Bivariate Relationships Between Variables

BUS 735: Business Decision Making and Research

1

Goals

- Specific goals:
  - Detect *relationships* between variables.
  - Be able to prescribe appropriate statistical methods for measuring relationship based on scale of measurement.
- Learning objectives:
  - LO1: Construct and test hypotheses using a variety of bivariate statistical methods to compare characteristics between two populations.
  - LO2: Construct and use advanced multivariate models to identify complex relationships among multiple variables; including regression models, limited dependent variable models, and analysis of variance and covariance models.

2 Correlation

2.1 Linear and Monotonic Relationships

Correlation

Correlation

Correlation: when two variables move together in some fashion.

Correlations measure *monotonic relationships*.

- Positive: When one variable increases, the other tends to increase.
- Negative: When one variable increases, the other tends to decrease.

Common Focus: Linear Relationships

Linear relationships: Visually illustrated with a straight line

Common monotonic relationships, but not linear:

- Employment experience and income
- Employment experience and productivity
- Wealth and consumer spending

2.2 Pearson vs Spearman Correlation

Pearson vs Spearman Correlation

Pearson linear correlation coefficient

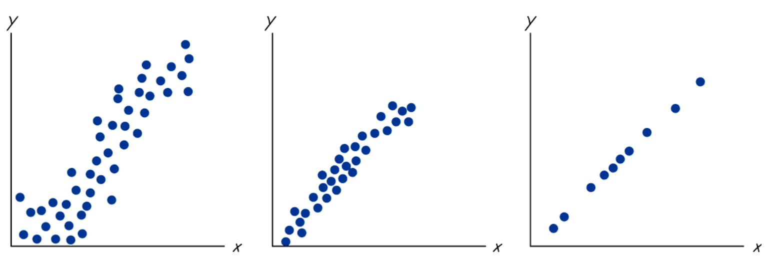
- Measure of the strength of the **linear relationship**
- Parametric test for interval or ratio data
- Null hypothesis: zero linear correlation between two variables.
- Alternative hypothesis: linear correlation exists (either positive or negative) between two variables.

Spearman linear correlation coefficient

- Measure of the strength of a **monotonic relationship**
- Non-parametric test for ordinal, interval, and ratio data
- Pearson computation with *ranks* instead of actual data
- Same hypotheses.

2.3 Strength of Correlation

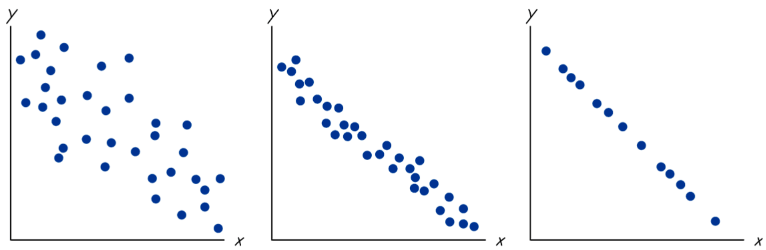
Positive linear correlation



(a) Positive correlation between x and y **(b)** Strong positive correlation between x and y **(c)** Perfect positive correlation between x and y

- Positive correlation: move in the same direction.
- Stronger correlation: closer to 1.0
- Perfect positive correlation: $\rho = 1.0$

Negative linear correlation



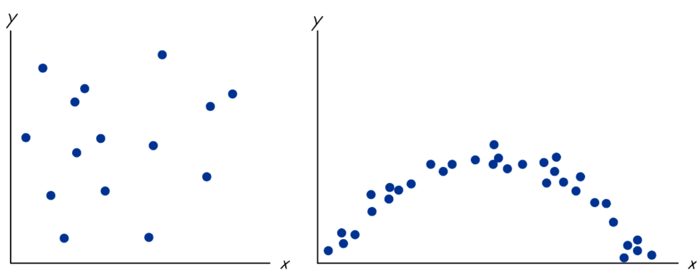
(d) Negative correlation between x and y

(e) Strong negative correlation between x and y

(f) Perfect negative correlation between x and y

- Negative correlation: move in opposite directions.
- Stronger correlation: closer to -1.0
- Perfect negative correlation: $\rho = -1.0$

No linear correlation



(g) No correlation between x and y

(h) Nonlinear relationship between x and y

- Panel (g): no relationship at all.
- Panel (h): strong relationship, but not a *linear* relationship.
  - Cannot use regular correlation to detect this.

3 Chi-Square Test of Independence

3.1 Definition and Example

Chi-Square Test for Independence

- Used to determine if two categorical variables (eg: nominal) are related.
- Example: Suppose a hotel manager surveys guest who indicate they will not return:

	Reason for Not Returning		
Reason for Stay	Price	Location	Amenities
Personal/Vacation	56	49	0
Business	20	47	27

- Data in the table are always frequencies that fall into individual categories.
- Could use this table to test if two variables are independent.

3.2 Hypothesis Test

Chi-Square Test of independence

- **Null hypothesis**: there is no relationship between the row variable and the column variable (independent)
- **Alternative hypothesis**: There is a relationship between the row variable and the column variable (dependent).

4 Bivariate Regression

4.1 Definition

Bivariate Regression

- Regression line: equation of the line that describes the linear relationship between variable x and variable y.
- Need to assume that *independent variables* influence *dependent variables*.
  - x: *independent* or *explanatory* variable.
  - y: *dependent* or *outcome* variable.
  - Variable x can influence variable y, but not vice versa.
- Example: How does advertising expenditures affect sales revenue?

4.2 Population vs. Sample

Regression line

Population regression line:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- The population coefficients β_0 and β_1 describing the relationship between x and y are unknown.
- Since x and y are not perfectly correlated, ϵ_i is the error term.

Sample regression line:

$$y_i = b_0 + b_1 x_i + e_i$$

- Not perfectly correlated, e_i is the sample error term.

4.3 Predicted Values and Residuals

Predicted Values and Residuals

For a given x_i, the **predicted value** for y_i, denoted $\hat{y}_i$, is...

$$\hat{y}_i = b_0 + b_1 x_i$$

- This is not likely be the actual value for y_i.

Residual is the difference *in the sample* between the actual value of y_i and the predicted value, $\hat{y}$.

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$$