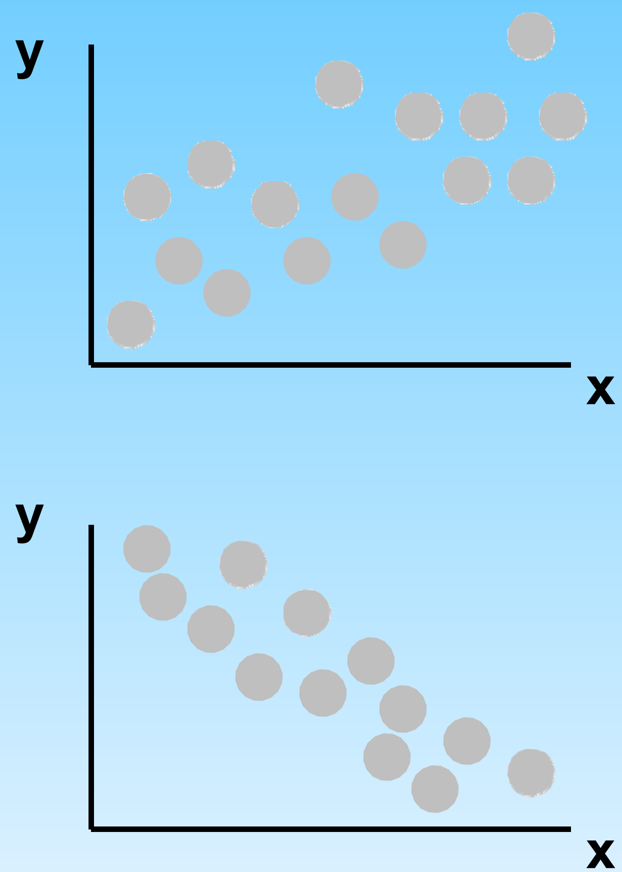
Regression and Correlation Analysis

Correlation vs. Scatter Plots

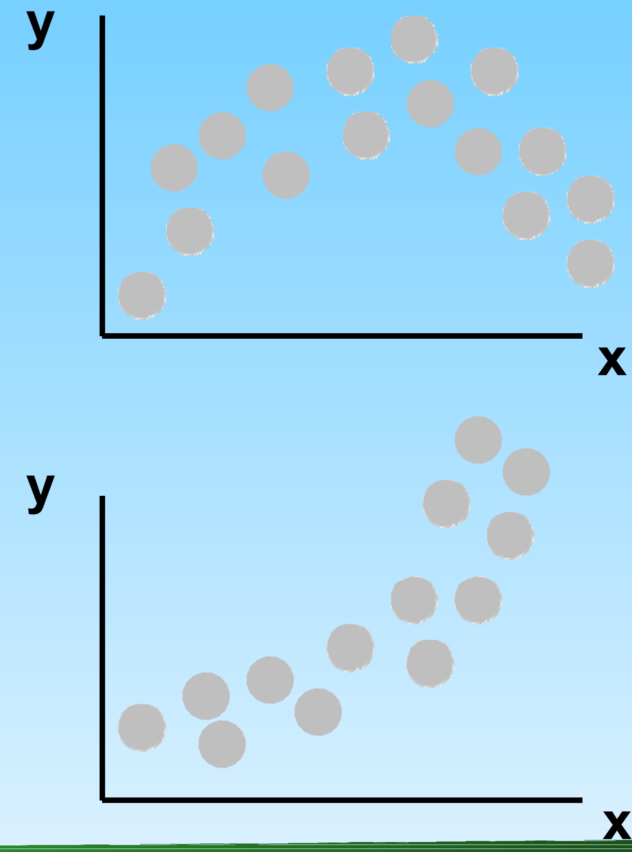
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Only concerned with strength of the relationship
  - No causal effect is implied
- A scatter plot (or scatter diagram) is used to show the relationship between two variables

Scatter Plot Examples

Linear relationships



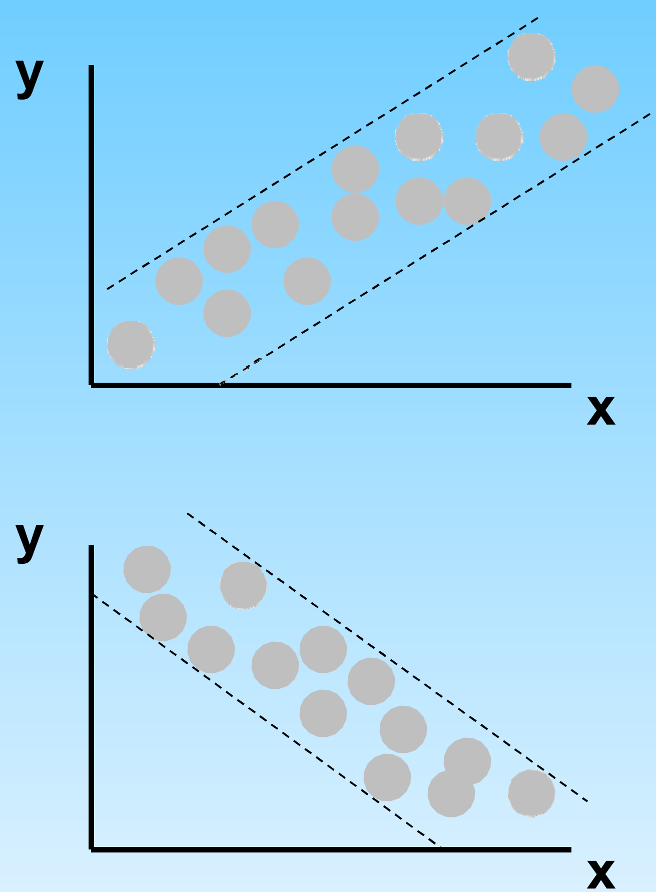
Curvilinear relationships



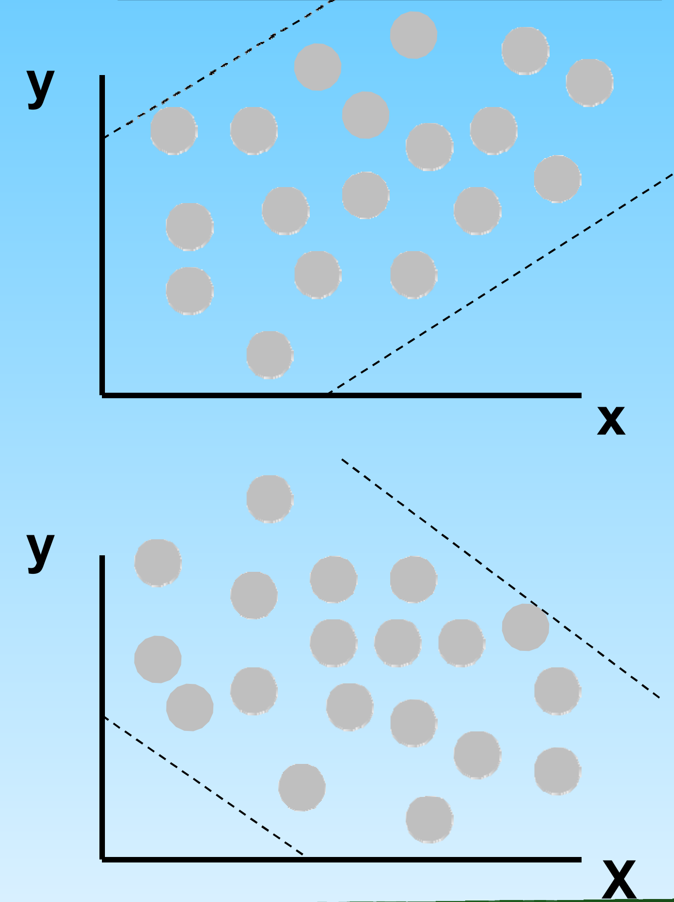
Scatter Plot Examples

(continued)

Strong relationships



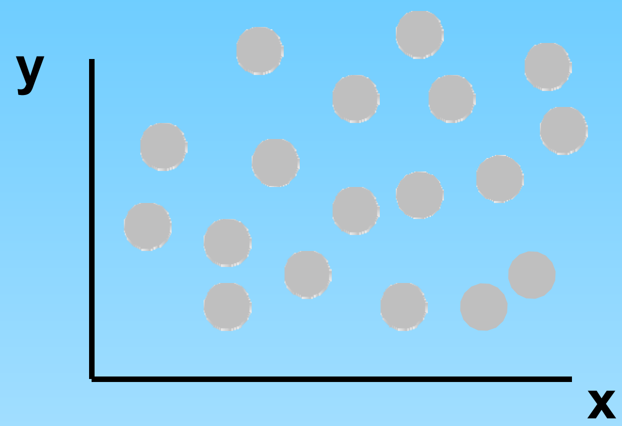
Weak relationships

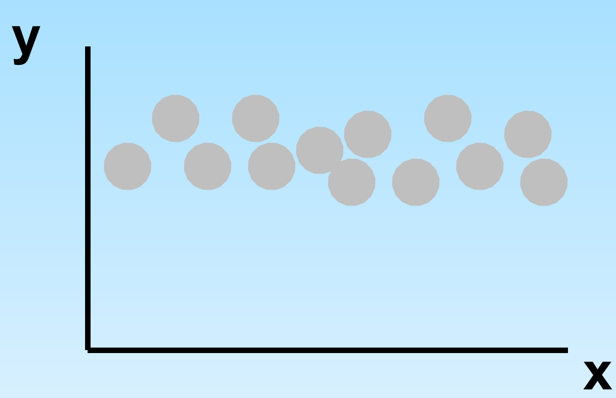


Scatter Plot Examples

(continued)

No relationship





Correlation Coefficient

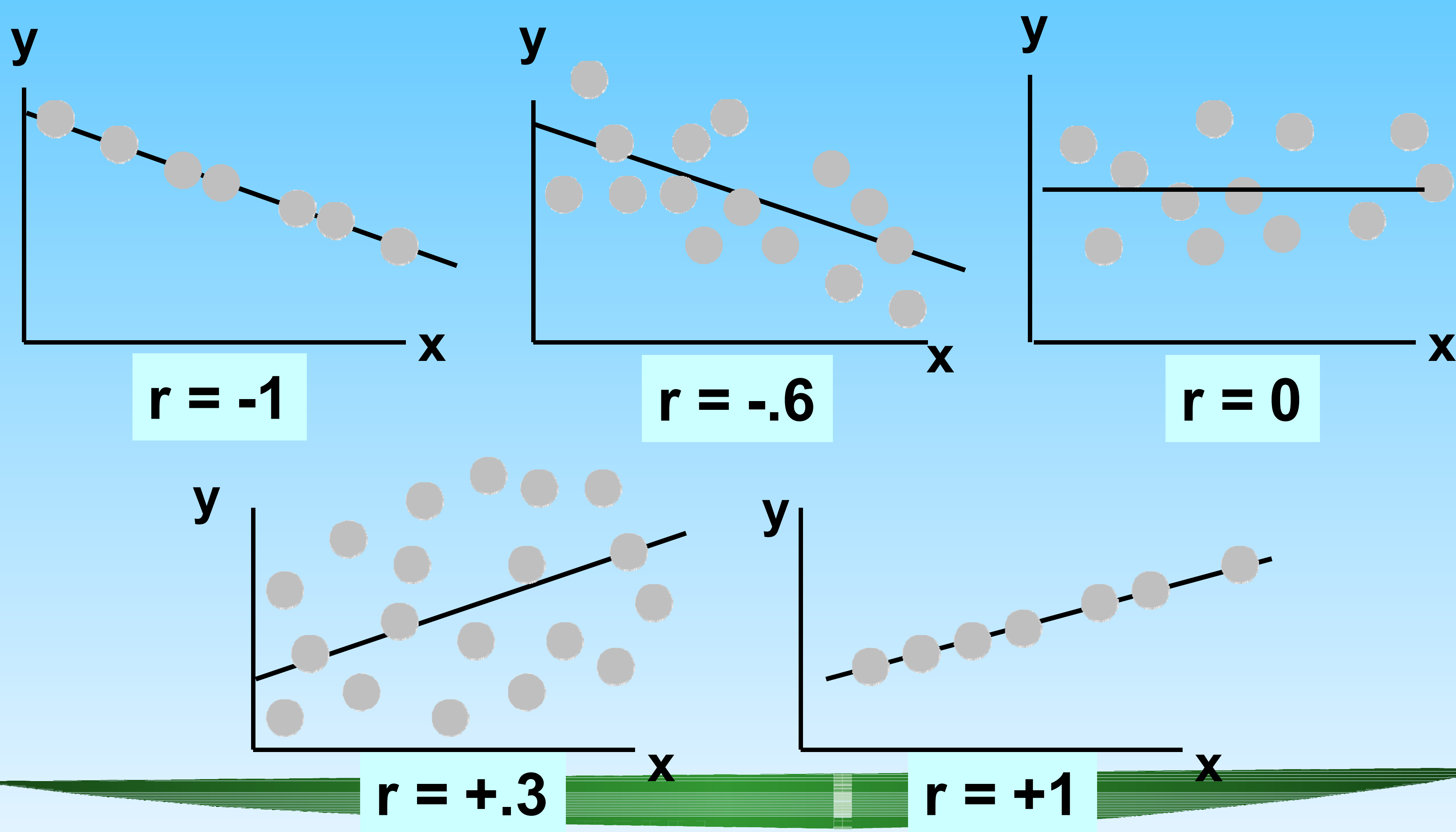
(continued)

- The population correlation coefficient ρ (rho) measures the strength of the association between the variables
- The sample correlation coefficient r is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations

Features of ρ and r

- Unit free
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

Examples of Approximate r Values



Calculating the Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

or the algebraic equivalent:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where:

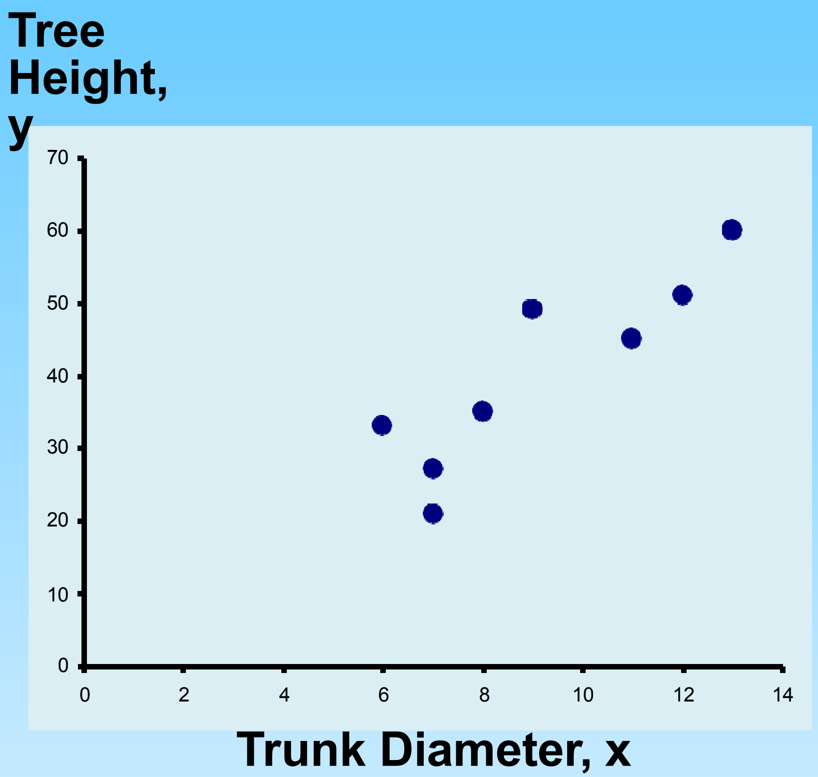
r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

Calculation Example

Tree Height	Trunk Diameter			
y	**x**	**xy**	**y^2**	**x^2**
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
Σ=**321**	Σ=**73**	Σ=**3142**	Σ=**14111**	Σ=**713**

Calculation Example

(continued)



$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$

$$= 0.886$$

r = 0.886 → relatively strong positive linear association between x and y

Excel Output

Excel Correlation Output

Tools / data analysis / correlation…

	Tree Height	Trunk Diameter
Tree Height	1	
Trunk Diameter	0.886231	1

Correlation between

Tree Height and Trunk Diameter

Significance Test for Correlation

- Hypotheses

  H_0: $\rho = 0$ (no correlation)

  H_A: $\rho \neq 0$ (correlation exists)

- Test statistic

  • $$t = \frac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$$ (with n – 2 degrees of freedom)

Example: Produce Stores

Is there evidence of a linear relationship between tree height and trunk diameter at the 0.05 level of significance?

H_0: $\rho = 0$ (No correlation)

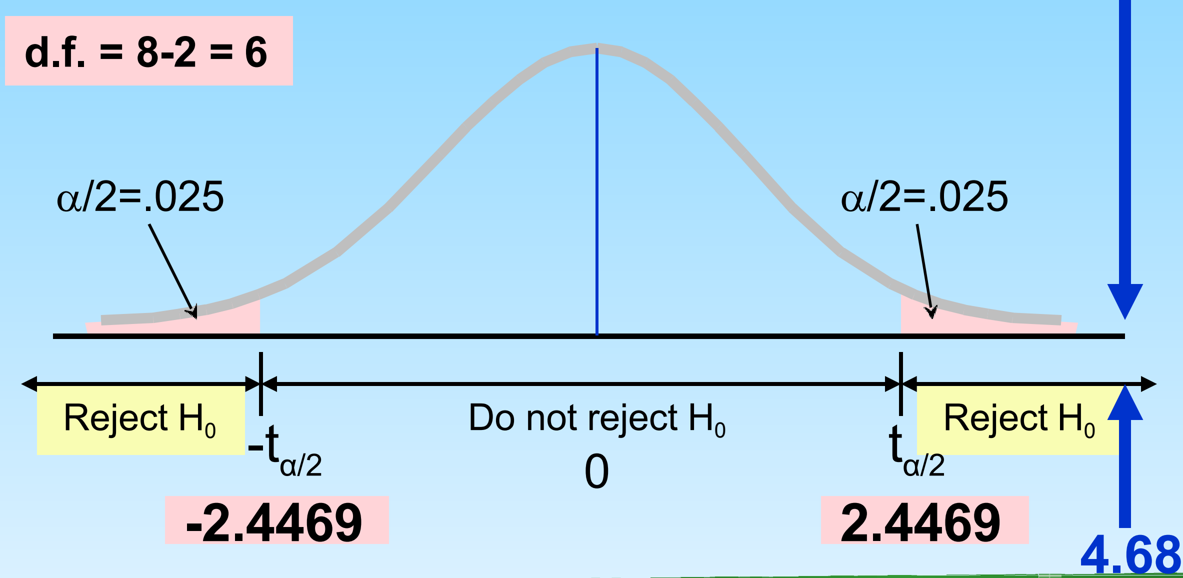
H_1: $\rho \neq 0$ (correlation exists)

$\alpha = .05$, df = 8 - 2 = 6

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{0.886}{\sqrt{\frac{1 - 0.886^2}{8 - 2}}} = 4.68$$

Example: Test Solution

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$



Decision:

Reject H_0

Conclusion:

There **is evidence** of a linear relationship at the 5% level of significance

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain

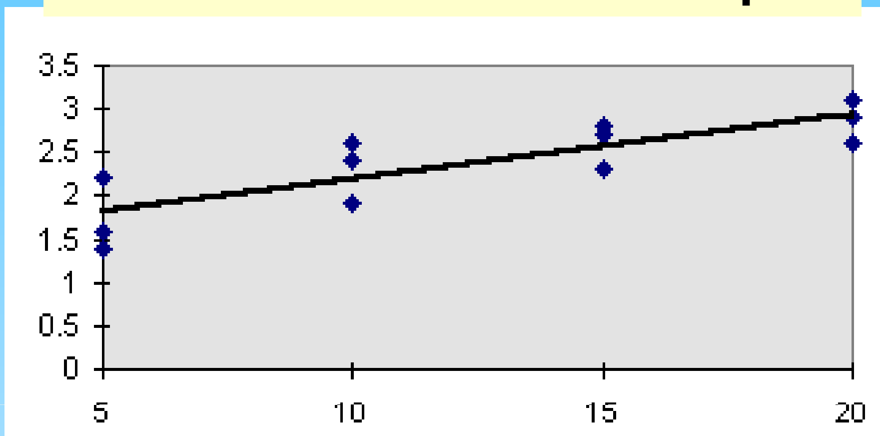
Independent variable: the variable used to explain the dependent variable

Simple Linear Regression Model

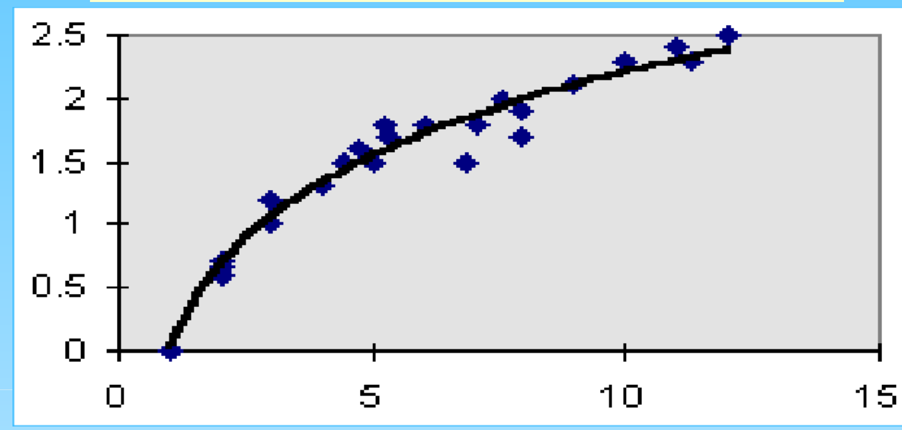
- Only **one** independent variable, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be caused by changes in x

Types of Regression Models

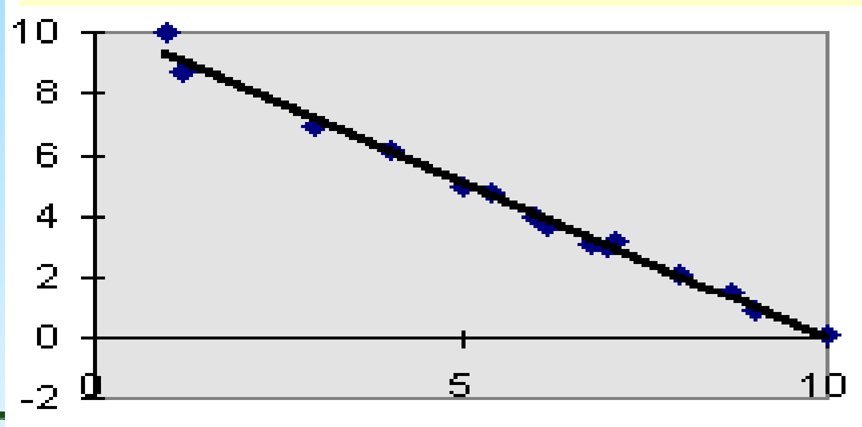
Positive Linear Relationship



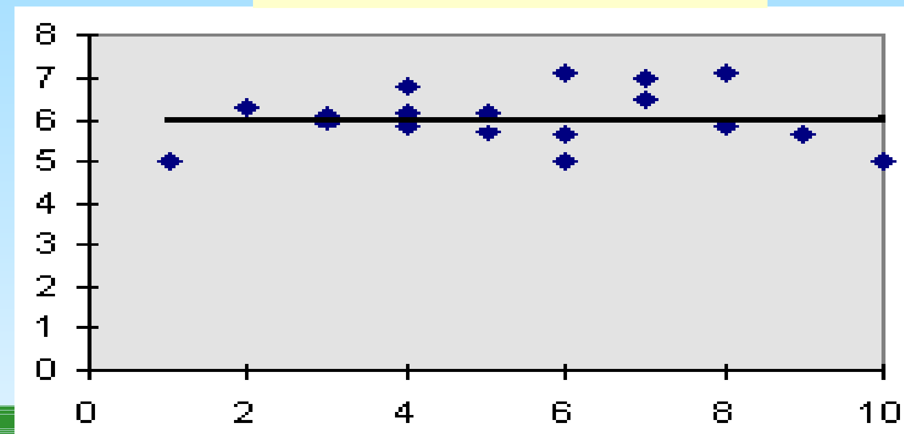
Relationship NOT Linear



Negative Linear Relationship



No Relationship



Population Linear Regression

The population regression model:

Dependent Variable

Population y intercept

Population Slope Coefficient

Independent Variable

Random Error term, or residual

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{Linear component}} + \underbrace{\varepsilon}_{\text{Random Error component}}$$

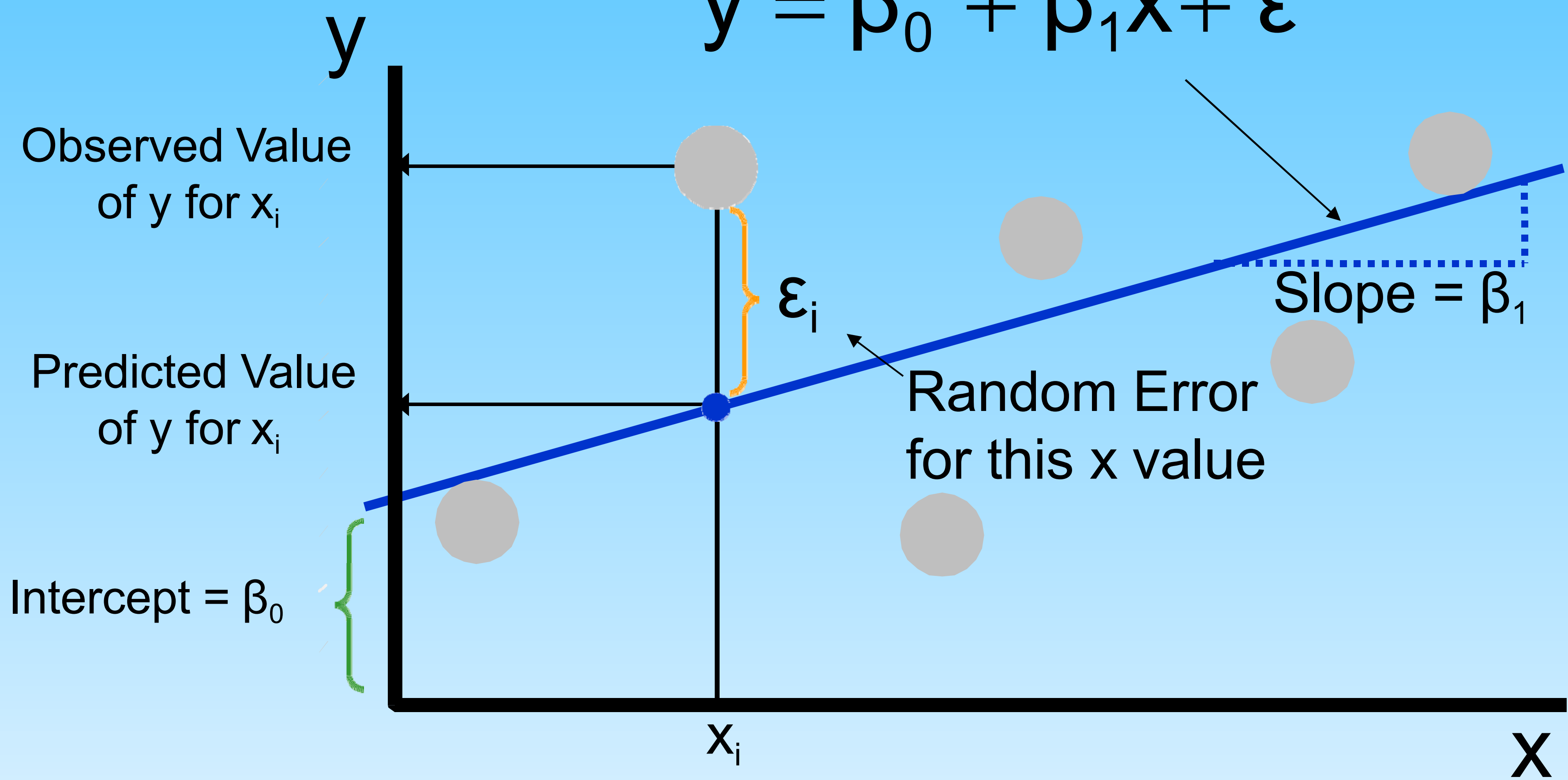
Linear Regression Assumptions

- Error values (ε) are statistically independent
- Error values are normally distributed for any given value of x
- The probability distribution of the errors is normal
- The probability distribution of the errors has constant variance
- The underlying relationship between the x variable and the y variable is linear

Population Linear Regression

(continued)

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



Estimated Regression Model

The sample regression line provides an estimate of the population regression line

Estimated (or predicted) y value

Estimate of the regression intercept

Estimate of the regression slope

Independent variable

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

The individual random error terms e_i have a mean of zero

Least Squares Criterion

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared residuals

$$\sum e^2 = \sum (y - \hat{y})^2$$

$$= \sum (y - (b_0 + b_1 x))^2$$

The Least Squares Equation

- The formulas for b_1 and b_0 are:

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

algebraic equivalent:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and

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

Interpretation of the Slope and the Intercept

- 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 change in x

Finding the Least Squares Equation

- The coefficients b_0 and b_1 will usually be found using computer software, such as Excel or Minitab

- Other regression measures will also be computed as part of computer-based regression analysis

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)

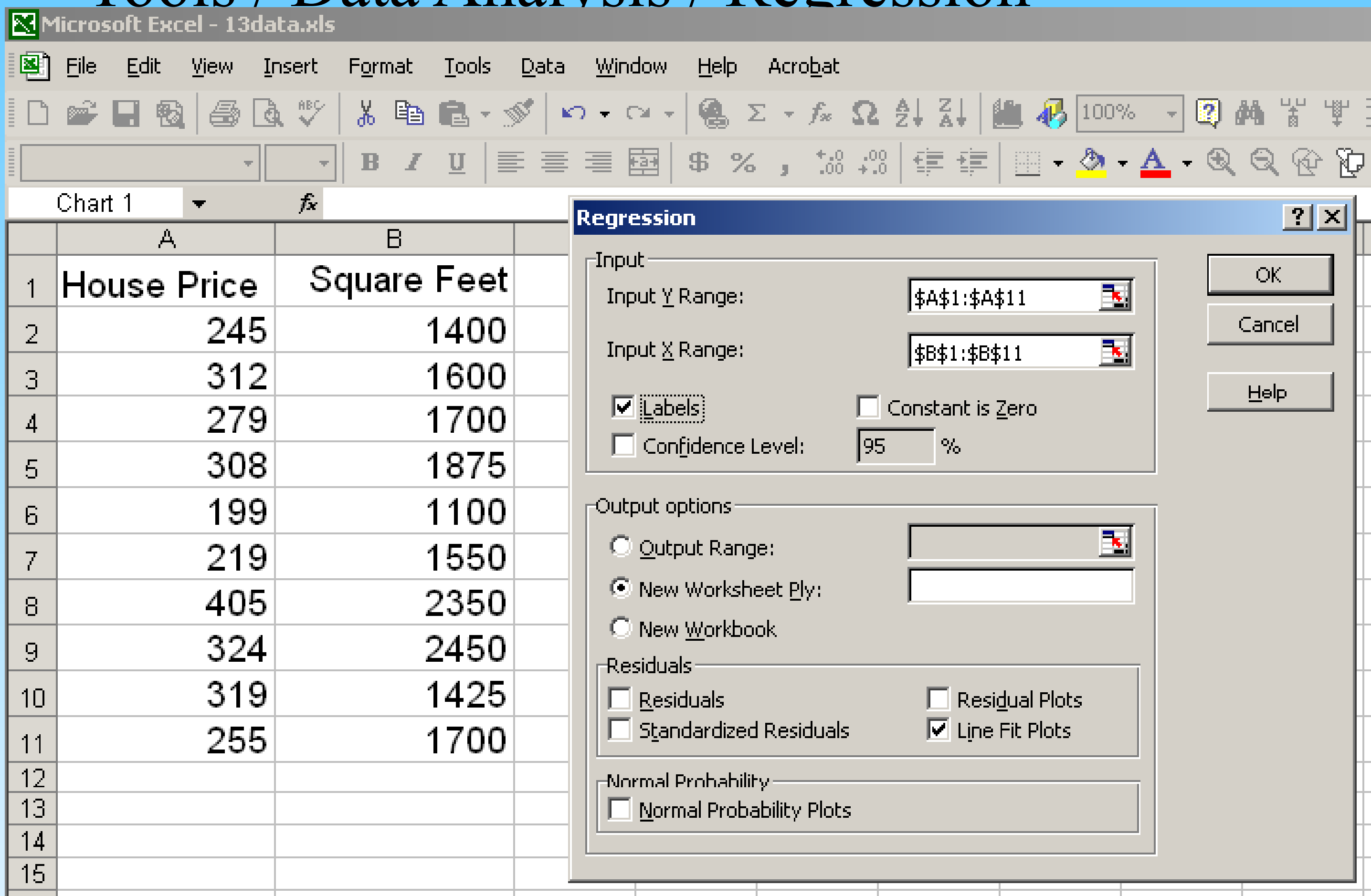
- A random sample of 10 houses is selected
  - Dependent variable (y) = house price in $1000s
  - Independent variable (x) = square feet

Sample Data for House Price Model

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

Regression Using Excel

- Tools / Data Analysis / Regression



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:

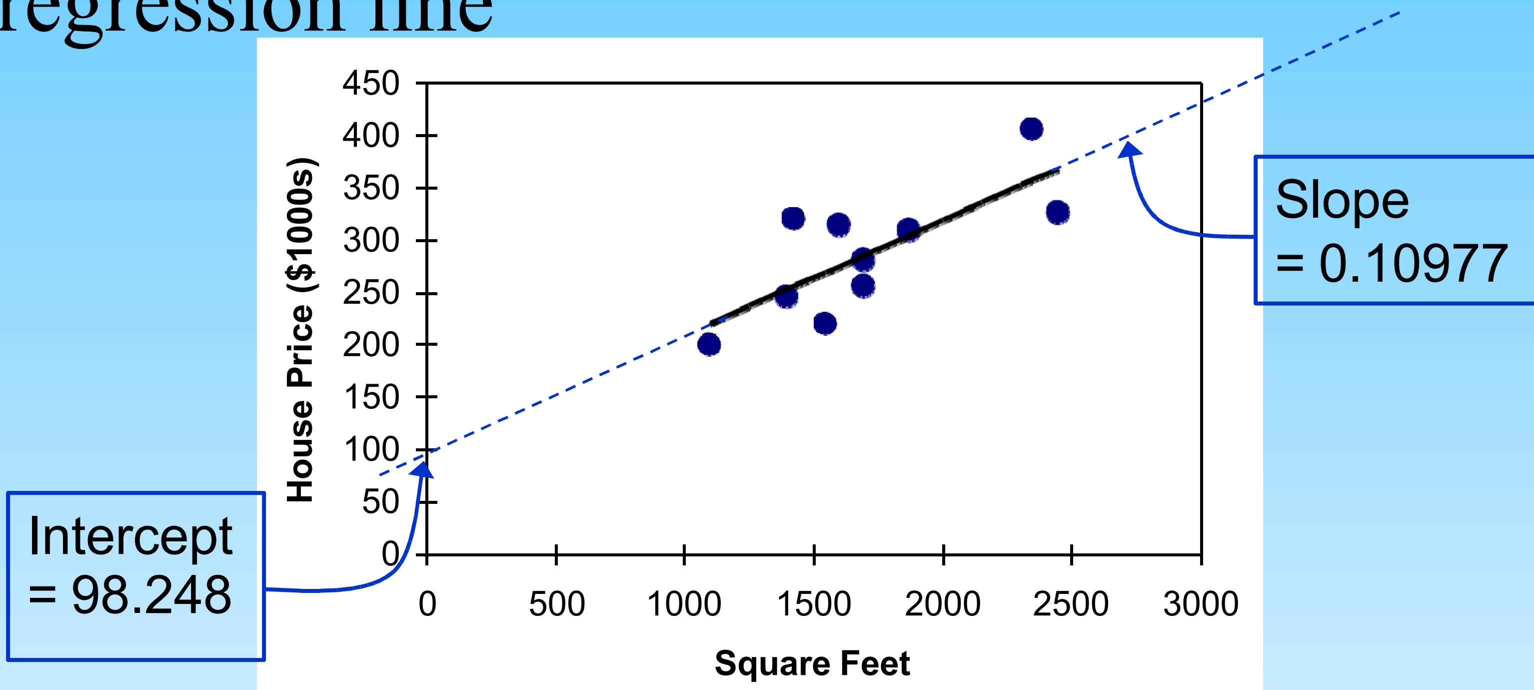
$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

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

Graphical Presentation

- House price model: scatter plot and regression line



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