to identify possible outliers (boxplots).
- Being aware of the impacts out potential outliers can help you make better statistical decisions.

Example 4

- Load the *Outliers.csv* data into your workspace to complete the following:
  1. Generate a boxplot for each of the four variables.
  2. Generate a scatterplot for each of the pairs of variables (Y1 vs X1 and Y2 vs X2).
     - Include the correlation coefficient in your plots.
  3. Generate a scatterplot for each of the pairs of variables (Y1 vs X1 and Y2 vs X2) **without** the outliers.
     - Include the correlation coefficient in your plots.
- What did you notice?

Example 5

- Using the *Football22.csv* data complete the following:
  1. Create the functions under the *Required Functions* heading.
  2. Create a pairs plot of all of the numeric variables.
  3. Create a pairs plot of all the numeric variables and include a LOWESS curve, the correlation coefficients, and the histograms.
- Does there appear to be any linear relationships between the variables?

Final Thoughts

- **Correlation only measures linear association**
  - The variables in Example 5 are related but not necessarily linearly.
- Based on correlation alone we can not determine causality (independent and dependent) variables.
  - ρ can not be used for prediction.
- We will use statistical models (regression) to investigate the relationship between variables of interest.

Exercise 1

- Load the *Cars93* dataset from the *MASS* R package and examine the linear relationships (covariance & correlation) between the following pairs of variables:

  1. *Price* and *RPM*
  2. *Price* and *EngineSize*
  3. *Horsepower* and *EngineSize*
  4. *Weight* and *Fuel.tank.capacity*
  5. *Length* and *Width*

Exercise 2

- Take some time to examine possible linear relationships (covariance & correlation) between continous variables in your project data.

References & Resources

1. Evans, J. R., Olson, D. L., & Olson, D. L. (2007). *Statistics, data analysis, and decision modeling.* Upper Saddle River, NJ: Pearson/Prentice Hall.
2. Devore, J. L., Berk, K. N., & Carlton, M. A. (2012). *Modern mathematical statistics with applications (Second Edition).* New York: Springer.

- https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient
- https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient