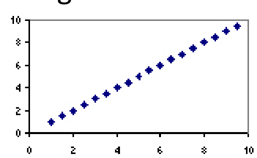
STAT 206: **3.5 (Coefficient of Correlation) and Chapter 13 (Simple Linear Regression)**

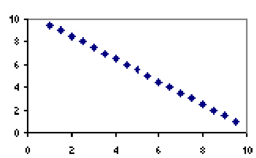
- **Ideas in 3.5 and Chapter 13**
  - Exploring the Association between Two Quantitative Variables
  - Association versus Causation
  - Regression and Least Square Regression
  - Calculating values using Excel

- **Three cases for exploring the association between two variables:**
  - Positive association: as values of x increase, values of y increase
  - Negative association: as values of x increase, values of y decrease
  - No association: values of x do not affect the values of y

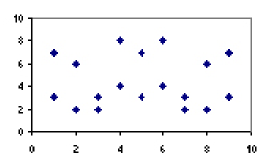
- **If a linear pattern is present in the scatterplot, calculate the correlation, denoted by r, to measure the strength and direction of the LINEAR relationship between x and y.**
  - Positive values of r indicate a positive relationship between the variables
  - Negative values of r indicate a negative relationship between the variables
  - r ranges from -1 to 1. The closer r is to 0, the weaker the relationship. The closer r is to 1 or -1, the stronger the relationship



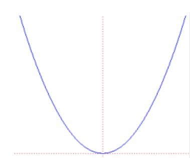
Maximum positive correlation
(r = 1.0)



Maximum negative correlation
(r = -1.0)



Zero correlation
(r = 0)



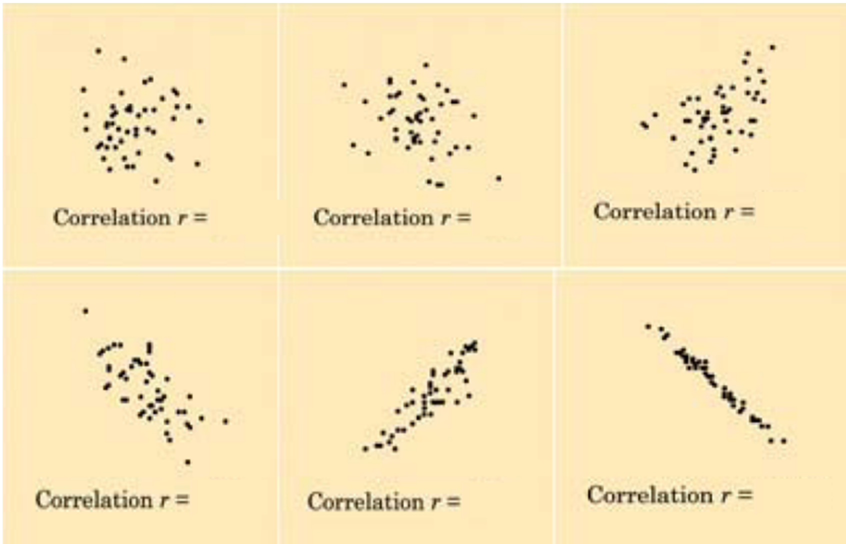
- **Calculating r (correlation coefficient)**

Calculating the Correlation r

$$r = \frac{1}{n-1}\Sigma z_x z_y = \frac{1}{n-1}\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

where n is the number of points, $\bar{x}$ and $\bar{y}$ are means, and s_x and s_y are standard deviations for x and y. The sum is taken over all n observations.

- **Correlation Examples:**



- **Features of the Coefficient of Correlation**
  - The **population coefficient of correlation** is referred as ***ρ*** (rho) and the **sample coefficient of correlation** is referred to as ***r***
  - Either ***ρ*** or ***r*** have the following features:
    - Range:
    - The _______ to –1, the stronger the __________ linear relationship
    - The ______ to 1, the stronger the ___________ linear relationship
    - The closer to 0, the _______ the linear relationship
  - **correlation does not imply causation – if two variables are associated with each other, it does not necessarily mean that changes in one variable cause changes in the other variable**
  - Examples on Google……
- **The Coefficient of Correlation Using Microsoft Excel Function**

The Coefficient of Correlation Using Microsoft Excel Function

Test #1 Score	Test #2 Score		Correlation Coefficient	
78	82		0.7332	=CORREL(A2:A11,B2:B11)
92	88			
86	91			
83	90			
95	92			
85	85			
91	89			
76	81			
88	96			
79	77			

The Coefficient of Correlation Using Microsoft Excel Data Analysis Tool

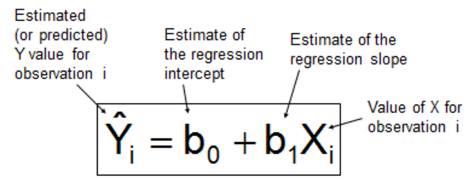
- **Introduction to Regression Analysis**
  - Regression analysis is used to:
    -
    -
  - Dependent variable:
  - Independent variable:

- **How can we predict the outcome of a Variable?**
  - **The regression line predicts the value for the response variable y as a straight line function of the value of x of the explanatory variable**
  - $\hat{y} = b_0 + b_1 x$
    - $\hat{y}$ **is**
    - b_0 **is**
    - b_1 **is**
  - **The slope is**
  - **y-intercept is**
  - **Linear Regression is appropriate when:**
    1.

    2.

- **Simple Linear Regression Model**
  - Only one independent variable, X
  - Relationship between X and Y is described by a linear function
  - Changes in Y are assumed to be related to changes in X

- **Simple Linear Regression Equation (Prediction Line)**
  - The simple linear regression equation provides an estimate of the population regression line



- **The Least Squares Method**
  - *b_0* and *b_1* are obtained by finding the values of that minimize the sum of the squared differences between Y and $\hat{Y}$:

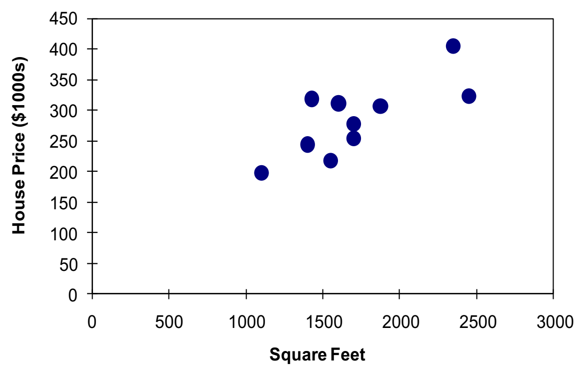
$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

  - The coefficients b_0 and b_1, and other regression results in this chapter, will be found using Excel

  - b_0 is the estimated average value of Y when the value of X is zero

  - b_1 is the estimated change in the average value of Y as a result of a one-unit increase in X

- **Simple Linear Regression Example:** A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
  - A random sample of 10 houses is selected
  - Independent variable (X) = square feet
  - Dependent variable (Y) = house price in $1000s



• House Price in $1000s	• Square Feet
• (Y)	• (X)
• 245	• 1400
• 312	• 1600
• 279	• 1700
• 308	• 1875
• 199	• 1100
• 219	• 1550
• 405	• 2350
• 324	• 2450
• 319	• 1425
• 255	• 1700

- **Simple Linear Regression Example: Interpretation of b_0 and b_1**
  - b_0

  - b_1

- **Making Predictions – Predict the price for a house with 2000 square feet:**

Simple Linear Regression Example: Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

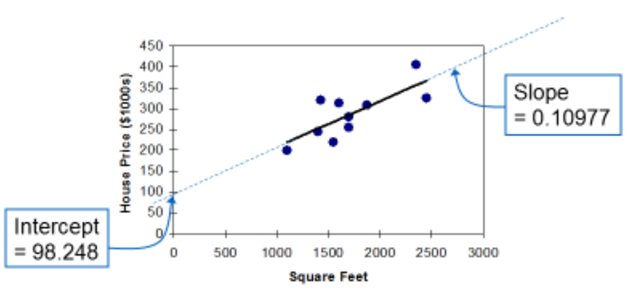
$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

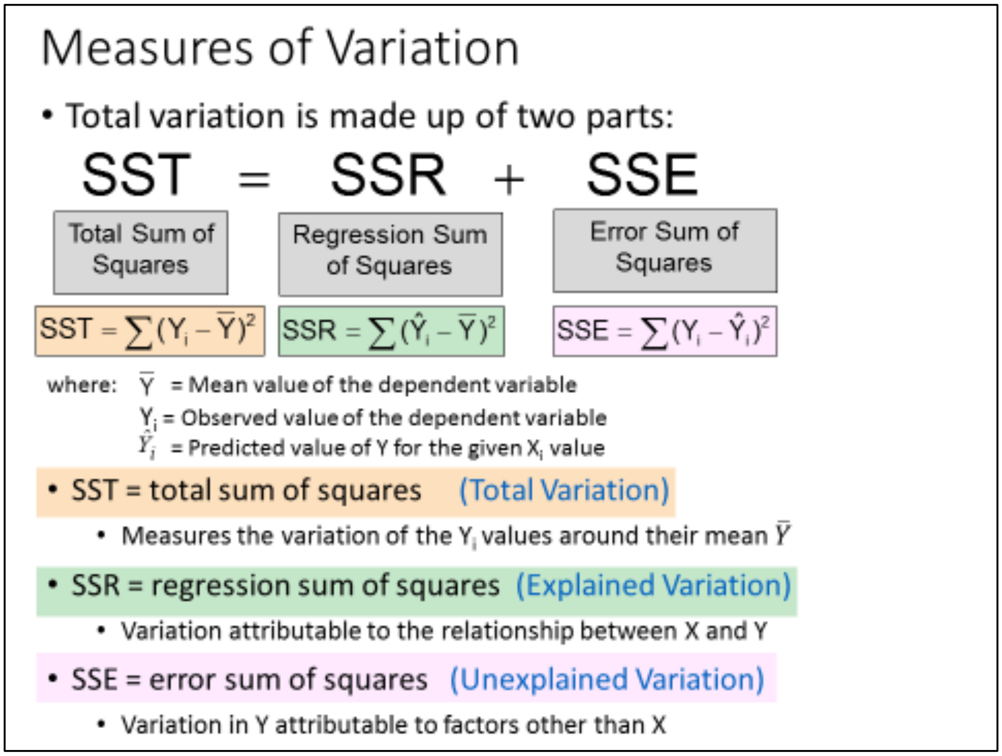
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

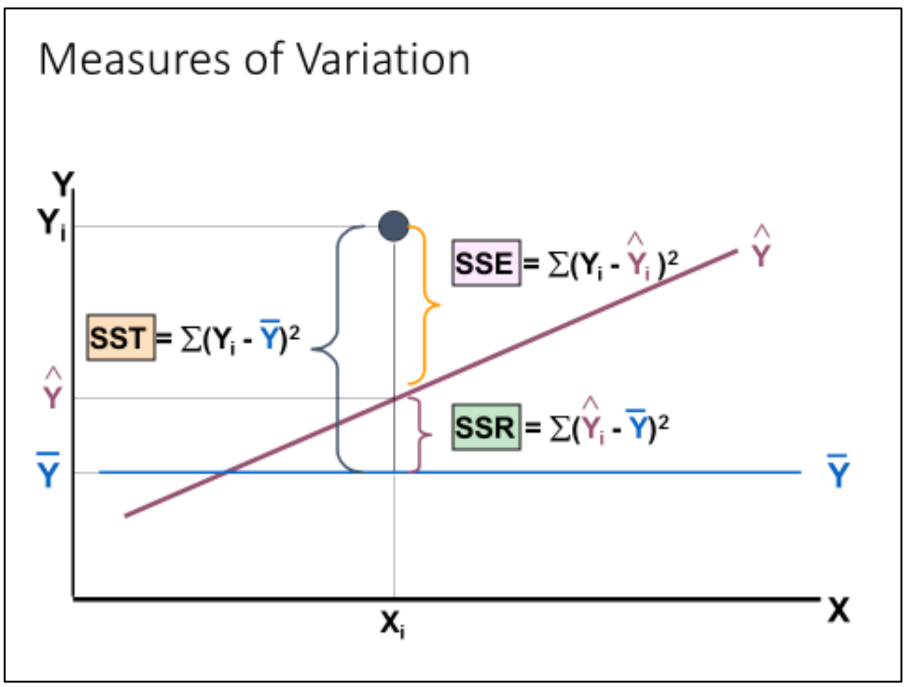
Simple Linear Regression Example: Graphical Representation

House price model: Scatter Plot and Prediction Line



$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

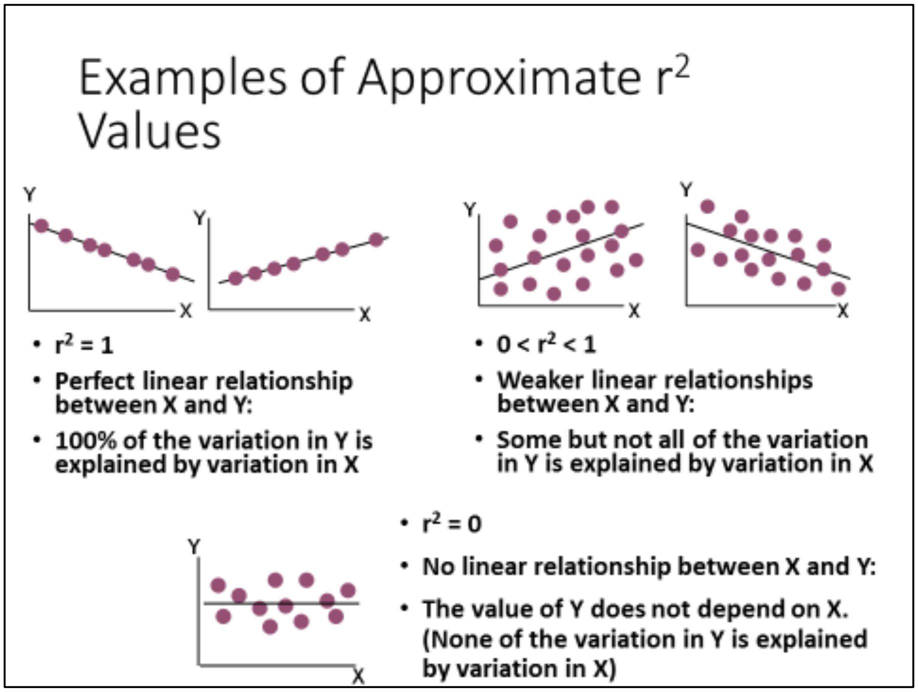




- **Coefficient of Determination, r^2**
  - The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
  - The coefficient of determination is also called r-squared and is denoted as r^2

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

  - NOTE: $0 \le r^2 \le 1$



- Simple Linear Regression Example: Coefficient of Determination, r^2 in Excel

Simple Linear Regression Example: Coefficient of Determination, r^2 in Excel

DCOVA

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580