BNAD 276
Lecture 10
Simple Linear Regression Model

Phuong Ho

May 30, 2017

Outline

Outline

Simple Linear Regression Model

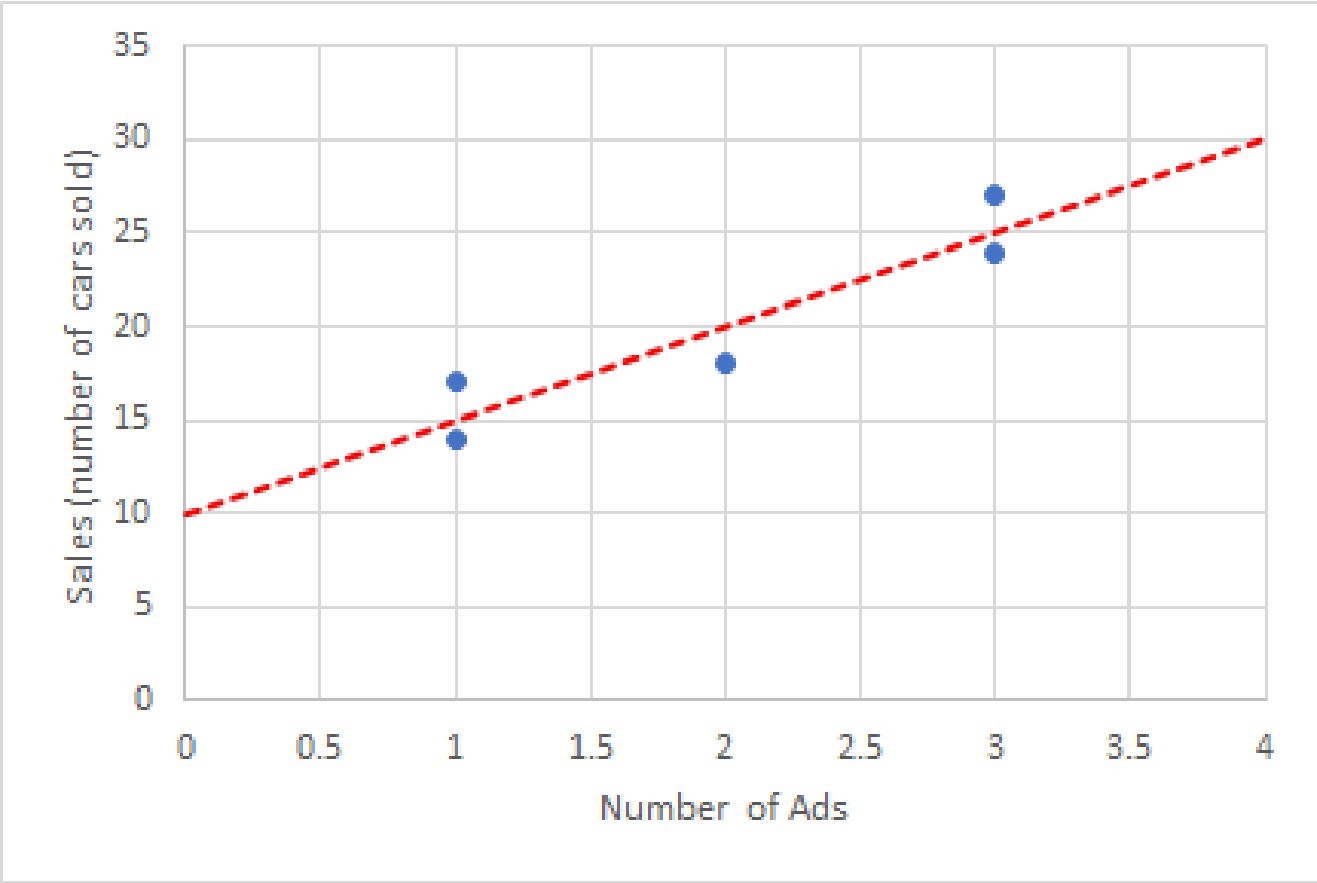
- Managerial decisions often are based on the relationship between two or more variables.
- Regression analysis can be used to develop an equation showing how the variables are related.
- The variable being predicted is called the dependent variable and is denoted by y.
- The variables being used to predict the value of the dependent variable are called the independent variables and are denoted by x.

Motivating Example

Reed Auto periodically has a special week-long sale. As part of the advertising campaign Reed runs one or more television commercials during the weekend preceding the sale. Data from a sample of 5 previous sales are shown on the next slide.

Number of TV Ads X	Number of Cars Sold Y
1	14
3	24
2	18
1	17
3	27

Motivating Example, cont'd



We want to examine the possible relation between Number of Ads and Sales. The process of finding the best linear line to fit that relation is called **Linear Regression**.

Simple Linear Regression

- **Simple linear regression** involves **one independent variable** and one dependent variable.
- The relationship between the two variables is approximated by a straight line.
- Regression analysis involving **two or more independent variables** is called **multiple regression**.

Simple Linear Regression Model

- The equation that describes how y is related to x and an error term is called the regression model.

Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where

- β_0, β_1 are called parameters of the model,
- ϵ is a random variable called the error term.

Simple Linear Regression Equation

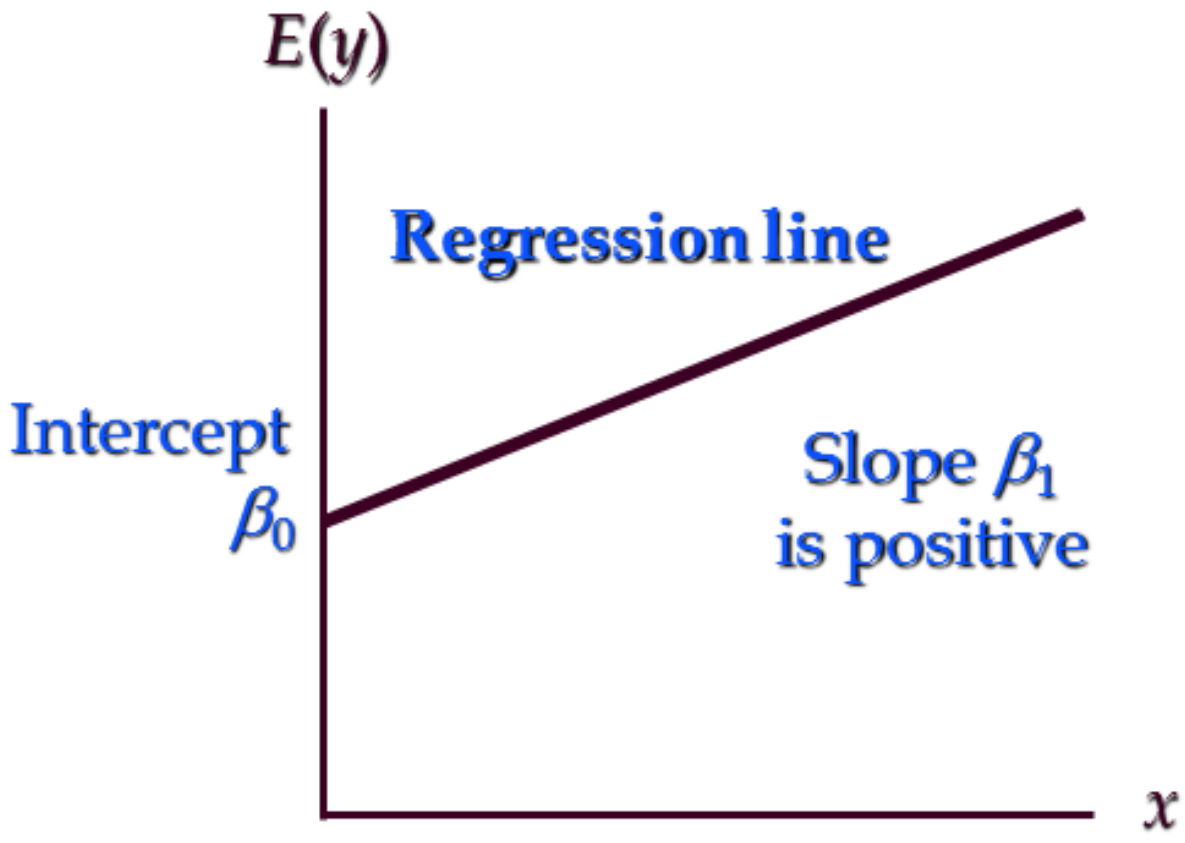
- The simple linear regression equation is:

$$E(Y) = \beta_0 + \beta_1 X$$

  where
  - β_0 is the y intercept of the regression line.
  - β_1 is the slope of the regression line.
  - $E(Y)$ is the expected value of Y for a given X value.
- Graph of the regression equation is a straight line.

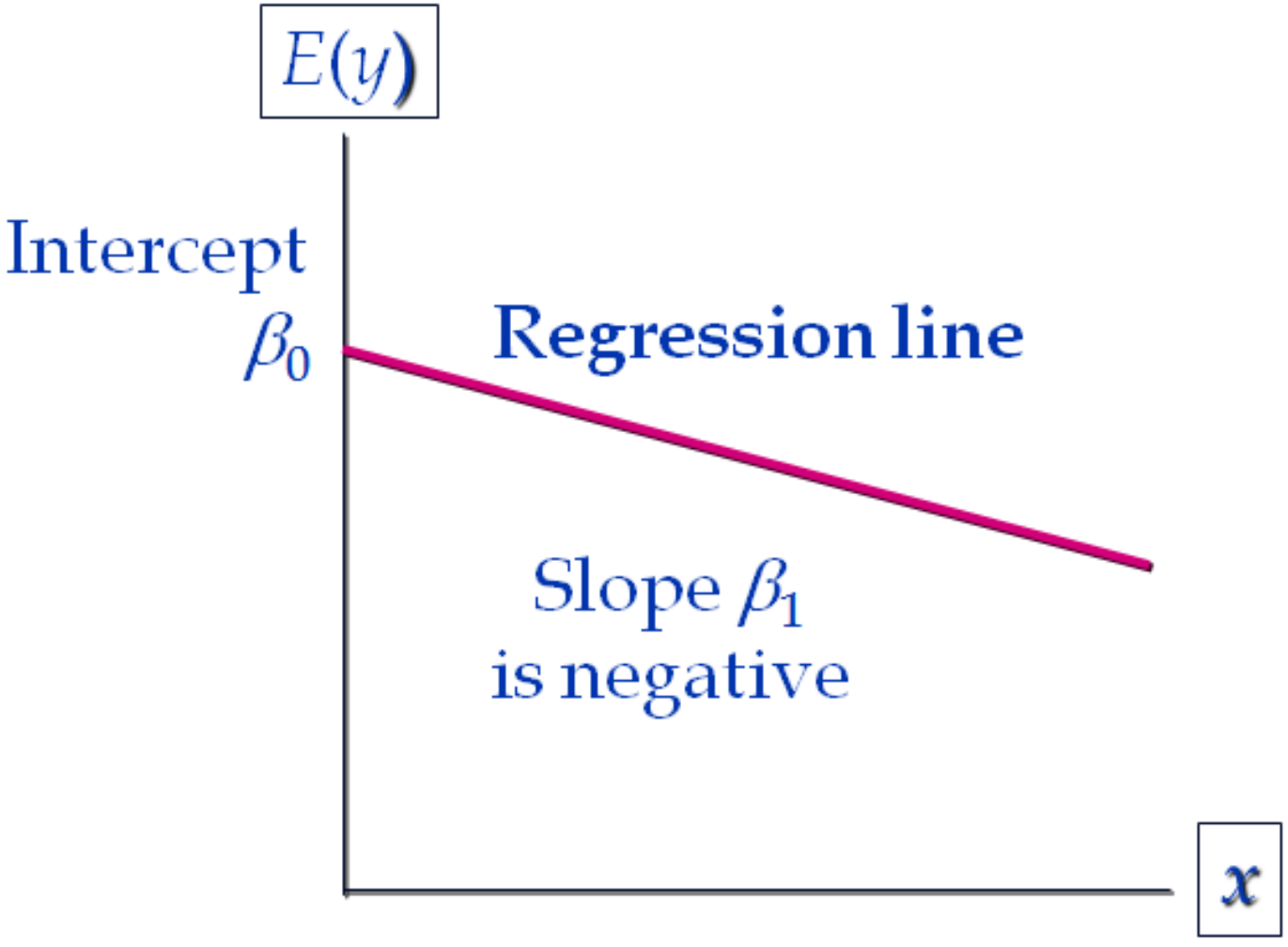
Simple Linear Regression Equation

- Positive Linear Relationship



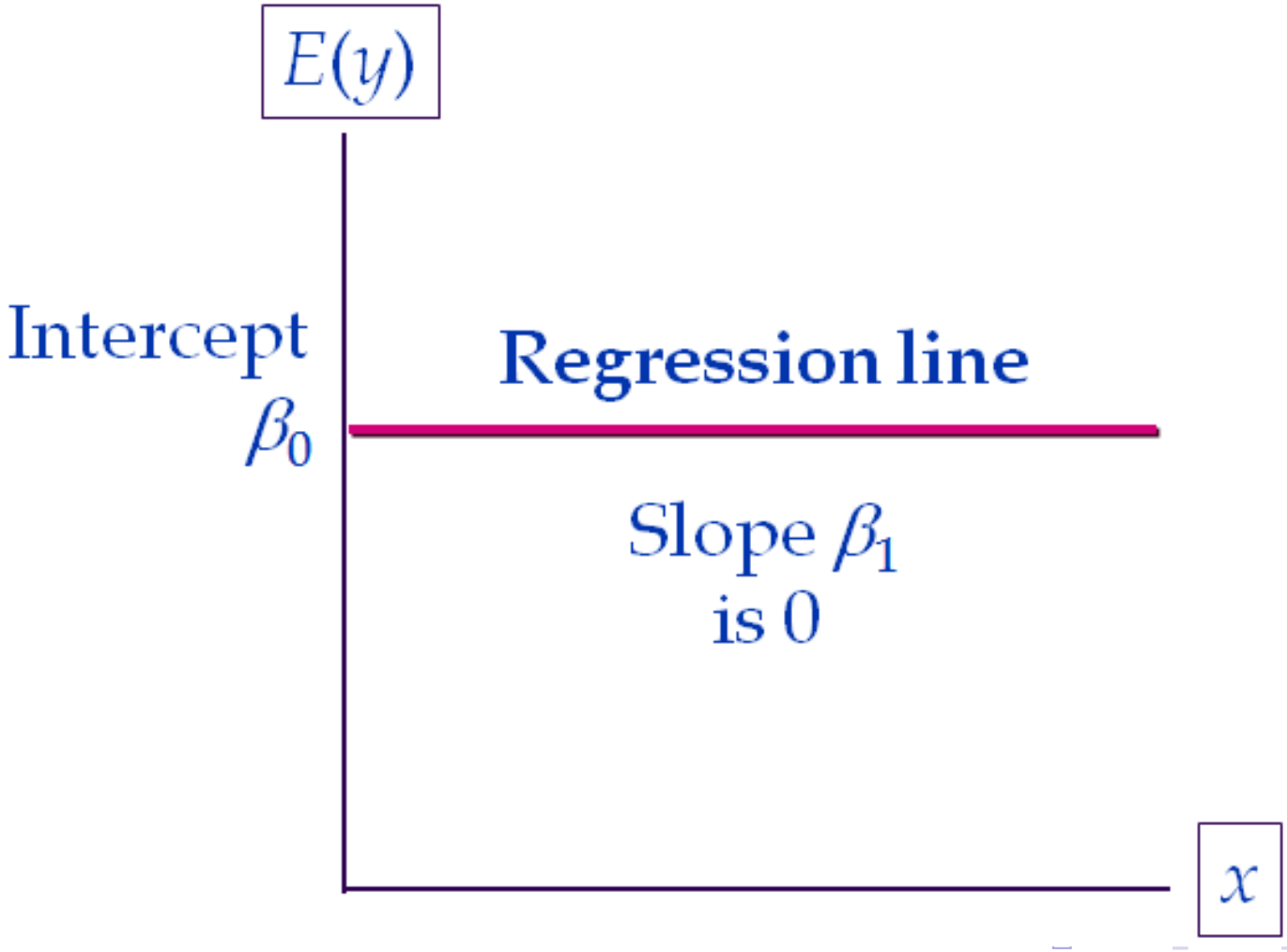
Simple Linear Regression Equation

- Negative Linear Relationship



Simple Linear Regression Equation

- Positive Linear Relationship



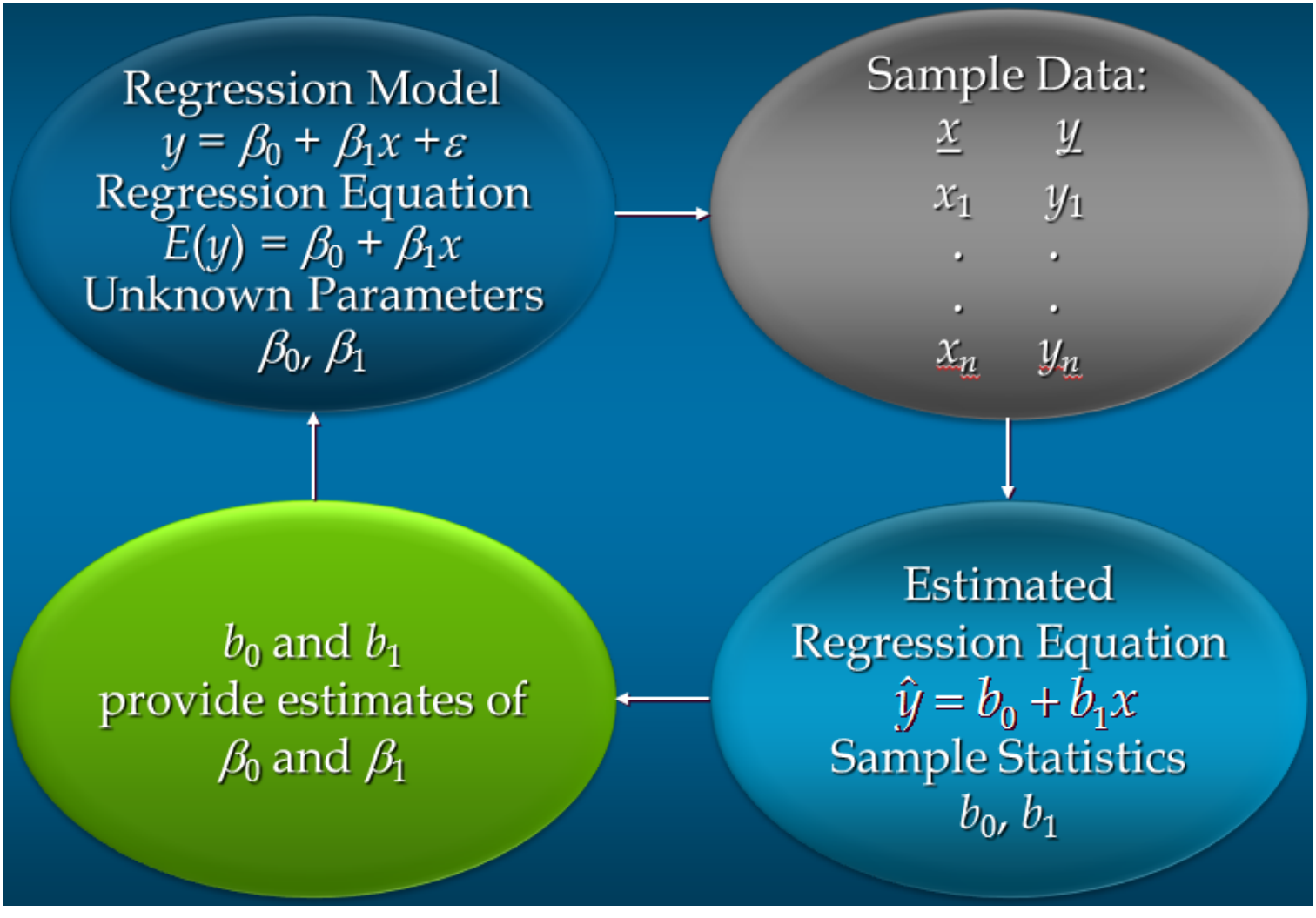
Estimated Simple Linear Regression Equation

- The estimated simple linear regression equation:

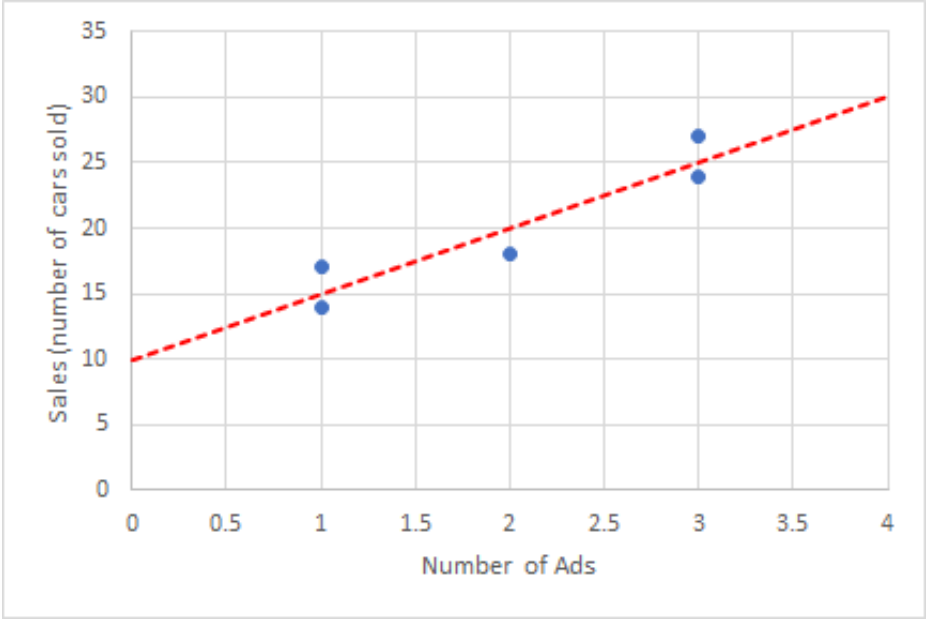
$$\hat{Y} = b_0 + b_1 X$$

- The graph is called the estimated regression line.
- b_0 is the y intercept of the line
- b_1 is the slope of the line.
- $\hat{Y}$ is the estimated value of Y for a given X value

Estimation Process



Example, cont'd



- We want to find the intercept and the slope of the best linear line above to fit that relation. That is, we want to find b_0, b_1 of the estimated regression equation: $\hat{Y} = b_0 + b_1 X$
- A method to find b_0, b_1 is to use Least Squares Method.

Outline

Least Squares Criterion

$$\min_{b_0,b_1} \sum (Y_i - \hat{Y}_i)^2$$

where

- $Y_i =$ observed value of the dependent variable for the i^{th} observation
- $\hat{Y}_i =$ estimated value of the dependent variable for the i^{th} observation

Slope and Intercept for the Estimated Regression Equation

Slope for the Estimated Regression Equation:

$$b_1 = \frac{\sum(X_i - \bar{X}_i)(Y_i - \bar{Y}_i)}{\sum(X_i - \bar{X}_i)^2}$$

where

- X_i = value of the independent variable for i^{th} observation
- Y_i = value of the dependent variable for i^{th} observation
- $\bar{X}$ = (sample) mean value for independent variable
- $\bar{Y}$ = (sample) mean value for dependent variable

Intercept for the Estimated Regression Equation:

$$b_0 = \bar{Y} - b_1\bar{X}$$

Back to our example

Reed Auto periodically has a special week-long sale. As part of the advertising campaign Reed runs one or more television commercials during the weekend preceding the sale. Data from a sample of 5 previous sales are shown on the next slide.

Number of TV Ads X	Number of Cars Sold Y
1	14
3	24
2	18
1	17
3	27
$\sum X = 10$	$\sum Y = 100$
$\bar{X} = 2$	$\bar{Y} = 20$

Estimated Regression Equation

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{20}{4} = 5$$

- y-intercept for the Estimated Regression Equation:

$$b_0 = \bar{Y} - b_1\bar{X} = 20 - 5(2) = 10$$

- Estimated Regression Equation

$$\hat{Y} = 10 + 5X$$

Interpret the slope b_1

- If $b_1 > 0$, we say: An increase by 1 unit of X leads to an increase by b_1 unit of Y.
- If $b_1 < 0$, we say: An increase by 1 unit of X leads to a decrease by b_1 unit of Y.

E.g. Continue the advertising example:
The estimated $b_1 = 5$ means that as we conduct one more advertising campaign, we can sell 5 more cars.

Explanation Power

We want to see the **goodness of fit** of the linear regression in explaining the relation between the two variables.

$$\begin{aligned} SST &= SSR + SSE \\ \sum (Y_i - \bar{Y})^2 &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \end{aligned}$$

where

- $SST =$ total sum of squares
- $SSR =$ sum of squares due to regression
- $SSE =$ sum of squares due to error

The coefficient of determination, a measure for the goodness of fit is

$$R^2 = \frac{SSR}{SST} \tag{1}$$

Explanation Power. Interpret R^2

Continue the advertising example.

$$R^2 = \frac{SSR}{SST} = \frac{100}{114} = 0.8772 \tag{2}$$

- The regression relationship is very strong; 87.72% of the variability in the number of cars sold can be explained by the linear relationship between the number of TV ads and the number of cars sold.
- OR, the linear model here can explain 87.72% of the variability in the number of cars sold.

Assumptions about the Error Term ϵ

In order to the estimated slope and intercept of the regression equation are consistent (getting close to the true slope and intercept of the model as the sample size increases), we need:

1. The error ϵ is a random variable with mean of zero.
2. The variance of ϵ, denoted by σ^2, is the same for all values of the independent variable.
3. The values of ϵ are independent.
4. The error ϵ is a normally distributed random variable.

Hypothesis Testing for the Estimated Coefficient

- Often, we want to test whether there does exist the relationship, i.e. whether the slope β_1 is significantly different from 0.
- This test is called testing for significance.
- One popular test, we learn in this course, is t **Test**.
- Formally, we are going test the following hypothesis.

Testing Hypothesis

$$H_0 : \beta_1 = 0 \textbf{ vs. } H_a : \beta_1 \neq 0 \tag{3}$$

Testing for Significance

1. Estimate the variance of the error term ϵ:

$$s^2 = \frac{SSE}{n-2} \tag{4}$$

2. Estimate the standard deviation of the error term ϵ:

$$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}} \tag{5}$$

3. Calculate the standard error for the estimated coefficient b_1:

$$s_{b_1} = \frac{s}{\sqrt{\sum (X_i - \bar{X})^2}} \tag{6}$$

4. Calculate $t-$test statistic:

$$t = \frac{b_1}{s_{b_1}} \tag{7}$$

Decision Rule

Decition Rule

Reject H_0 if p-value $\leq \alpha$, or $t \leq -t_{\alpha/2}$, or $t \geq t_{\alpha/2}$
where $t_{\alpha/2}$ is based on a t distribution with $n-2$ degrees of freedom.

Eg. If we want to test for the significance coefficient hypothesis with the level of significance $\alpha = 5\%$, then
we reject H_0 if p-value ≤ 0.05 or $|t| \geq 3.182$ with 3 degrees of freedom.

- If we reject H_0, this means that the slope of the relationship is significantly different from 0.