

# Regression Analysis

BUS 735: Business Decision Making and Research

## 1

### Goals of this section

#### Specific goals

- Learn how to detect relationships between ordinal and categorical variables.
- Learn how to estimate a linear relationship between many variables.

#### Learning objectives

- LO2: Be able to construct and use multiple regression models (including some limited dependent variable models) to construct and test hypotheses considering complex relationships among multiple variables.
- LO6: Be able to use standard computer packages such as SPSS and Excel to conduct the quantitative analyses described in the learning objectives above.
- LO7: Have a sound familiarity of various statistical and quantitative methods in order to be able to approach a business decision problem and be able to select appropriate methods to answer the question.

## 2 Relationships Between Two Variables

### 2.1 Correlation

#### Correlation

- Pearson linear correlation coefficient: a value between -1 and +1 that is used to measure the strength of a positive or negative linear relationship.
  - Valid for interval or ratio data.
  - Not appropriate for ordinal or nominal data.
  - Test depends on assumptions behind the central limit theorem (CLT)

- Spearman rank correlation: non-parametric test.
  - Valid for small sample sizes (when assumptions of CLT are violated)
  - Appropriate for interval, ratio, and even ordinal data.
  - Still makes no sense to use for nominal data.

## 2.2 Chi-Squared Test of Independence

### Chi-Squared 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 Stay	Reason for Not Returning		
	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.

### Test of independence

- **Null hypothesis:** there is no relationship between the row variable and the column variable.
- **Alternative hypothesis:** The two variables are dependent.
- Test statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- $O$ : observed frequency in a cell from the contingency table.
- $E$ : expected frequency assuming variables are independent.
- Large  $\chi^2$  values indicate variables are dependent (reject the null hypothesis).

## 3 Regression

### 3.1 Single Variable Regression

#### 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* variable.
  - Variable  $x$  can influence the value for variable  $y$ , but not vice versa.
- Example: How does smoking affect lung capacity?
- Example: How does advertising affect sales?

### Regression line

- Population regression line:

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

- The actual coefficients  $\beta_0$  and  $\beta_1$  describing the relationship between  $x$  and  $y$  are unknown.
- Use sample data to come up with an estimate of the regression line:

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

- Since  $x$  and  $y$  are not perfectly correlated, still need to have an error term.

### Predicted values and residuals

- Given a value for  $x_i$ , can come up with a **predicted value** for  $y_i$ , denoted  $\hat{y}_i$ .

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

- 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,  $\hat{y}_i$ .

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

## 3.2 Multiple Regression

### Multiple Regression

- Multiple regression line (population):

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1} + \epsilon_i$$

- Multiple regression line (sample):

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

- $k$ : number of parameters (coefficients) you are estimating.
- $\epsilon_i$ : error term, since linear relationship between the  $x$  variables and  $y$  are not perfect.
- $e_i$ : residual = the difference between the predicted value  $\hat{y}$  and the actual value  $y_i$ .

### Least Squares Estimate

- How should we obtain the “best fitting line”.
- Ordinary least squares (OLS) method.
- Choose sample estimates for the regression coefficients that minimizes:

$$\sum_{i=0}^n (y_i - \hat{y}_i)^2$$

### Interpreting the slope

- Interpreting the slope,  $\beta$ : 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  $y$ .
- SPSS reports sample estimates for coefficients, along with...
  - Estimates of the standard errors.
  - T-test statistics for  $H_0 : \beta = 0$ .
  - P-values of the T-tests.
  - Confidence intervals for the coefficients.

### Example: Public Expenditure

- Data from 1960 about public expenditures per capita, and variables that may influence it.
- In SPSS, choose **Analyze** menu and select **Regression** and **Linear**.
- Select **EX** (Expenditure per capita) as your dependent variable. This is the variable you are interested in explaining.
- Select your independent (aka explanatory) variables. These are the variables that you think can explain the dependent variable. I suggest you select these:

- ECAB: Economic Ability
- MET: Metropolitan
- GROW: Growth rate of population
- WEST: Western state = 1.

### Example: Public Expenditure

- If the percentage of the population living in metropolitan areas is expected to increase by 1%, what change should we expect in public expenditure?
- Is this change statistically significantly different from zero?
- Accounting for economic ability, metropolitan population, and population growth, how much more do Western states spend on public expenditure per capita?

## 3.3 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.

$$R_{\text{adj}}^2 = 1 - \frac{n-1}{n-k-1} (1 - R^2)$$

- 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.
- $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$ .
- $H_1$  : 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.

### Example: Public Expenditure

- In the previous example, how much of the variability in public expenditure is explained by the following four variables:
  - ECAB: Economic Ability
  - MET: Metropolitan
  - GROW: Growth rate of population
  - WEST: Western state = 1.
- Is the combination of these variables significant in explaining public expenditure?
- Re-run the regression, this time also including:
  - YOUNG: Percentage of population that is young.
  - OLD: Percentage of population that is old.

### Example: Public Expenditure

- What happened to the coefficient of determination?
- What happened to the adjusted coefficient of determination? What is your interpretation?
- What happened to the estimated effect of the other variables: metropolitan area? Western state?

## 4 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 Crucial Assumptions for Regression

### Crucial Assumptions for Regression

- 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?

## 4.3 Multicollinearity

### Multicollinearity

- **Multicollinearity:** when two or more of the explanatory variables are highly correlated.
- With multicollinearity, it is difficult to determine the effect coming from a specific individual variable.
- Correlated variables will have standard errors for coefficients will be large (coefficients will be statistically insignificant).
- Examples:
  - experience and age used to predict productivity
  - size of store (sq feet) and store sales used to predict demand for inventories.
  - parent's income and parent's education used to predict student performance.
- Perfect multicollinearity - when two variables are perfectly correlated.



## 4.4 Homoscedasticity

### Homoscedasticity

- **Homoscedasticity:** when the variance of the error term is constant (it does not depend on other variables).
- Counter examples (heteroscedasticity):
  - Impact of income on demand for houses.
  - Many economic and financial variables related to income suffer from this.
- Heteroscedasticity is not too problematic:
  - Estimates will still be unbiased.
  - Your standard errors will be downward biased (reject more than you should).
- May be evidence of a bigger problem: linearity or stationarity.