Simple Linear Regression

Sara Esfahani

Economics 30330

Spring 2017

Introduction

- How to better understand the relationship between two or more variables using regression analysis
- Let's say we want to study the relationship between advertising and sales
- Usually we would expect to see as we spend more on advertising, we sell more
  - But by how much?
- We would like to develop a model to show how the variables are related and to predict sales for a given level of advertising for example

Introduction

- In regression we define
  - A dependent variable (y)
    - Sales – what we are trying to predict
  - An independent variable (x)
    - Advertising expenditures – what we use to predict sales

What is a simple linear regression?

- The word “simple” refers to the number of independent variables in the model
  - A simple linear regression only has one independent variable and one dependent variable (one x and one y)
  - If we have more than one independent variable, we call the model a multiple regression
- The term “Linear” refers to the relationship that will be approximated using a straight line

Simple Linear Regression Model

- Regression line

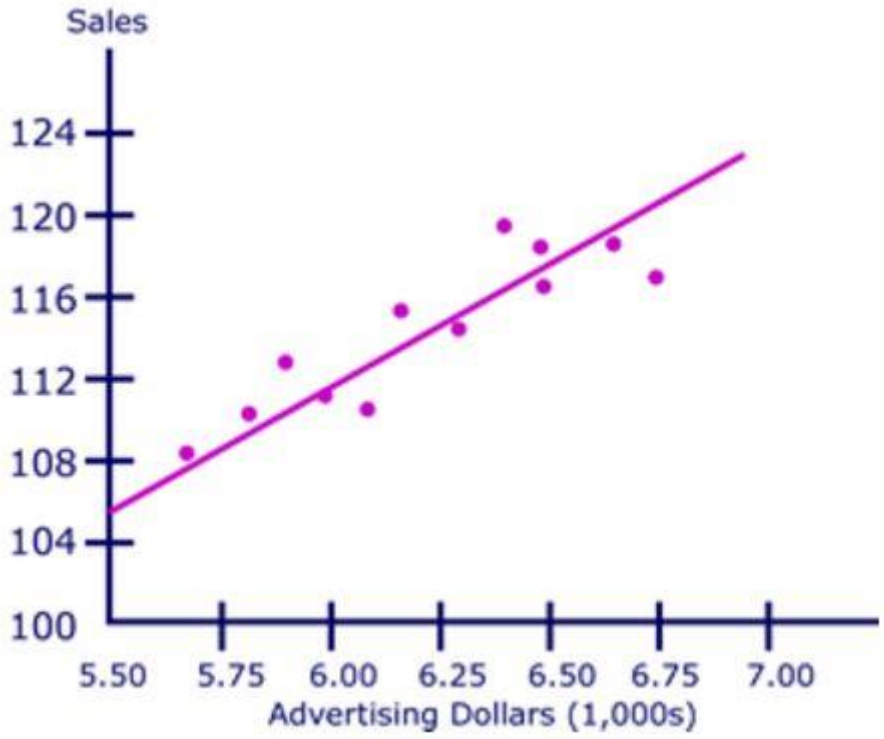
$$y = \beta_0 + \beta_1 x + \epsilon$$

- β_0 is the y-intercept of the regression line
- β_1 is the slope of the regression line
  - Increasing or decreasing
  - How steep?
- ϵ is the error term

Simple Linear Regression Model

- Regression line

$$y = \beta_0 + \beta_1 x + \epsilon$$



Simple Linear Regression Model

- Regression line

$$y = \beta_0 + \beta_1 x + \epsilon$$

- β_0 and β_1 are the population parameters
- In estimated simple linear regression equation, b_0 and b_1 are the sample statistics used to estimate β_0 and β_1

$$\hat{y} = b_0 + b_1 x$$

  - $\hat{y}$ (y-hat) is the estimated/predicted value of y for a given value of x
  - b_0 is the y intercept of the line
  - b_1 is the slope of the line

Example: Data

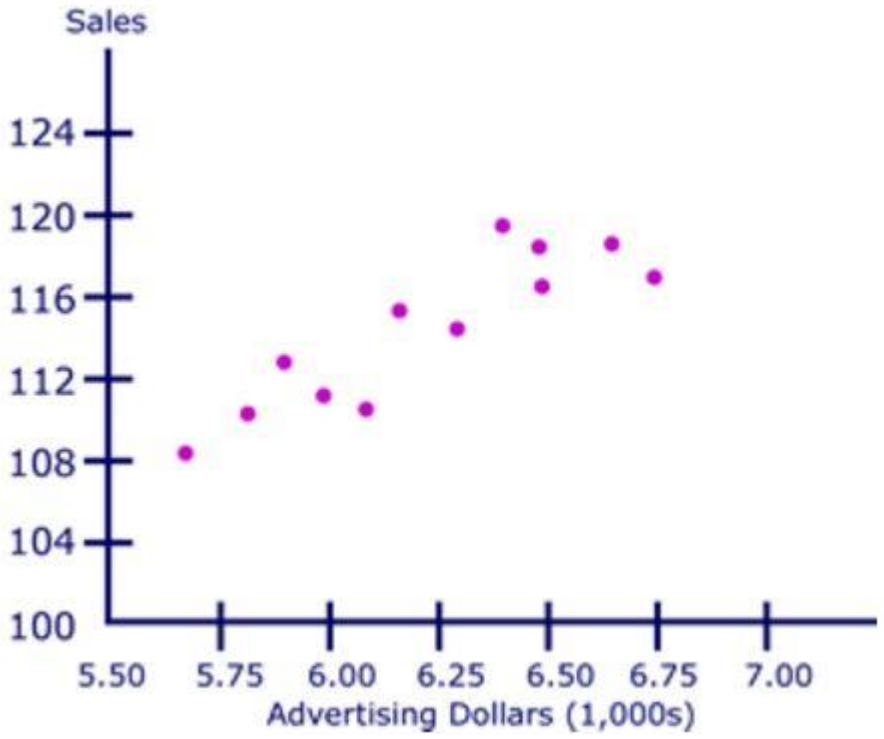
- The following table lists the monthly sales and advertising expenditures for all of last year by a digital electronics company.

Month	Sales (in 1000s)	Advertising Dollars (1000s)
January	100	5.5
February	110	5.8
March	112	6
April	115	5.9
May	117	6.2
June	116	6.3
July	118	6.5
August	120	6.6
September	121	6.4
October	120	6.5
November	117	6.7
December	123	6.8

Example: Scatter Plot

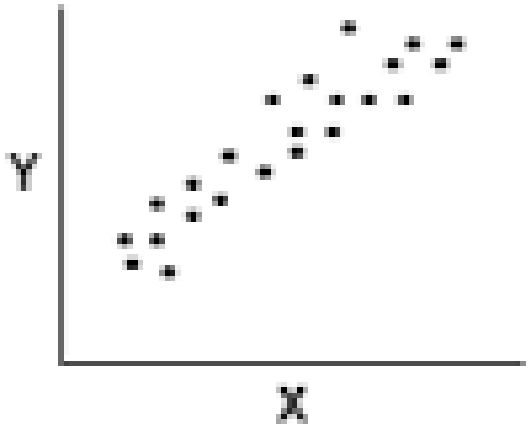
- In this case, you would plot last year's data for monthly sales and advertising expenditures as shown on the scatter plot below. (Data for independent and dependent variables must be from the same period of time.)

- Scatter plots are effective in visually identifying relationships between variables.



Scatter Diagram

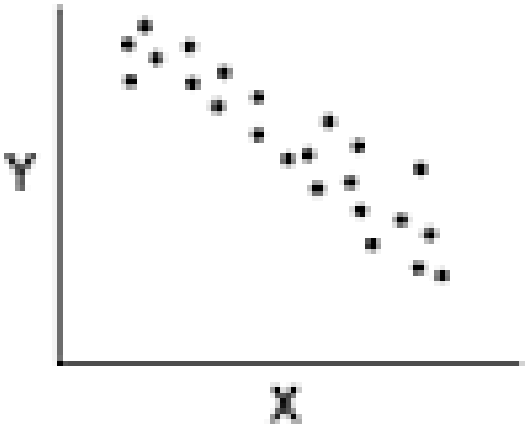
(A)



Positive linear relationship between x and y

For an increase in x, there is corresponding increase in y

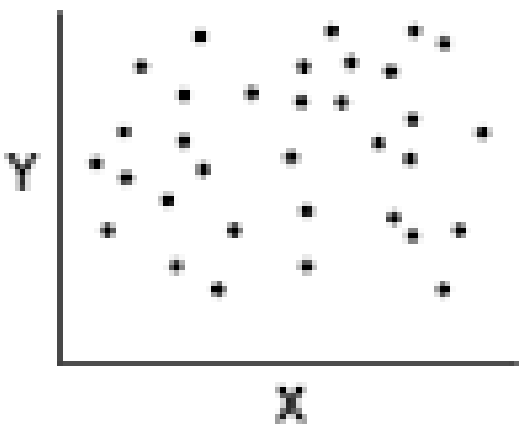
(B)



Negative linear relationship between x and y

For an increase in x, there is corresponding decrease in y

(C)



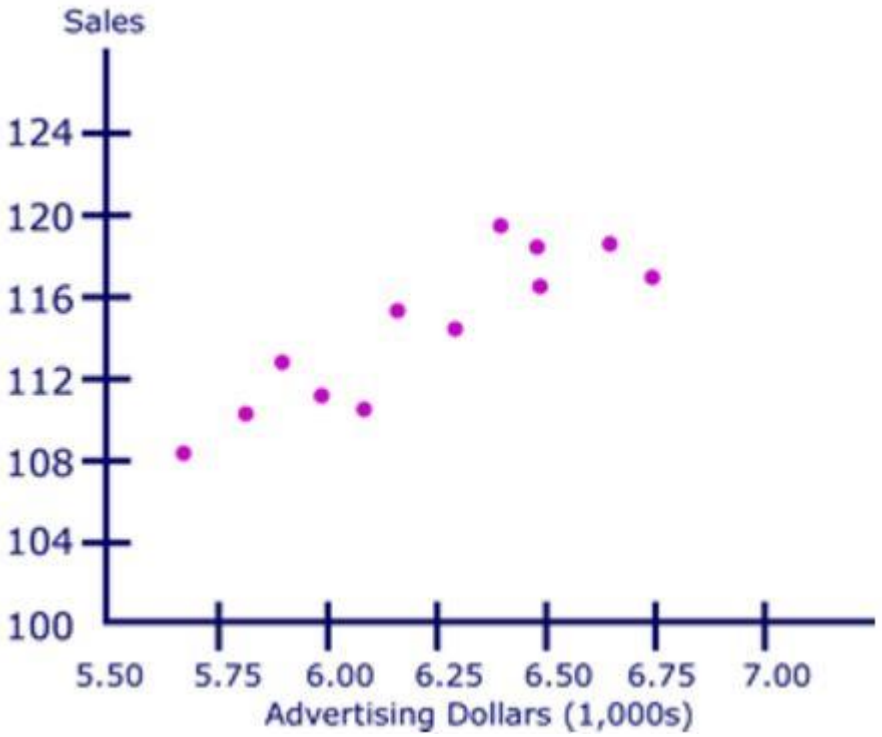
No linear relationship between x and y

For an increase in x, there is no corresponding reaction in y

Example: Scatter Plot

- These relationships can be expressed mathematically in terms of a **correlation coefficient**, which is commonly referred to as a correlation.

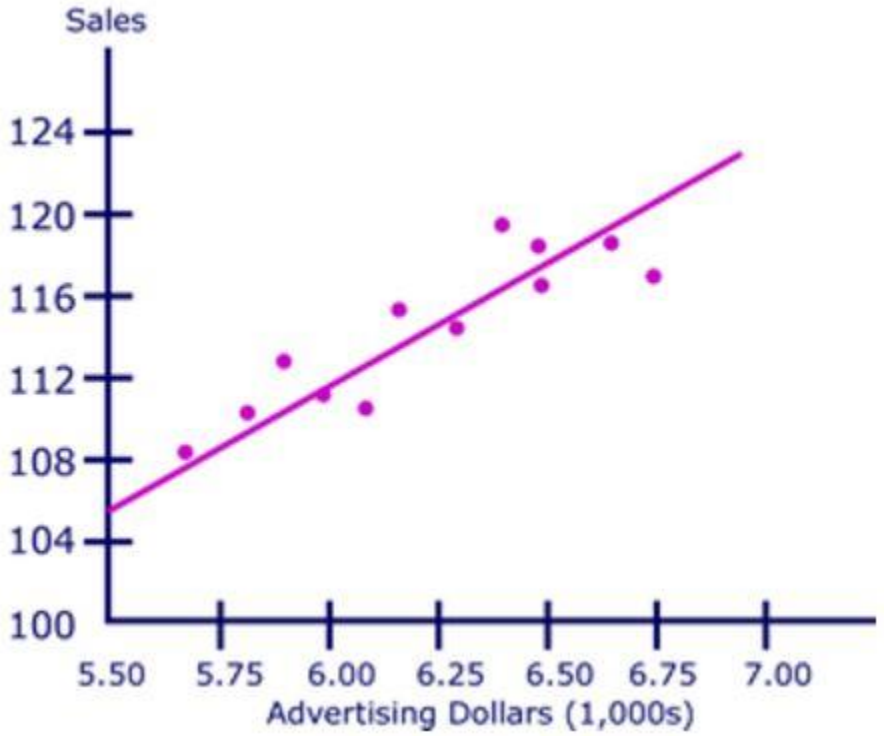
- Correlations are indices of the **strength of the relationship** between two variables. They can be any value from –1 to +1.



- Viewing the scatter plot above, you can see that there appears to be some degree of correlation between the level of advertising expenditure and product awareness.

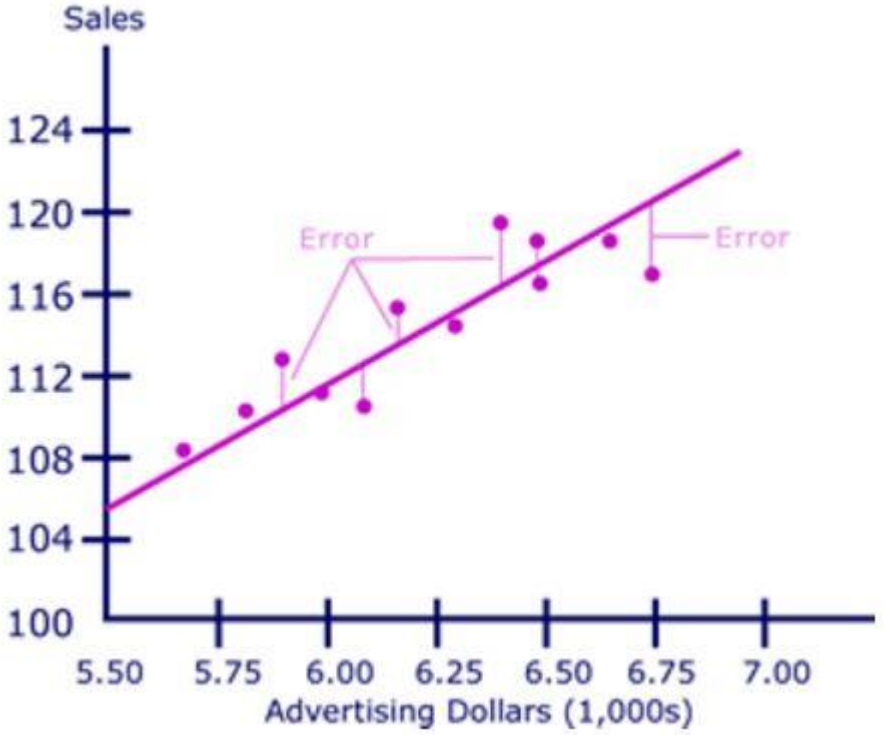
Example: Regression Line

- This figure is the same as the scatter plot on the previous slide, with the addition of a **regression line** fitted to the historical data.
- The regression line is the line with the **smallest possible set of distances** between itself and each data point.
- The distances of the data points from the regression line are called **error terms**



Example: Error Terms

- A regression line will always contain error terms because, in reality, independent variables are never perfect predictors of the dependent variables.
- There are many uncontrollable factors in the business world.
- Some predictive capacity will always be absent, particularly in simple regression.



How do we find the fitted line?

- The typical procedure for finding the line of best fit is called the **least-squares method**.

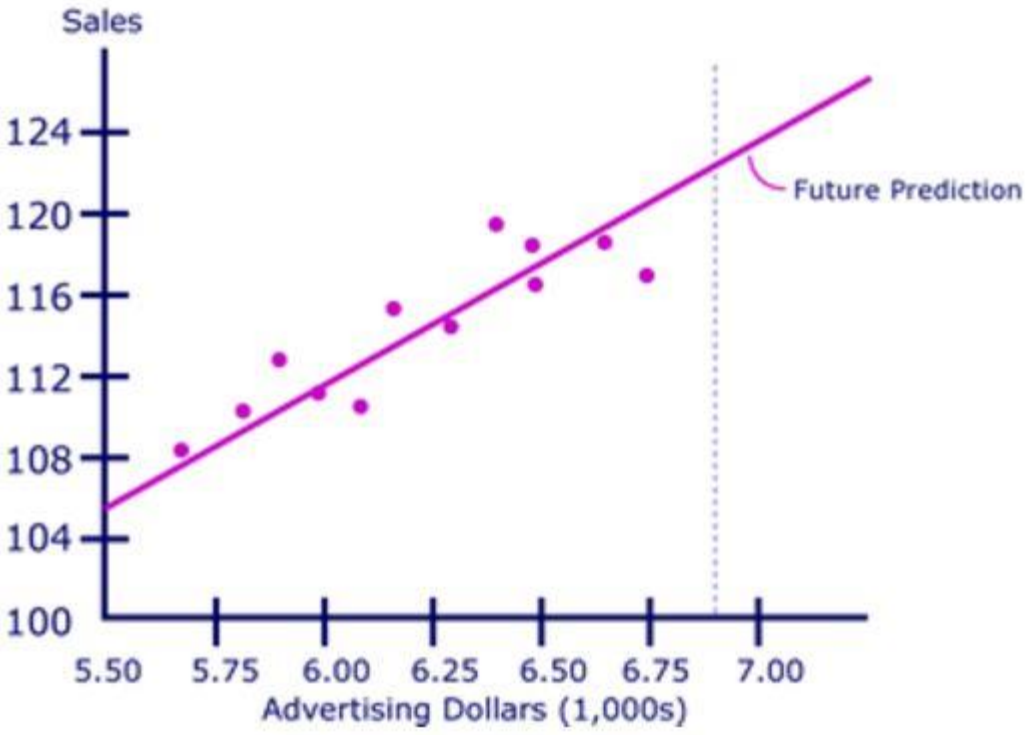
- The least-squares method is based upon the principle that the sum of the squared errors should be made as small as possible so the regression line has the least error.

- This calculation is usually performed using computer software.

Example: Prediction

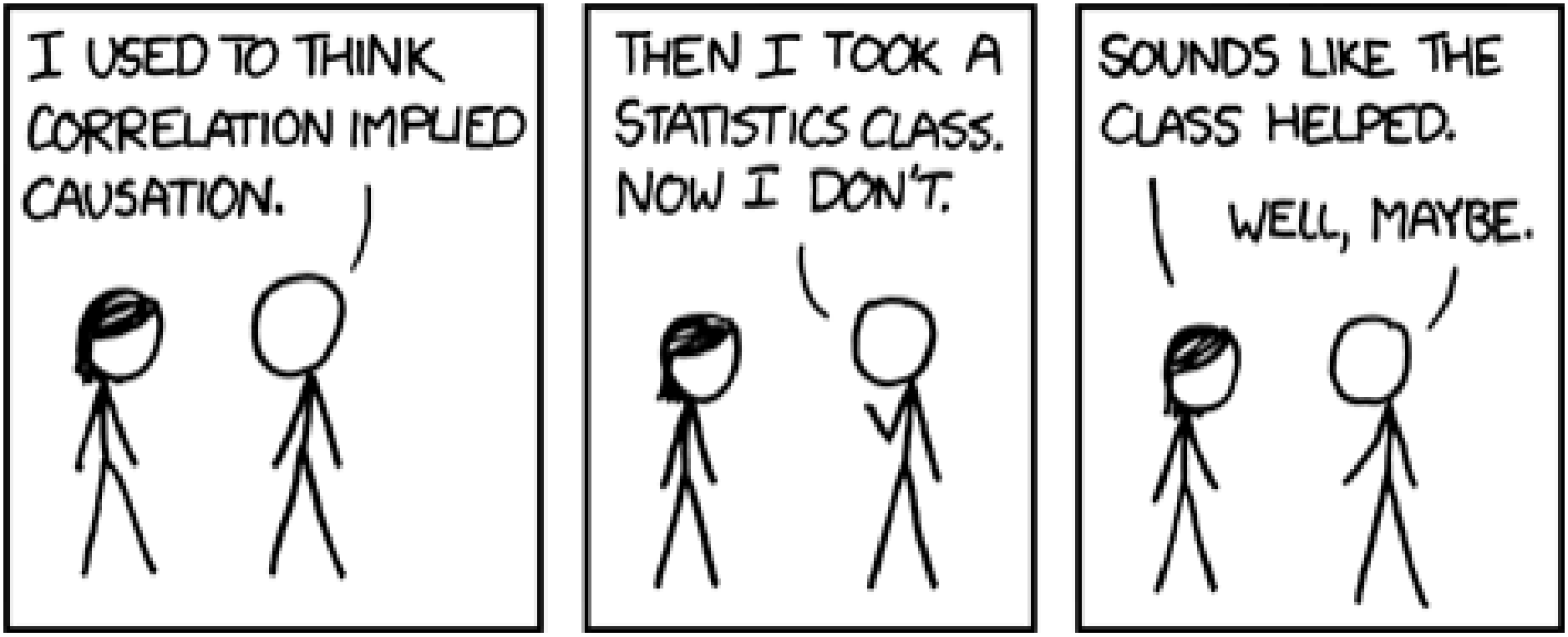
- Once the regression line is determined, it can be extended beyond the historical data to **predict future levels** of product awareness, given a particular level of advertising expenditure.

- The extension of the line of regression requires the <u>assumption</u> that the underlying process causing the relationship between the two variables is valid beyond the range of the sample data.



Regression Analysis

- When you run a regression in Excel or in a statistics program, the program will provide you with a report.
- We will discuss some of the terms in this report in our class. But, most of the details of these reports, and the definition of all the terms included in the report, are some of the concepts you'll learn in econometrics class.



Source: xkcd.com

Correlation vs. Causation

- Before we get to estimation of the fitted line to data, it is important for us to differentiate between a correlational and a causal relationship

Correlation vs. Causation

Regression Analysis

- Generate a model that helps measure the impact of x on y
  - The model allows us to measure the strength of the statistical association between two variables
  - This is an accurate estimate ONLY IF we have measured the causal relationship
- In most cases, authors want to estimate the **causal impact** of 'x on y'
  - In very few cases do they satisfy the necessary assumptions (you'll talk about this more in econometrics)

Equation of Regression Line

- You may recall the equation of a straight line from your algebra class:

$$f(x) = mx + b$$

Where x represents the independent variable

f(x) is the dependent variable

the constant b denotes the y-intercept

and the coefficient m is the slope (movement in dependent variable as a results of a movement in independent variable)

- Similarly, for the equation of the regression line we have:

$$\hat{y} = b_0 + b_1 x$$

Least Squares Method

- As it was pointed out, the typical procedure for finding the line of best fit is called the **least-squares method**
- The least-squares method is based upon the principle that the sum of the squared errors should be made as small as possible so the regression line has the least error

$$\min_{b_0,b_1} \sum_i (y_i - \hat{y}_i)^2$$

Where y_i is the observed value of the dependent variable for the *i*th observation ($y = \beta_0 + \beta_1 x + \epsilon$)

And $\hat{y}_i$ is the predicted value of the dependent variable for the *i*th observation ($\hat{y} = b_0 + b_1 x$)

Notation

- True model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - We observe data points (y_i, x_i)
  - The parameters β_0 and β_1 are unknown
  - Because we do not know β_0 and β_1, the actual error (ϵ_i) is unknown

- We can however estimate values of ϵ_i by estimating values of β_0 and β_1

- Estimated model: $\hat{y}_i = b_0 + b_1 x_i + e_i$
  - (b_0, b_1) are estimates for the parameters (β_0, β_1)
  - e_i is an estimate of ϵ_i where $e_i = y_i - b_0 - b_1 x_i$
  - e_i is the **residual**

- How do you estimate b_0 and b_1?

Solution to Minimization Problem

- By minimizing the sum of squared residuals (SSR), the slope and y-intercept for the estimated regression equation are

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Where

x_i = value of the independent variable for the *i*th observation

y_i = value of the dependent variable for the *i*th observation

$\bar{x}$ = mean value for the independent variable

$\bar{y}$ = mean value for the dependent variable

Notes

- The sum of squared residuals (SSR) is also called the sum of squared errors (SSE)

- Why do we square the residuals?
  - To treat positive and negative errors equally
  - Over or under predict by some value is the same magnitude of error
  - Functions reach min or max values when derivatives are zero

Example

Month	Sales (in 1000s)	Advertising Dollars (1000s)	y-ybar	x-xbar	(y-ybar)(x-xbar)	(x-xbar)^2
Jan	100	5.5	-15.75	-0.77	12.08	0.59
Feb	110	5.8	-5.75	-0.47	2.68	0.22
Mar	112	6	-3.75	-0.27	1.00	0.07
Apr	115	5.9	-0.75	-0.37	0.28	0.13
May	117	6.2	1.25	-0.07	-0.08	0.00
Jun	116	6.3	0.25	0.03	0.01	0.00
Jul	118	6.5	2.25	0.23	0.53	0.05
Aug	120	6.6	4.25	0.33	1.42	0.11
Sep	121	6.4	5.25	0.13	0.70	0.02
Oct	120	6.5	4.25	0.23	0.99	0.05
Nov	117	6.7	1.25	0.43	0.54	0.19
Dec	123	6.8	7.25	0.53	3.87	0.28
Sum	1389	75.2			24.00	1.73
Means	115.75	6.27				

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{24}{1.73} = 13.90$$

$$b_0 = \bar{y} - b_1\bar{x} = 115.75 - 13.90 \times 6.27 = 28.65$$

Example: Equation of the fitted line

- So, we calculated

$$b_1 = 13.90$$

$$b_0 = 28.65$$

- Which can be used to write down the equation of the fitted line that minimizes the sum of squared residuals:

$$\hat{y}_i = 28.65 + 13.90\, x_i$$

Example: Interpretation

- Equation line:

$$\hat{y}_i = 28.65 + 13.90\, x_i$$

- Historical data shows that every $1000 increase in advertising expenditure leads to $13,900 increase in sales
  - That is if x goes up from say 5.8 to 6.8 (a 1 unit increase – which is in 1000s), then y goes up from 110 to 123 (see the data table)
- With no advertising expenditure, this company seems to have $28,650 in sales
  - That is if x is zero, then y is 28.65 (in 1000s)

Coefficient of Determination

- Let's define SSE as the Sum of Squares due to Error

$$SSE = \sum(y_i - \hat{y}_i)^2$$

- And SST as the Total Sum of Squares

$$SST = \sum(y_i - \bar{y})^2$$

- Then SSR is the Sum of Squares due to Regression such that

$$SST = SSR + SSE$$

and

$$SSR = \sum(\hat{y}_i - \bar{y})^2$$

- The ratio of SSR/SST is defined as the **coefficient of determination** and is denoted by r^2

Coefficient of Determination

- The coefficient of determination

$$r^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

  takes a value between 0 and 1.

- When r^2 is expressed in percentage, it can be interpreted as the percentage of the total sum of squares that can be explained by using the estimated regression equation (goodness of the fit).

- If in our example $r^2 = 0.79$, it means 79% of the variation in sales can be explained by the linear relationship between advertising expenditure and sales that we estimated.

Correlation Coefficient

- The correlation coefficient that we learned in chapter 3 is related to the coefficient of determination

$$r_{xy} = (sign\ of\ b_1)\sqrt{r^2}$$

Where r^2 is the coefficient of determination

And r_{xy} is the correlation coefficient

- From our sales and advertising expenditure example we get $r_{xy} = 0.89$
- Generally a correlation coefficient above 0.5 in absolute terms indicates a strong linear relationship.

The End

(of Linear Regression)

What We've Learned

Part 1: Basics of Organizing Information

- Data basics
- Graphical representation
- Descriptive statistics
- Basic probability theory

Part 2: Generalizing Results

- Discrete probability distributions
- Continuous probability distributions
- Sampling distributions

Part 3: (Dis)proving Statements

- Confidence intervals
- Hypothesis testing
- Importance of statistics and probability

"What I should have learned in ECON 30330"

- Discrete random variables
- Continuous random variables
- PDFs and CDFs
- Discrete distributions
- Continuous distributions
- Properties of expectations
- Estimation
- Variance
- Standard normal distribution
- Normal distribution
- Conditional probabilities
- Independent events
- Covariance
- Correlation coefficient
- Linear combination of random variables
- Variance of the sample mean
- Central Limit Theorem
- Testing hypotheses about a sample mean or proportion
- Testing for equality of means/proportions across two samples
- Confidence intervals
- Simple linear regression

Listen to Hal

“I keep saying that the sexy job in the next 10 years will be statisticians,” said Hal Varian, chief economist at Google. “And I’m not kidding.”

– *New York Times*, 8/6/2009

"Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.”

– Hal Varian, January 2009

- Thanks everyone! It’s been a great semester!