Finding Relationships Among Variables

BUS 735: Business Decision Making and Research

Goals

- Specific goals:
  - Detect how outcome variables can be explained by multiple explanatory variables.
- Learning objectives:
  - 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.

Goals

- Specific goals:
    - Detect how outcome variables can be explained by multiple explanatory variables.
- Learning objectives:
    - 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.

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

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

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

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

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 Coefficients

- 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.

Interpreting Coefficients

- 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.

Interpreting Coefficients

- 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.

Interpreting Coefficients

- 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.

Interpreting Coefficients

- 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.

Interpreting Coefficients

- 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.

Interpreting Coefficients

- 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.

Interpreting Coefficients

- 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.

Interpreting Coefficients

- 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.

Sum of Squares Measures of Variation

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

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

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

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

Sum of Squares Measures of Variation

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

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

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

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

Sum of Squares Measures of Variation

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

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

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

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

Sum of Squares Measures of Variation

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

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

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

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

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.

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.

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.

$$R^2 = \frac{SSE}{SST}$$

- 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.

Coefficient of determination

- The **coefficient of determination** is the percentage of variability in y that is explained by x.

$$R^2 = \frac{SSE}{SST}$$

- 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.

Coefficient of determination

- The **coefficient of determination** is the percentage of variability in y that is explained by x.

$$R^2 = \frac{SSE}{SST}$$

- 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.

Coefficient of determination

- The **coefficient of determination** is the percentage of variability in y that is explained by x.

$$R^2 = \frac{SSE}{SST}$$

- 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

Problem: Adding variables not always good

- R^2 will likely increase (slightly) even by adding nonsense variables.
- Adding such variables increases in-sample fit by chance
- Adding nonsense hurts out-of-sample forecasting accuracy

Adjusted R^2

- Adjusted R^2 penalizes R^2 for additional variables.
- When adjusted R^2 increases $\rightarrow$
  Additional variable helps explain outcome variable.
- When adjusted R^2 decreases $\rightarrow$
  Additional variable *does not* help explain outcome variable.

Adjusted R^2

Problem: Adding variables not always good

- R^2 will likely increase (slightly) even by adding nonsense variables.
- Adding such variables increases in-sample fit by chance
- Adding nonsense hurts out-of-sample forecasting accuracy

Adjusted R^2

- Adjusted R^2 penalizes R^2 for additional variables.
- When adjusted R^2 increases $\rightarrow$
  Additional variable helps explain outcome variable.
- When adjusted R^2 decreases $\rightarrow$
  Additional variable *does not* help explain outcome variable.

Adjusted R^2

Problem: Adding variables not always good

- R^2 will likely increase (slightly) even by adding nonsense variables.
- Adding such variables increases in-sample fit by chance
- Adding nonsense hurts out-of-sample forecasting accuracy

Adjusted R^2

- Adjusted R^2 penalizes R^2 for additional variables.
- When adjusted R^2 increases $\rightarrow$
  Additional variable helps explain outcome variable.
- When adjusted R^2 decreases $\rightarrow$
  Additional variable *does not* help explain outcome variable.

Adjusted R^2

Problem: Adding variables not always good

- R^2 will likely increase (slightly) even by adding nonsense variables.
- Adding such variables increases in-sample fit by chance
- Adding nonsense hurts out-of-sample forecasting accuracy

Adjusted R^2

- Adjusted R^2 penalizes R^2 for additional variables.
- When adjusted R^2 increases $\rightarrow$
  Additional variable helps explain outcome variable.
- When adjusted R^2 decreases $\rightarrow$
  Additional variable *does not* help explain outcome variable.

Adjusted R^2

Problem: Adding variables not always good

- R^2 will likely increase (slightly) even by adding nonsense variables.
- Adding such variables increases in-sample fit by chance
- Adding nonsense hurts out-of-sample forecasting accuracy

Adjusted R^2

- Adjusted R^2 penalizes R^2 for additional variables.
- When adjusted R^2 increases $\rightarrow$
  Additional variable helps explain outcome variable.
- When adjusted R^2 decreases $\rightarrow$
  Additional variable *does not* help explain outcome variable.

Adjusted R^2

Problem: Adding variables not always good

- R^2 will likely increase (slightly) even by adding nonsense variables.
- Adding such variables increases in-sample fit by chance
- Adding nonsense hurts out-of-sample forecasting accuracy

Adjusted R^2

- Adjusted R^2 penalizes R^2 for additional variables.
- When adjusted R^2 increases $\rightarrow$
  Additional variable helps explain outcome variable.
- When adjusted R^2 decreases $\rightarrow$
  Additional variable *does not* help explain outcome variable.

Adjusted R^2

Problem: Adding variables not always good

- R^2 will likely increase (slightly) even by adding nonsense variables.
- Adding such variables increases in-sample fit by chance
- Adding nonsense hurts out-of-sample forecasting accuracy

Adjusted R^2

- Adjusted R^2 penalizes R^2 for additional variables.
- When adjusted R^2 increases $\rightarrow$
  Additional variable helps explain outcome variable.
- When adjusted R^2 decreases $\rightarrow$
  Additional variable *does not* help explain outcome variable.

Adjusted R^2

Problem: Adding variables not always good

- R^2 will likely increase (slightly) even by adding nonsense variables.
- Adding such variables increases in-sample fit by chance
- Adding nonsense hurts out-of-sample forecasting accuracy

Adjusted R^2

- Adjusted R^2 penalizes R^2 for additional variables.
- When adjusted R^2 increases $\rightarrow$
  Additional variable helps explain outcome variable.
- When adjusted R^2 decreases $\rightarrow$
  Additional variable *does not* help explain outcome variable.

F-test for Regression Fit

Test if the regression line explains the data

Hypotheses

- $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$.
- H_1 : At least one of the variables explains outcome (i.e. at least one $\beta_j \neq 0$).

Test Statistic

$$F = \frac{SSE/k}{SSR/(n-k-1)}$$

- k: number of explanatory variables
- Ratio of *explained variation* relative to *unexplained variation*

F-test for Regression Fit

Test if the regression line explains the data

Hypotheses

- $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$.
- H_1 : At least one of the variables explains outcome (i.e. at least one $\beta_j \neq 0$).

Test Statistic

$$F = \frac{SSE/k}{SSR/(n-k-1)}$$

- k: number of explanatory variables
- Ratio of *explained variation* relative to *unexplained variation*

F-test for Regression Fit

Test if the regression line explains the data

Hypotheses

- $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$.
- H_1 : At least one of the variables explains outcome (i.e. at least one $\beta_j \neq 0$).

Test Statistic

$$F = \frac{SSE/k}{SSR/(n-k-1)}$$

- k: number of explanatory variables
- Ratio of *explained variation* relative to *unexplained variation*

F-test for Regression Fit

Test if the regression line explains the data

Hypotheses

- $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$.
- H_1 : At least one of the variables explains outcome (i.e. at least one $\beta_j \neq 0$).

Test Statistic

$$F = \frac{SSE/k}{SSR/(n-k-1)}$$

- k: number of explanatory variables
- Ratio of *explained variation* relative to *unexplained variation*

F-test for Regression Fit

Test if the regression line explains the data

Hypotheses

- $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$.
- H_1 : At least one of the variables explains outcome (i.e. at least one $\beta_j \neq 0$).

Test Statistic

$$F = \frac{SSE/k}{SSR/(n-k-1)}$$

- k: number of explanatory variables
- Ratio of *explained variation* relative to *unexplained variation*

F-test for Regression Fit

Test if the regression line explains the data

Hypotheses

- $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0.$
- H_1 : At least one of the variables explains outcome (i.e. at least one $\beta_j \neq 0$).

Test Statistic

$$F = \frac{SSE/k}{SSR/(n-k-1)}$$

- k: number of explanatory variables
- Ratio of *explained variation* relative to *unexplained variation*

Assumptions from the CLT

Large Sample Size

- Useful for normality result from the Central Limit Theorem
- Also necessary as you increase the number of explanatory variables.

Normally Distributed Variables

Useful for small sample sizes, but not essential as sample size increases.

Scale of Measurement

- Dependent variable must be interval or ratio.
- Independent variable can be interval, ratio, *or a dummy variable.*

Assumptions from the CLT

Large Sample Size

- Useful for normality result from the Central Limit Theorem
- Also necessary as you increase the number of explanatory variables.

Normally Distributed Variables

Useful for small sample sizes, but not essential as sample size increases.

Scale of Measurement

- Dependent variable must be interval or ratio.
- Independent variable can be interval, ratio, *or a dummy variable.*

Assumptions from the CLT

Large Sample Size

- Useful for normality result from the Central Limit Theorem
- Also necessary as you increase the number of explanatory variables.

Normally Distributed Variables

Useful for small sample sizes, but not essential as sample size increases.

Scale of Measurement

- Dependent variable must be interval or ratio.
- Independent variable can be interval, ratio, *or a dummy variable.*

Assumptions from the CLT

Large Sample Size

- Useful for normality result from the Central Limit Theorem
- Also necessary as you increase the number of explanatory variables.

Normally Distributed Variables

Useful for small sample sizes, but not essential as sample size increases.

Scale of Measurement

- Dependent variable must be interval or ratio.
- Independent variable can be interval, ratio, *or a dummy variable.*

Assumptions from the CLT

Large Sample Size

- Useful for normality result from the Central Limit Theorem
- Also necessary as you increase the number of explanatory variables.

Normally Distributed Variables

Useful for small sample sizes, but not essential as sample size increases.

Scale of Measurement

- Dependent variable must be interval or ratio.
- Independent variable can be interval, ratio, *or a dummy variable.*

Assumptions from the CLT

Large Sample Size

- Useful for normality result from the Central Limit Theorem
- Also necessary as you increase the number of explanatory variables.

Normally Distributed Variables

Useful for small sample sizes, but not essential as sample size increases.

Scale of Measurement

- Dependent variable must be interval or ratio.
- Independent variable can be interval, ratio, *or a dummy variable*.

Assumptions from the CLT

Large Sample Size

- Useful for normality result from the Central Limit Theorem
- Also necessary as you increase the number of explanatory variables.

Normally Distributed Variables

Useful for small sample sizes, but not essential as sample size increases.

Scale of Measurement

- Dependent variable must be interval or ratio.
- Independent variable can be interval, ratio, *or a dummy variable*.

Regression-Specific Assumptions

Linearity

- Straight line describes the relationship
- Exceptions: experience/productivity

Stationarity

- The mean and variance must exist and converge
- Big issue in economic and financial time series

Exogeneity

- Dependent variable must not influence explanatory variables
- Omitted variables must not influence both outcome and explanatory variables
- Examples: Advertising/Sales, Violent Crime/Ice Cream

Regression-Specific Assumptions

Linearity

- Straight line describes the relationship
- Exceptions: experience/productivity

Stationarity

- The mean and variance must exist and converge
- Big issue in economic and financial time series

Exogeneity

- Dependent variable must not influence explanatory variables
- Omitted variables must not influence both outcome and explanatory variables
- Examples: Advertising/Sales, Violent Crime/Ice Cream

Regression-Specific Assumptions

Linearity

- Straight line describes the relationship
- Exceptions: experience/productivity

Stationarity

- The mean and variance must exist and converge
- Big issue in economic and financial time series

Exogeneity

- Dependent variable must not influence explanatory variables
- Omitted variables must not influence both outcome and explanatory variables
- Examples: Advertising/Sales, Violent Crime/Ice Cream

Regression-Specific Assumptions

Linearity

- Straight line describes the relationship
- Exceptions: experience/productivity

Stationarity

- The mean and variance must exist and converge
- Big issue in economic and financial time series

Exogeneity

- Dependent variable must not influence explanatory variables
- Omitted variables must not influence both outcome and explanatory variables
- Examples: Advertising/Sales, Violent Crime/Ice Cream

Regression-Specific Assumptions

Linearity

- Straight line describes the relationship
- Exceptions: experience/productivity

Stationarity

- The mean and variance must exist and converge
- Big issue in economic and financial time series

Exogeneity

- Dependent variable must not influence explanatory variables
- Omitted variables must not influence both outcome and explanatory variables
- Examples: Advertising/Sales, Violent Crime/Ice Cream

Regression-Specific Assumptions

Linearity

- Straight line describes the relationship
- Exceptions: experience/productivity

Stationarity

- The mean and variance must exist and converge
- Big issue in economic and financial time series

Exogeneity

- Dependent variable must not influence explanatory variables
- Omitted variables must not influence both outcome and explanatory variables
- Examples: Advertising/Sales, Violent Crime/Ice Cream

Regression-Specific Assumptions

Linearity

- Straight line describes the relationship
- Exceptions: experience/productivity

Stationarity

- The mean and variance must exist and converge
- Big issue in economic and financial time series

Exogeneity

- Dependent variable must not influence explanatory variables
- Omitted variables must not influence both outcome and explanatory variables
- Examples: Advertising/Sales, Violent Crime/Ice Cream

Regression-Specific Assumptions

Linearity

- Straight line describes the relationship
- Exceptions: experience/productivity

Stationarity

- The mean and variance must exist and converge
- Big issue in economic and financial time series

Exogeneity

- Dependent variable must not influence explanatory variables
- Omitted variables must not influence both outcome and explanatory variables
- Examples: Advertising/Sales, Violent Crime/Ice Cream

Regression-Specific Assumptions

Linearity

- Straight line describes the relationship
- Exceptions: experience/productivity

Stationarity

- The mean and variance must exist and converge
- Big issue in economic and financial time series

Exogeneity

- Dependent variable must not influence explanatory variables
- Omitted variables must not influence both outcome and explanatory variables
- Examples: Advertising/Sales, Violent Crime/Ice Cream

Regression-Specific Assumptions

Linearity

- Straight line describes the relationship
- Exceptions: experience/productivity

Stationarity

- The mean and variance must exist and converge
- Big issue in economic and financial time series

Exogeneity

- Dependent variable must not influence explanatory variables
- Omitted variables must not influence both outcome and explanatory variables
- Examples: Advertising/Sales, Violent Crime/Ice Cream