Finding Relationships Among 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.
  - Detect how outcome variables can be explained by one or more explanatory variables.
- 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 Relationships Between Two Variables

2.1 Correlation

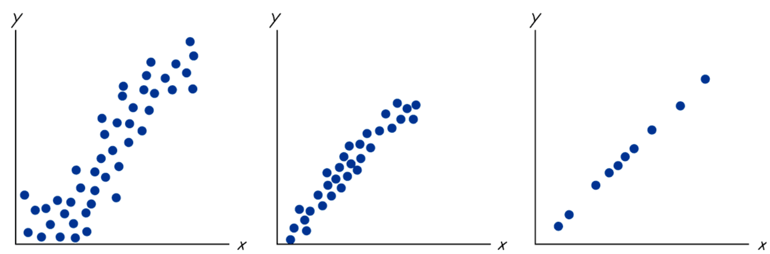
Correlation

- A **correlation** exists between two variables when one of them is related to the other in some way.
- **Pearson linear correlation coefficient** is a measure of the strength of the linear relationship between two variables.
  - Parametric test for interval or ratio data
  - Null hypothesis: there is zero linear correlation between two variables.

- Alternative hypothesis: there is a linear correlation (either positive or negative) between two variables.
- Measures strength of *linear* relationship

- **Spearman linear correlation coefficient**
  - Non-parametric test for ordinal, interval, and ratio data
  - Pearson computation with *ranks* instead of actual data
  - Same hypotheses.
  - Measures strength of *linear* relationship in *ranks*, more general monotonic relationships in interval/ratio data are permitted.

Positive linear correlation



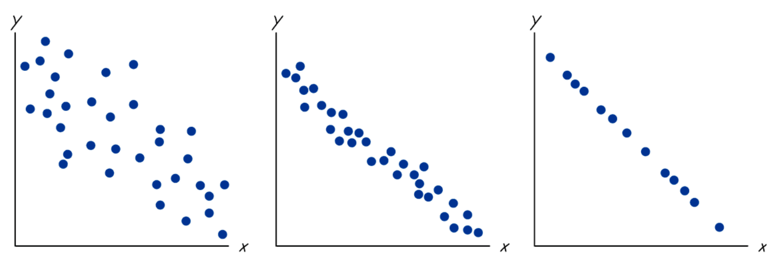
(a) Positive correlation between x and y

(b) Strong positive correlation between x and y

(c) Perfect positive correlation between x and y

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

Negative linear correlation



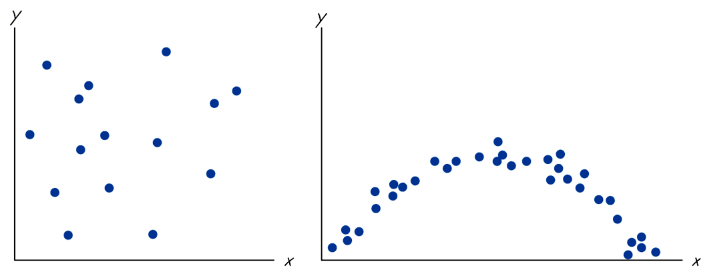
(d) Negative correlation between x and y

(e) Strong negative correlation between x and y

(f) Perfect negative correlation between x and y

- Negative correlation: two variables move in opposite directions.
- Stronger correlation: closer the correlation coefficient is 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.

2.2 Chi-Square Test of Independence

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.

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).

2.3 Bivariate Regression

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?

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.

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, $\hat{y}$.

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

3 Multiple Regression

3.1 Functional Form

Multiple Regression

Multiple regression line **(population)**:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_2 + ... + \beta_k x_k + \epsilon_i$$

Multiple regression line **(sample)**:

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_2 + ... + b_k x_k + e_i$$

- k: number of explanatory variables

Interpreting the slope

- Interpreting the slope, β: amount the y is predicted to increase when increasing x by one unit.
- When $\beta < 0$ there is a negative linear relationship.
- When $\beta > 0$ there is a positive linear relationship.
- When $\beta = 0$ there is no linear relationship between x and y.
- Statistical packages report sample estimates for coefficients, along with...
  - Standard errors of the coefficients
  - T-test statistics for $H_0 : \beta = 0$.
  - P-values of the T-tests.
  - Confidence intervals for the coefficients.

3.2 Variance Decomposition

Sum of Squares Measures of Variation

- **Sum of Squares Regression (SSR)**: measure of the amount of variability in the dependent (Y) variable that is explained by the independent variables (X's).

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- **Sum of Squares Error (SSE)**: measure of the unexplained variability in the dependent variable.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Sum of Squares Measures of Variation

- **Sum of Squares Total (SST)**: measure of the total variability in the dependent variable.
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
- SST = SSR + SSE.

Coefficient of determination

- The **coefficient of determination** is the percentage of variability in y that is explained by x.
$$R^2 = \frac{SSR}{SST}$$
- R^2 will always be between 0 and 1. The closer R^2 is to 1, the better x is able to explain y.
- The more variables you add to the regression, the higher R^2 will be.

Adjusted R^2

- R^2 will likely increase (slightly) even by adding nonsense variables.
- Adding such variables increases in-sample fit, but will likely hurt out-of-sample forecasting accuracy.
- The Adjusted R^2 penalizes R^2 for additional variables.
$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right)$$
- When the adjusted R^2 increases when adding a variable, then the additional variable really did help explain the dependent variable.
- When the adjusted R^2 decreases when adding a variable, then the additional variable does not help explain the dependent variable.

F-test for Regression Fit

- F-test for Regression Fit: Tests if the regression line explains the data.
- Very, very, very similar to ANOVA F-test.
- $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$.
- H_1 : At least one of the variables has explanatory power (i.e. at least one coefficient is not equal to zero).
$$F = \frac{SSR/(k-1)}{SSE/(n-k)}$$
- Where k is the number of explanatory variables.

4 Regression Assumptions

4.1 Assumptions from the CLT

Assumptions from the CLT

- Using the normal distribution to compute p-values depends on results from the Central Limit Theorem.
- Sufficiently large sample size (much more than 30).
  - Useful for normality result from the Central Limit Theorem
  - Also necessary as you increase the number of explanatory variables.
- Normally distributed dependent and independent variables
  - Useful for small sample sizes, but not essential as sample size increases.
- Types of data:
  - Dependent variable must be interval or ratio.
  - Independent variable can be interval, ratio, *or a dummy variable.*

4.2 Regression-Specific Assumptions

Regression-Specific Assumptions

- Linearity: a straight line reasonably describes the data.
  - Exceptions: experience on productivity, ordinal data like education level on income.
  - Consider transforming variables.
- Stationarity:
  - The central limit theorem: behavior of statistics as sample size approaches infinity!
  - The mean and variance must exist and be constant.
  - Big issue in economic and financial time series.
- Exogeneity of explanatory variables.
  - Dependent variable must not influence explanatory variables.
  - Explanatory variables must not be influenced by excluded variables that can influence dependent variable.
  - Example problem: how does advertising affect sales?