Introduction to Correlation Analysis and Simple Linear Regression

Josefina V. Almeda

Stat 115

2014

Goals

After this, you should be able to:

- Calculate and interpret the simple correlation between two variables
- Determine whether the correlation is significant
- Calculate and interpret the simple linear regression equation for a set of data
- Understand the assumptions behind regression analysis
- Determine whether a regression model is significant

Goals

(continued)

After this, you should be able to:

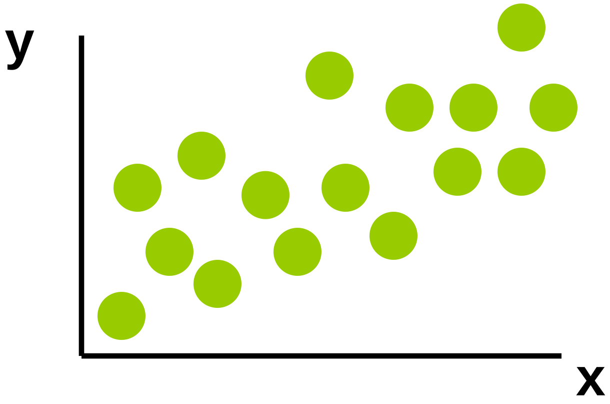
- Calculate and interpret confidence intervals for the regression coefficients
- Recognize regression analysis applications for purposes of prediction and description
- Recognize some potential problems if regression analysis is used incorrectly
- Recognize nonlinear relationships between two variables

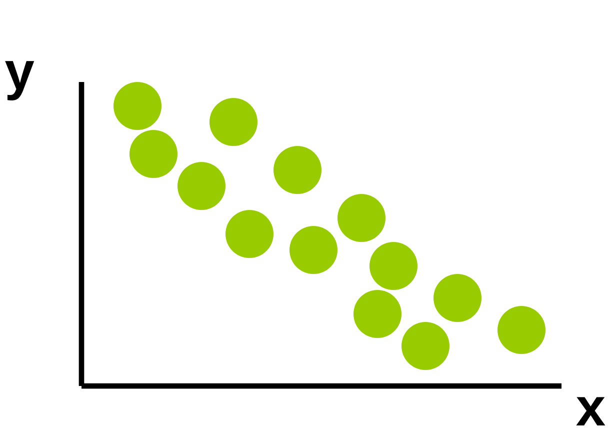
Scatter Plots and Correlation

- A scatter plot (or scatter diagram) is used to show the relationship between two variables
- 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

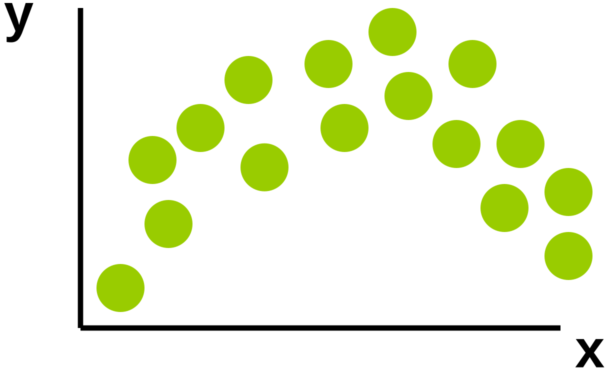
Scatter Plot Examples

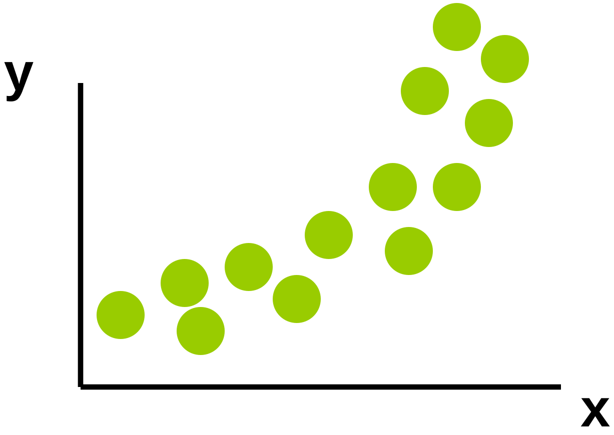
Linear relationships





Curvilinear relationships

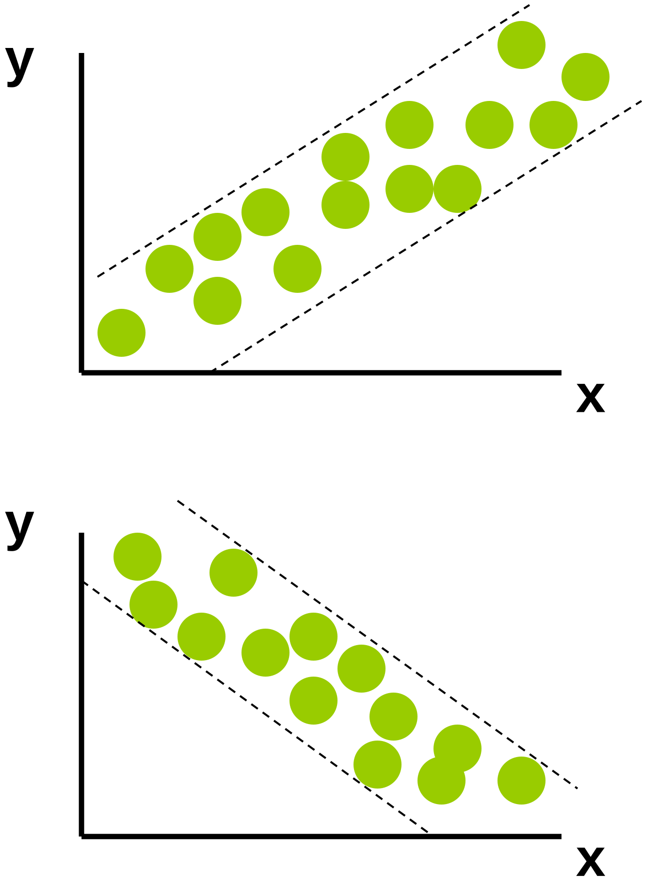




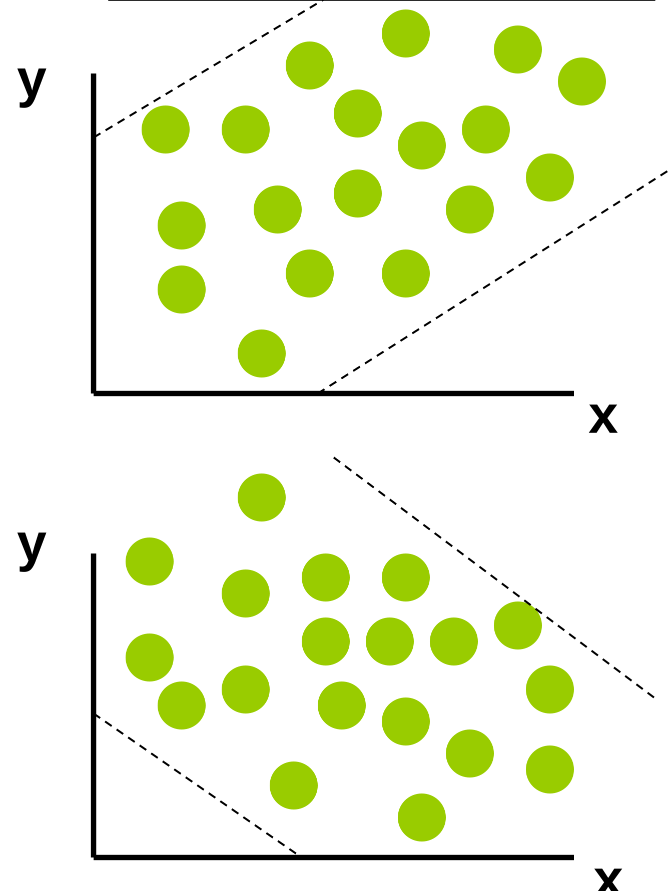
Scatter Plot Examples

(continued)

Strong relationships



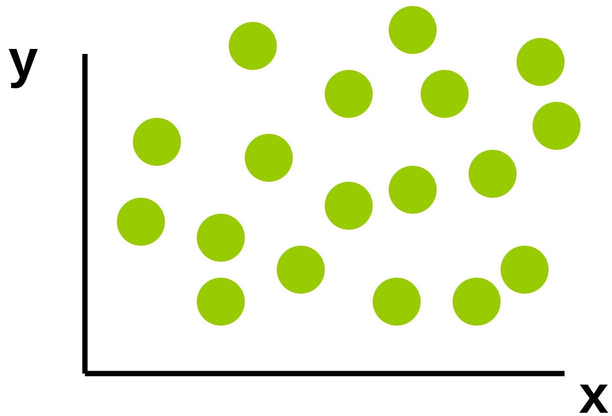
Weak relationships

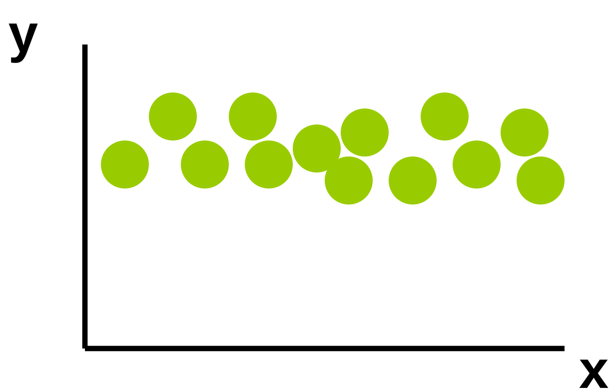


Scatter Plot Examples

(continued)

No relationship





Correlation Coefficient

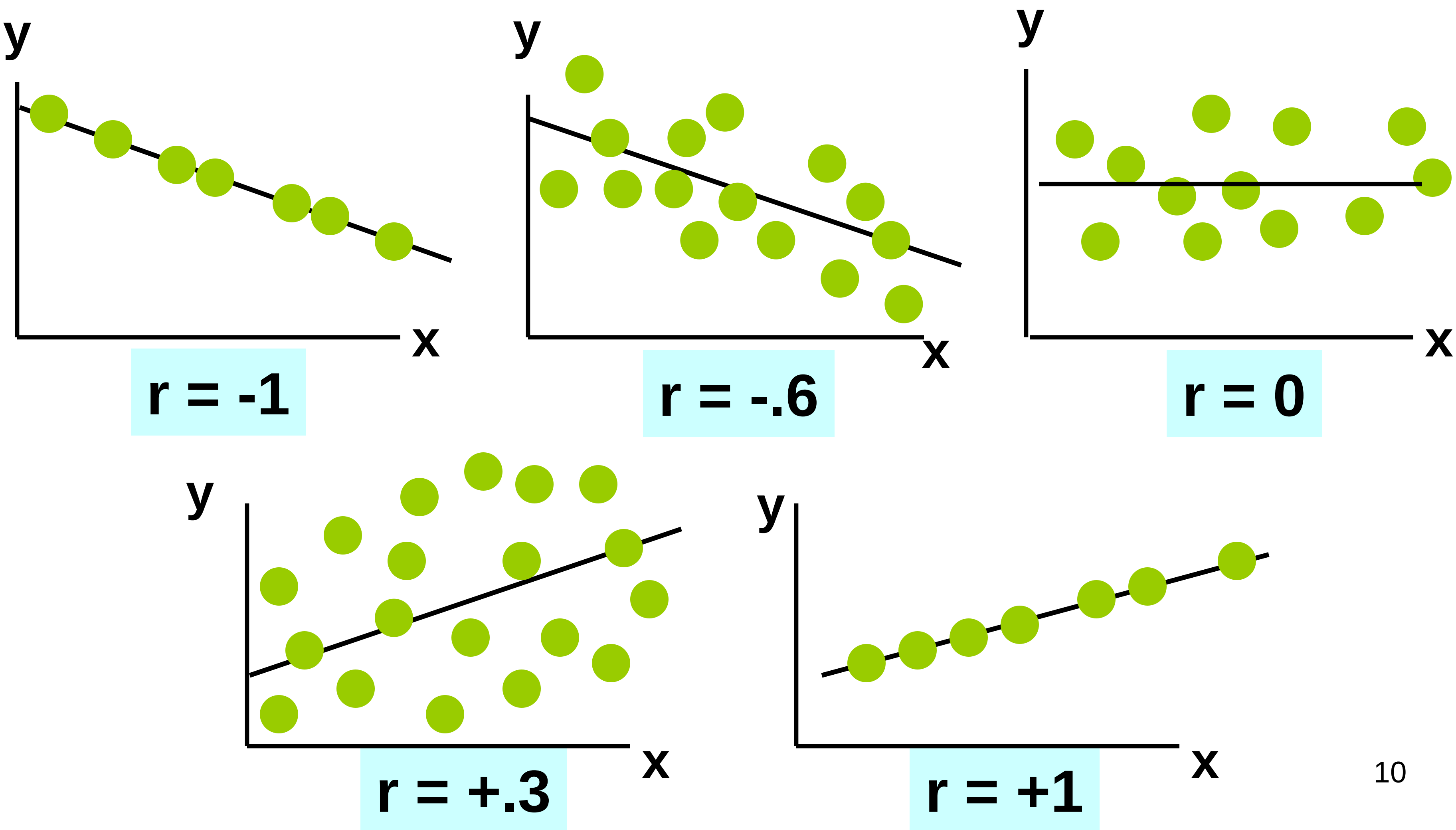
(continued)

- 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:

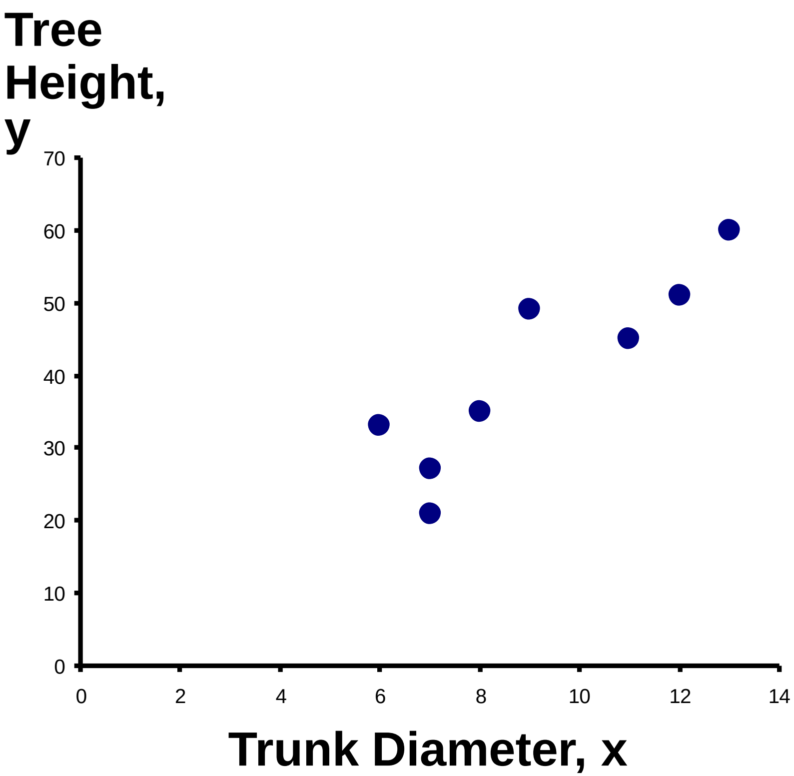
r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

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
$\Sigma=321$	$\Sigma=73$	$\Sigma=3142$	$\Sigma=14111$	$\Sigma=713$

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
  –
$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

(with n – 2 degrees of freedom)

Example: Produce Stores

Is there evidence of a linear relationship between tree height and trunk diameter at the .05 level of significance?

H_0: ρ = 0 (No correlation)

H_1: ρ ≠ 0 (correlation exists)

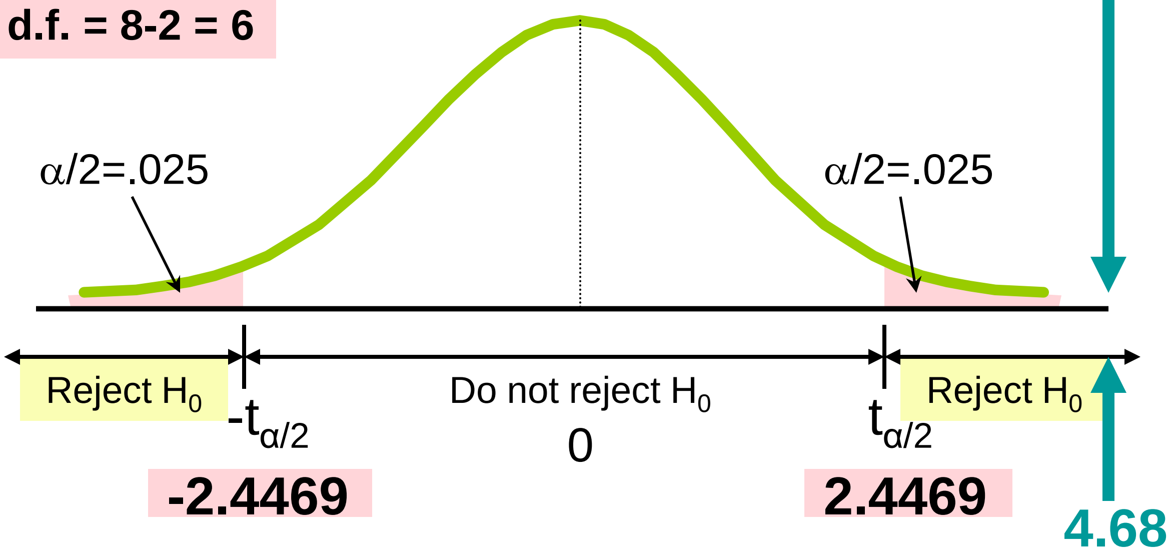
$\alpha = .05$, df = 8 - 2 = 6

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

Example: Test Solution

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

d.f. = 8-2 = 6



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

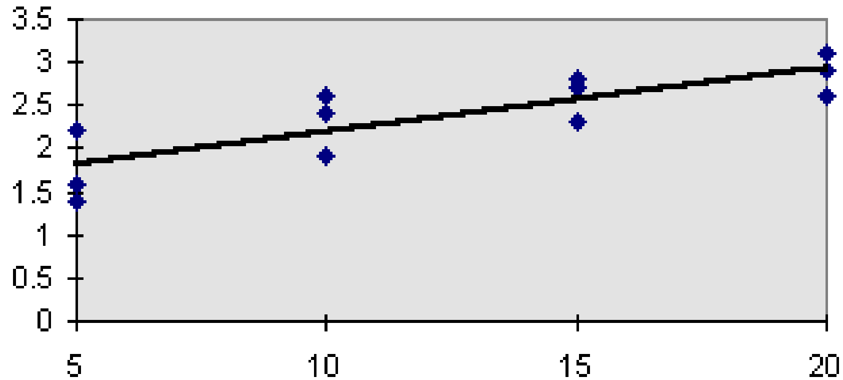
Independent variable: the variable used to explain the dependent variable

Simple Linear Regression Model

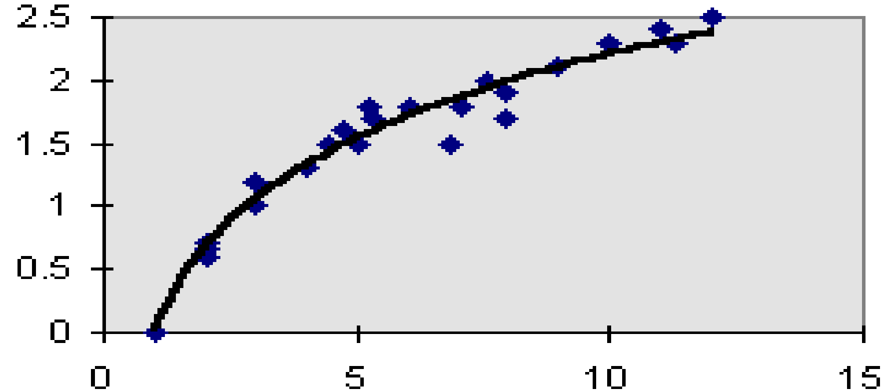
- Only **one** independent variable, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be caused by changes in x

Types of Regression Models

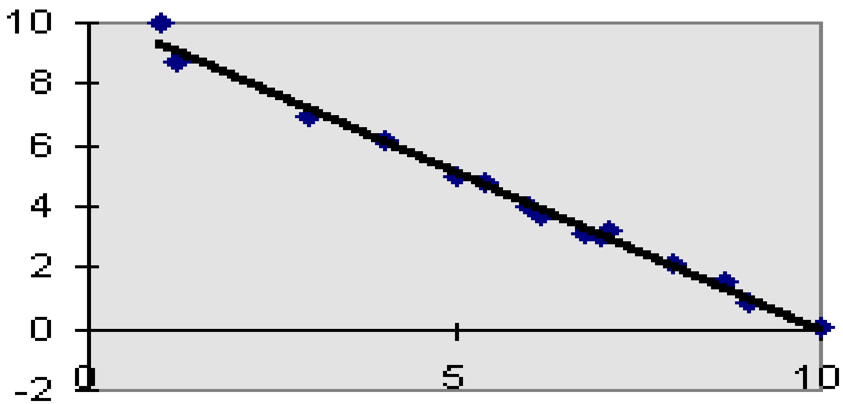
Positive Linear Relationship



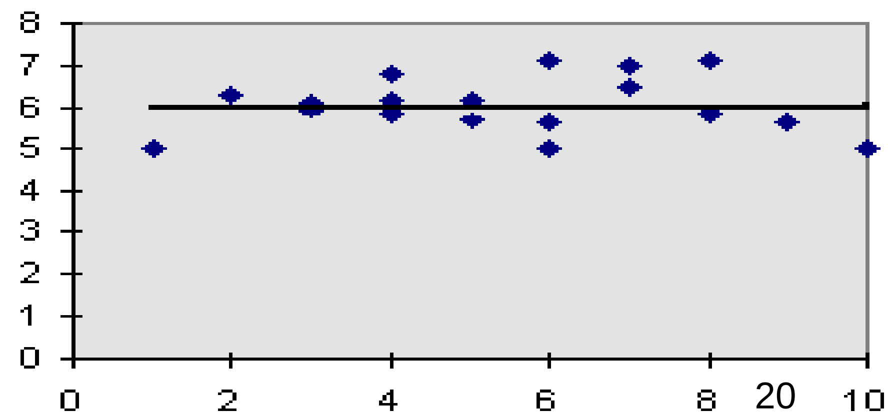
Relationship NOT Linear



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

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

Linear component

Random Error component

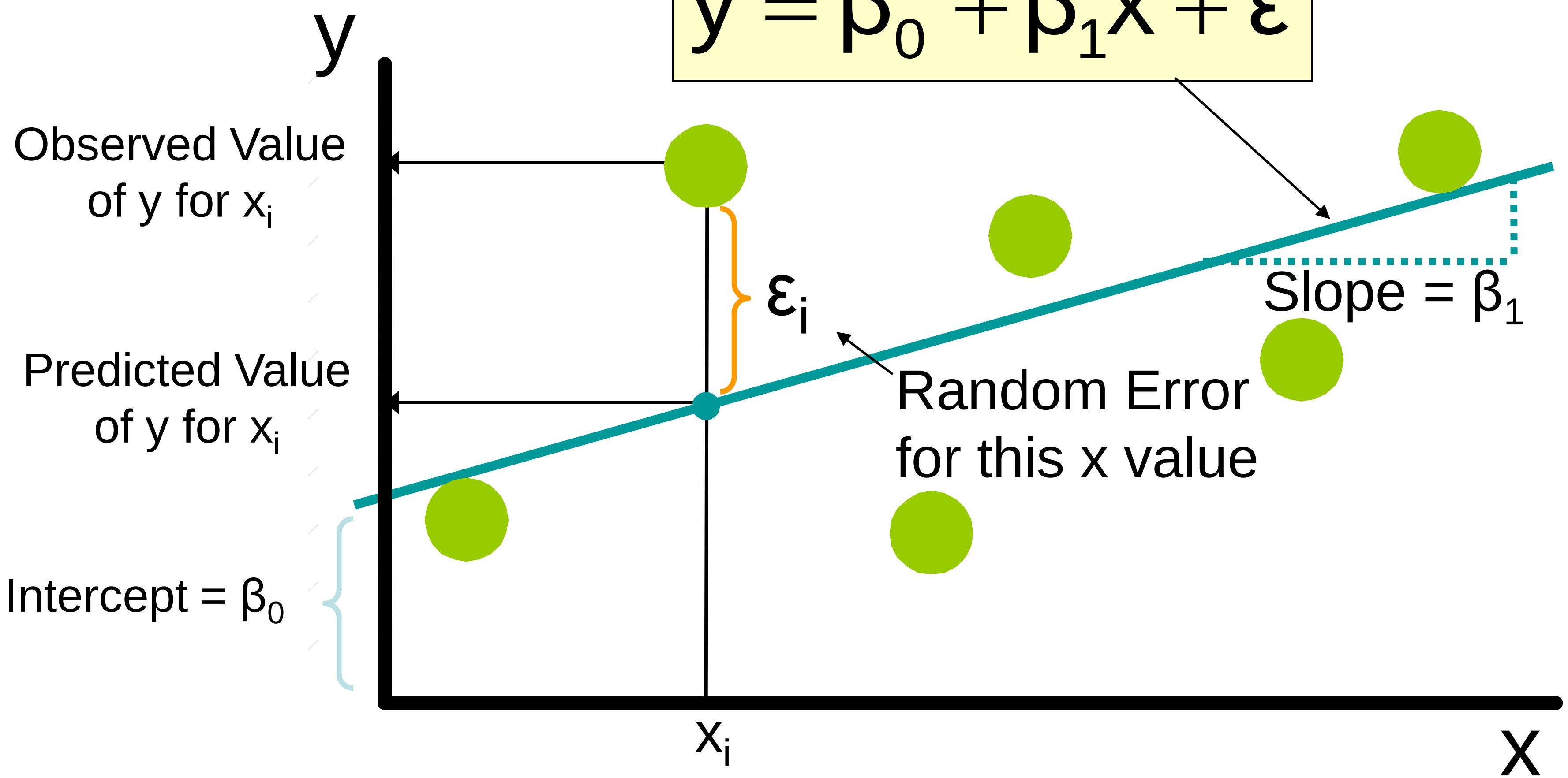
Linear Regression Assumptions

- 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)

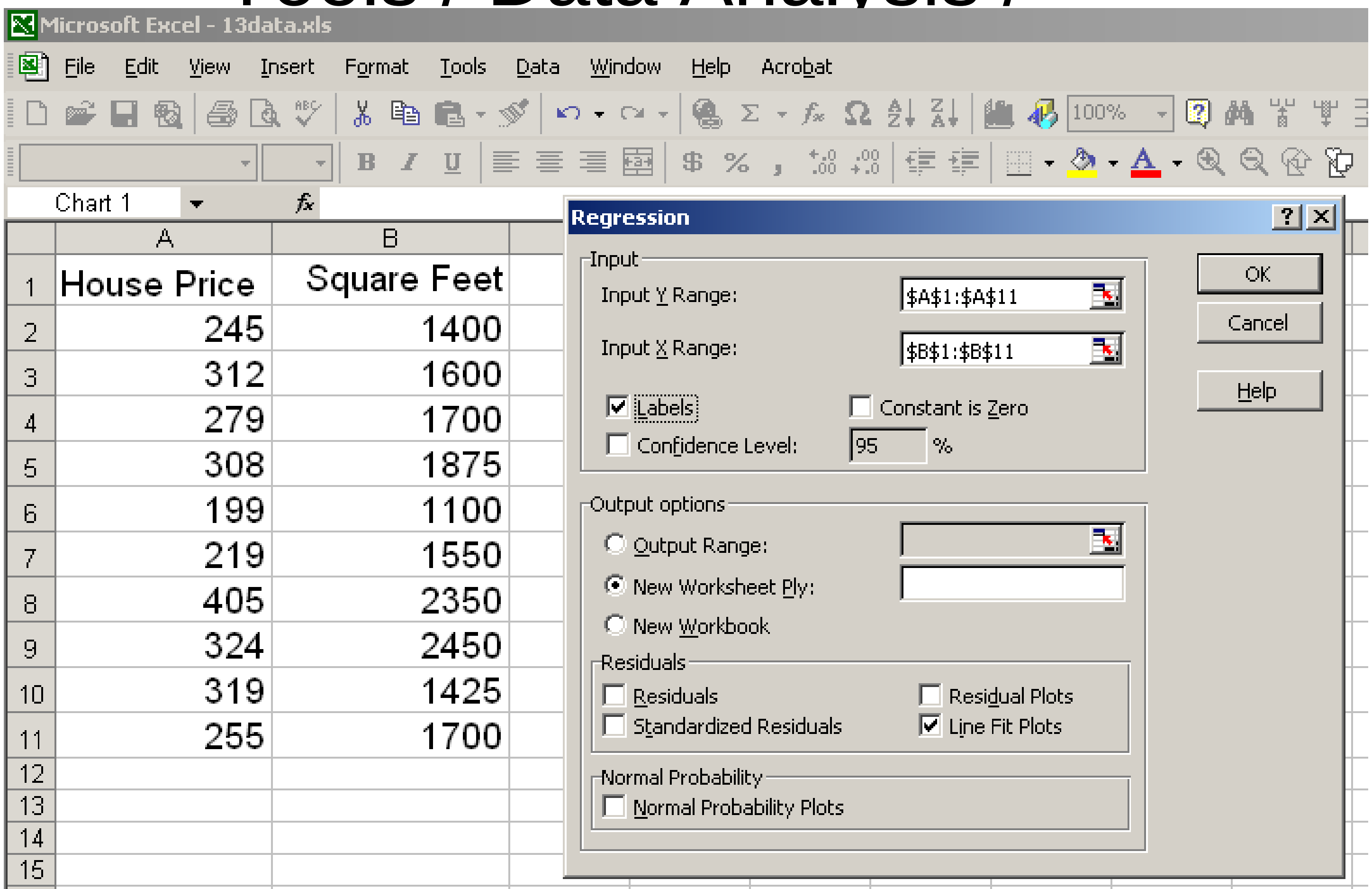
- A random sample of 10 houses is selected
  - Dependent variable (y) = house price in $1000s
  - Independent variable (x) = square feet

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 /



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:

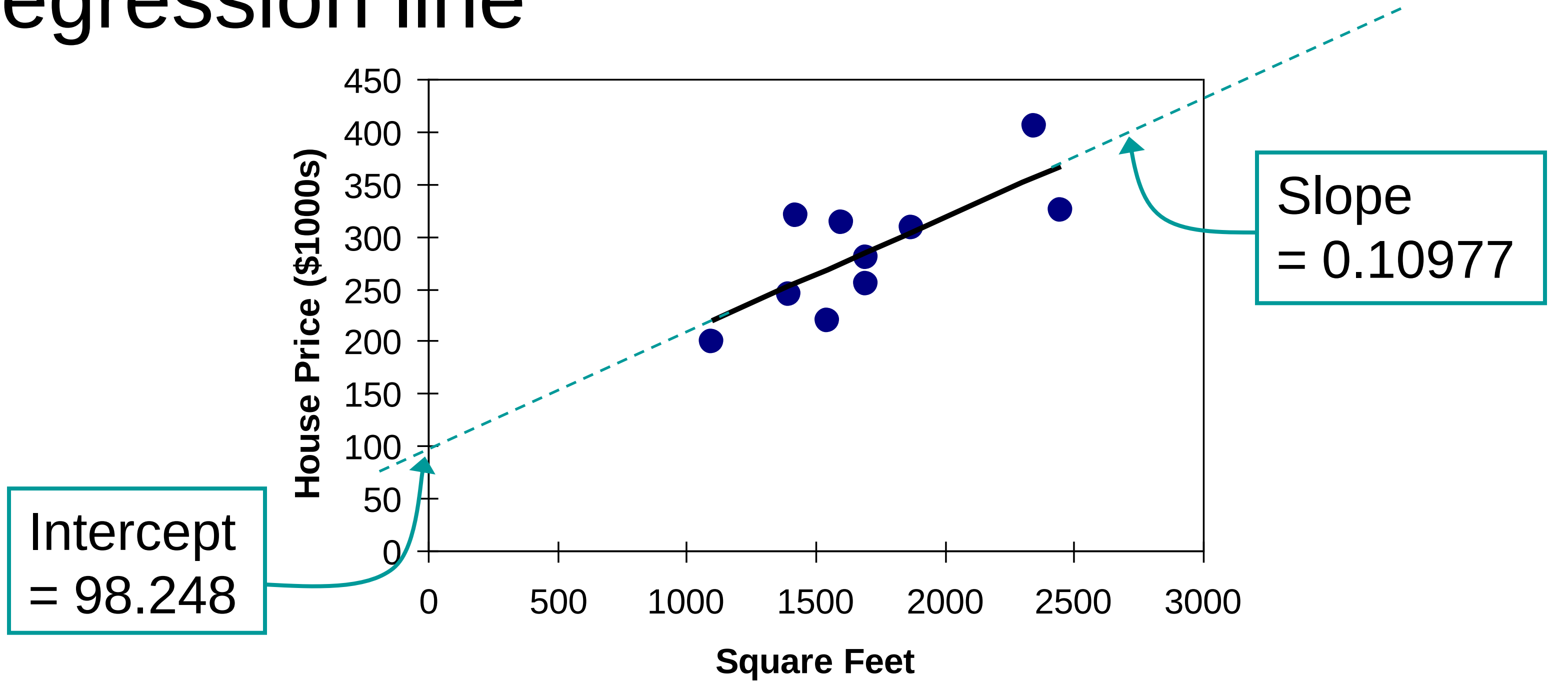
$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

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



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

Interpretation of the Intercept, b_0

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

- b_0 is the estimated average value of Y when the value of X is zero (if x = 0 is in the range of observed x values)
  - Here, no houses had 0 square feet, so b_0 = 98.24833 just indicates that, for houses within the range of sizes observed, \$98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient, b_1

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

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $b_1 = .10977$ tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Least Squares Regression Properties

- The sum of the residuals from the least squares regression line is 0 ($\sum (y-\hat{y}) = 0$)
- The sum of the squared residuals is a minimum (minimized $\sum (y-\hat{y})^2$)
- The simple regression line always passes through the mean of the y variable and the mean of the x variable
- The least squares coefficients are unbiased estimates of β_0 and β_1

Explained and Unexplained Variation

- Total variation is made up of two parts:

$$SST = SSE + SSR$$

Total sum of Squares	Sum of Squares Error	Sum of Squares Regression
$SST = \sum (y - \bar{y})^2$	$SSE = \sum (y - \hat{y})^2$	$SSR = \sum (\hat{y} - \bar{y})^2$

where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

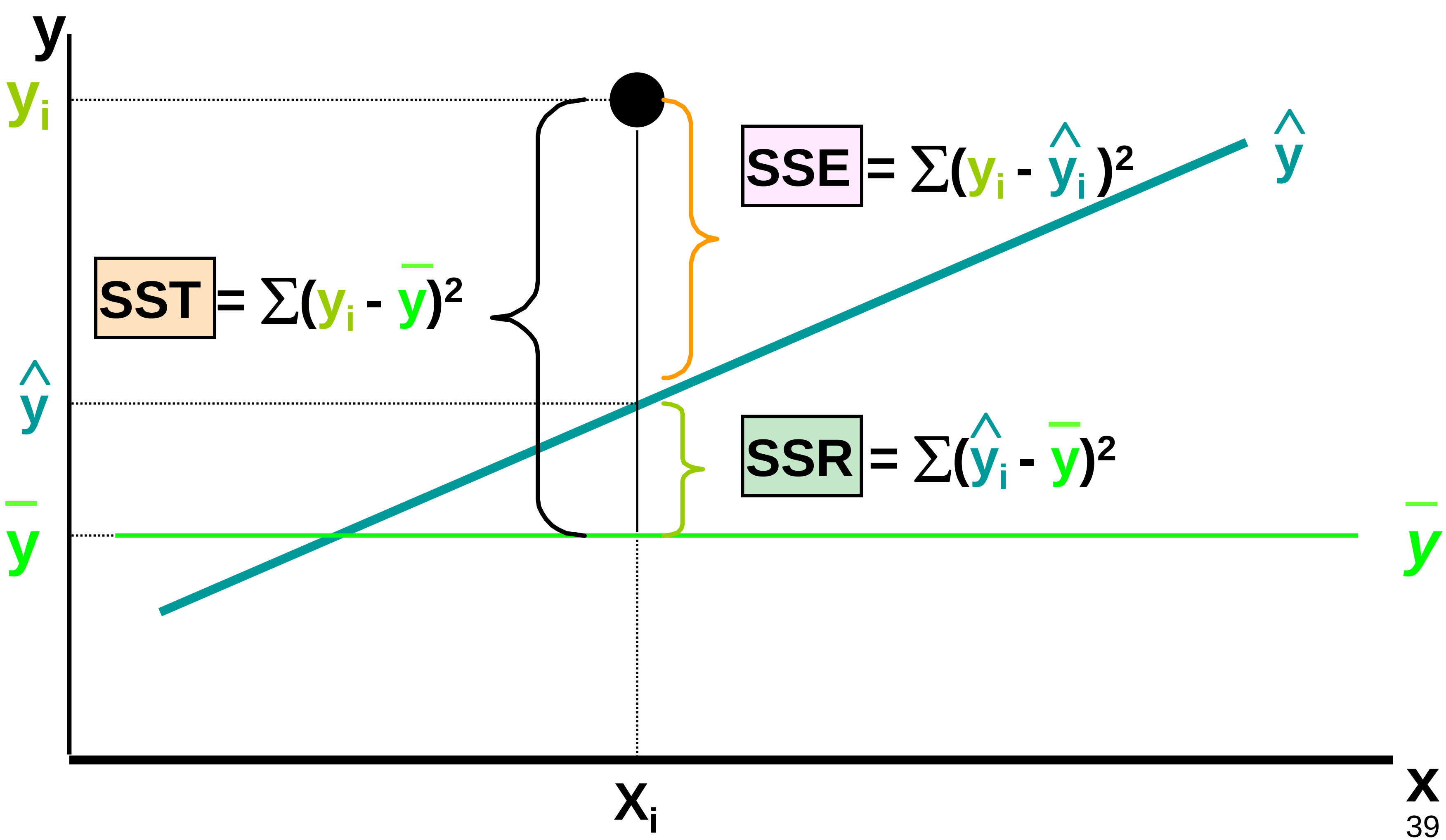
Explained and Unexplained Variation

(continued)

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean $\bar{y}$
- SSE = error sum of squares
  - Variation attributable to factors other than the relationship between x and y
- SSR = regression sum of squares
  - Explained variation attributable to the relationship between x and y

Explained and Unexplained Variation

(continued)



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

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

where

$$0 \leq R^2 \leq 1$$

Coefficient of Determination, R^2

(continued)

Coefficient of determination

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

Note: In the single independent variable case, the coefficient of determination is

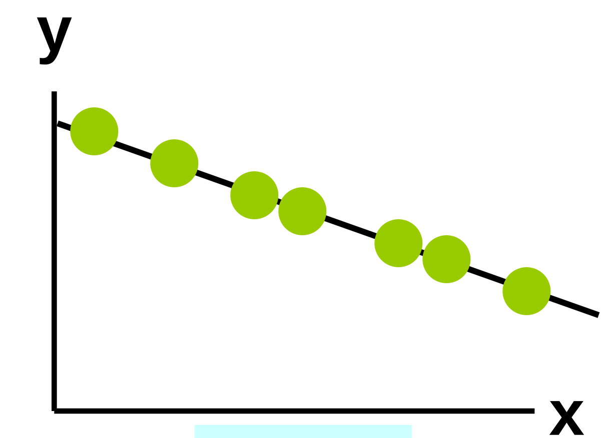
$$R^2 = r^2$$

where:

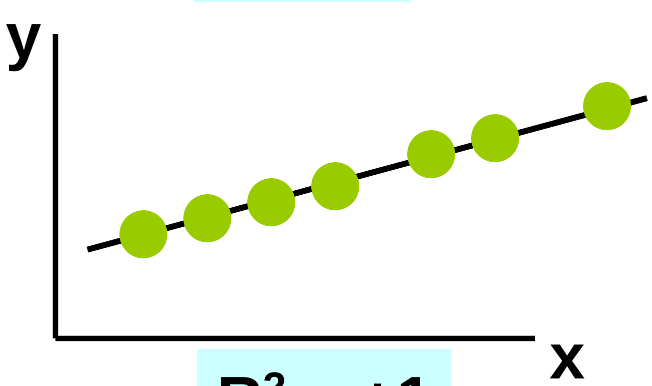
R^2 = Coefficient of determination

r = Simple correlation coefficient

Examples of Approximate R^2 Values



$R^2 = 1$



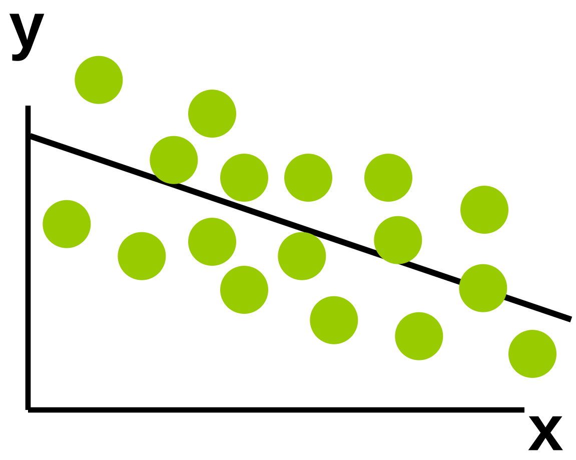
$R^2 = +1$

$R^2 = 1$

Perfect linear relationship between x and y:

100% of the variation in y is explained by variation in x

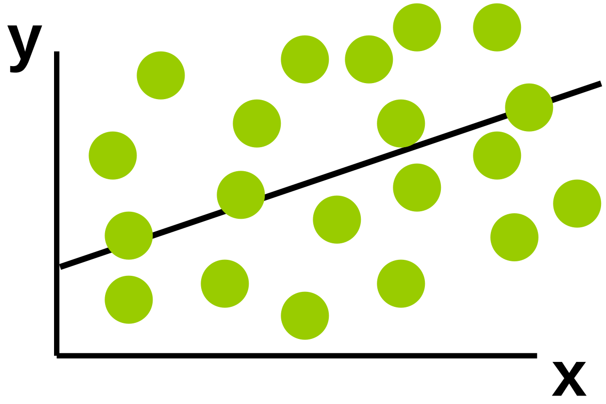
Examples of Approximate R^2 Values



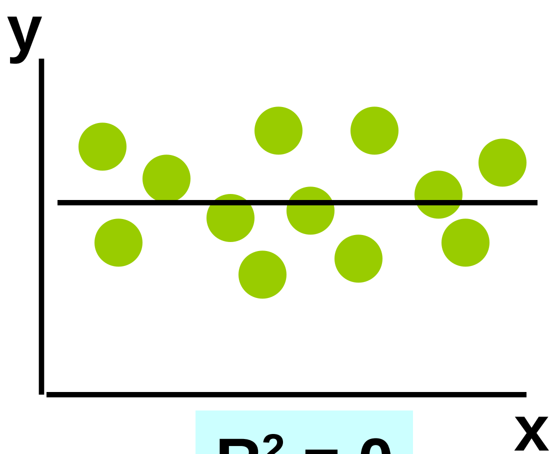
$0 < R^2 < 1$

Weaker linear relationship between x and y:

Some but not all of the variation in y is explained by variation in x



Examples of Approximate R^2 Values



$R^2 = 0$

$R^2 = 0$

No linear relationship between x and y:

The value of Y does not depend on x. (None of the variation in y is explained by variation in x)

Excel Output

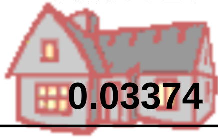
Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

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

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

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



Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$s_{\varepsilon} = \sqrt{\frac{SSE}{n-k-1}}$$

Where

SSE = Sum of squares error
n = Sample size
k = number of independent variables in the model

The Standard Deviation of the Regression Slope

- The standard error of the regression slope coefficient (b_1) is estimated by

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

where:

s_{b_1} = Estimate of the standard error of the least squares slope

$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ = Sample standard error of the estimate

Excel Output

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

$$s_\varepsilon = 41.33032$$

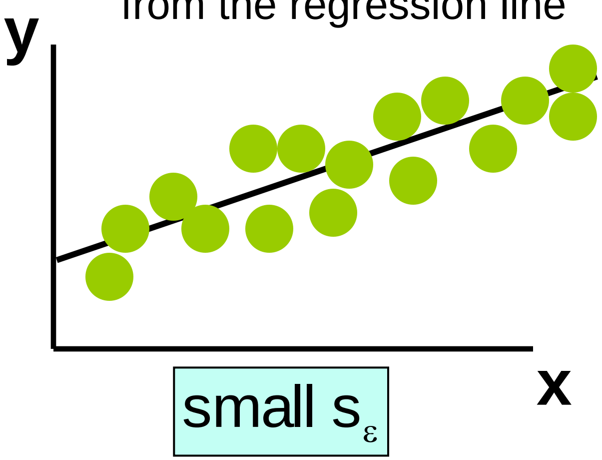
$$s_{b_1} = 0.03297$$

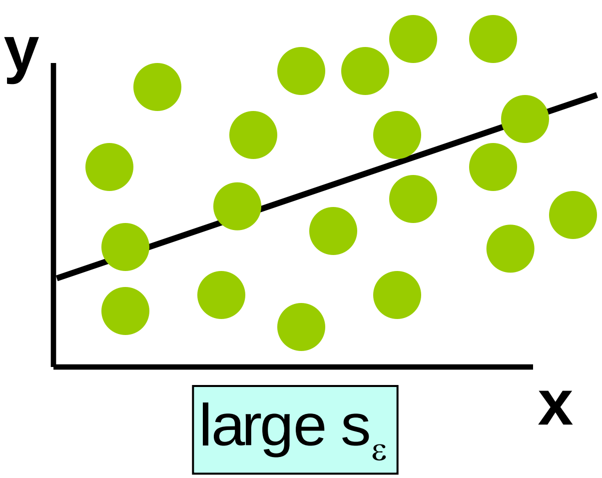
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	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

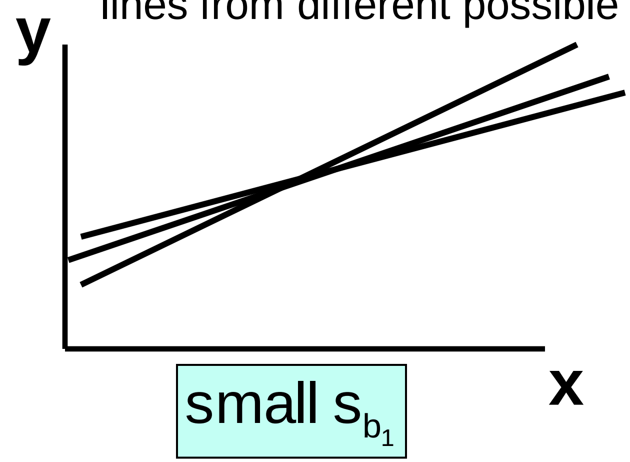
Comparing Standard Errors

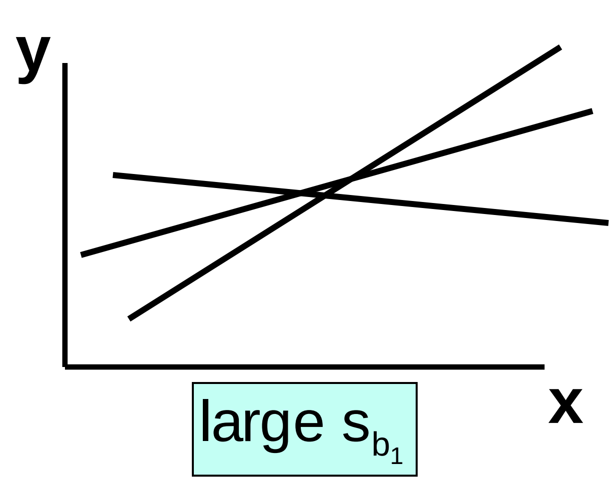
Variation of observed y values from the regression line





Variation in the slope of regression lines from different possible samples





Inference about the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between x and y?
- Null and alternative hypotheses
  - H_0: $\beta_1 = 0$ (no linear relationship)
  - H_1: $\beta_1 \neq 0$ (linear relationship does exist)
- Test statistic
  -

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

$$\text{d.f.} = n - 2$$

where:

b_1 = Sample regression slope coefficient

β_1 = Hypothesized slope

s_{b1} = Estimator of the standard error of the slope

Inference about the Slope: t Test

(continued)

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

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

The slope of this model is 0.1098

Does square footage of the house affect its sales price?

Inferences about the Slope: t Test Example

Test Statistic: t = 3.329

$H_0: \beta_1 = 0$

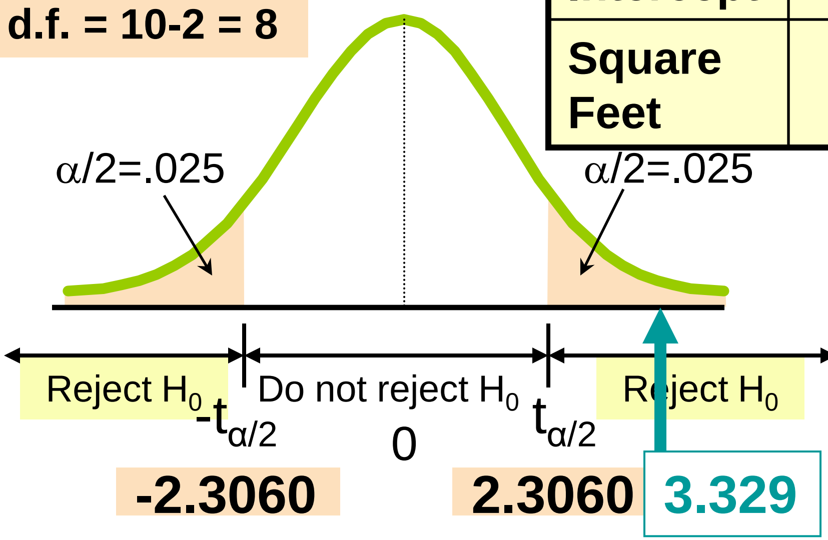
$H_A: \beta_1 \neq 0$

From Excel output:

b_1 s_{b_1} t

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

d.f. = 10-2 = 8



Decision: Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Regression Analysis for Description

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

d.f. = n - 2

Excel Printout for House Prices:

	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**

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

Regression Analysis for Description

	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

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

Confidence Interval for the Average y, Given x

Confidence interval estimate for the **mean of y** given a particular x_p

Size of interval varies according to distance away from mean, $\bar{x}$

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

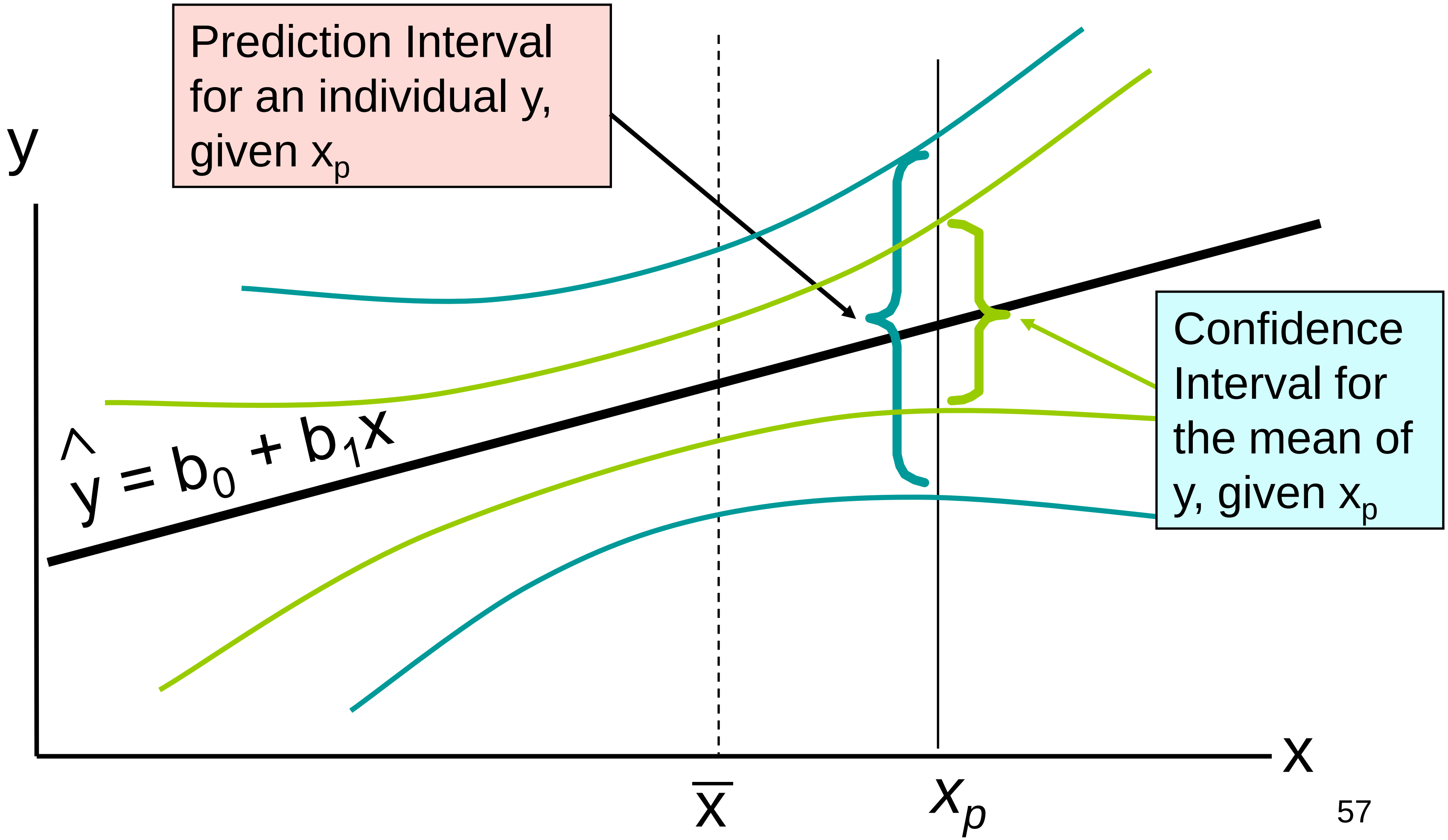
Confidence Interval for an Individual y, Given x

Confidence interval estimate for an **Individual value of y** given a particular x_p

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case

Interval Estimates for Different Values of x



Example: House Prices

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

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

Predict the price for a house with 2000 square feet

Example: House Prices

(continued)

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Estimation of Mean Values: Example

Confidence Interval Estimate for $E(y)|x_p$

Find the 95% confidence interval for the average price of 2,000 square-foot houses

Predicted Price $\hat{Y}_i = 317.85$ ($1,000s)

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints are 280.66 -- 354.90, or from $280,660 -- $354,900

Estimation of Individual Values: Example

Prediction Interval Estimate for $y|x_p$

Find the 95% confidence interval for an individual house with 2,000 square feet

Predicted Price $\hat{Y}_i$ = 317.85 ($1,000s)

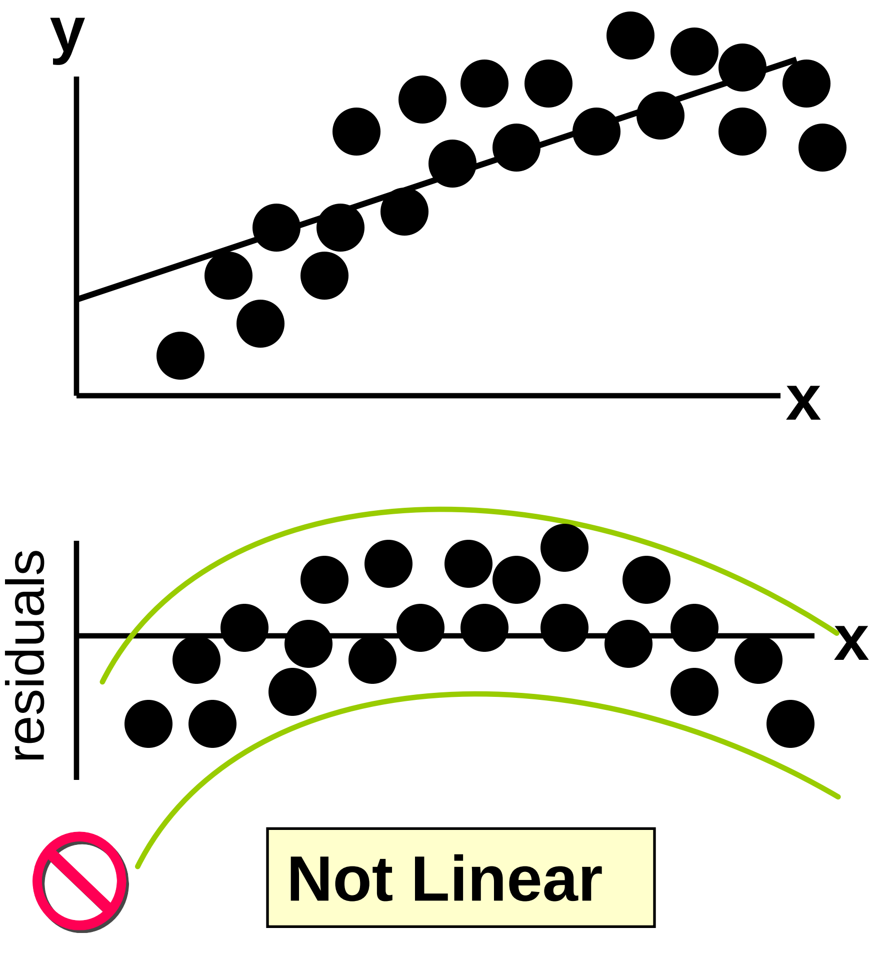
$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}} = 317.85 \pm 102.28$$

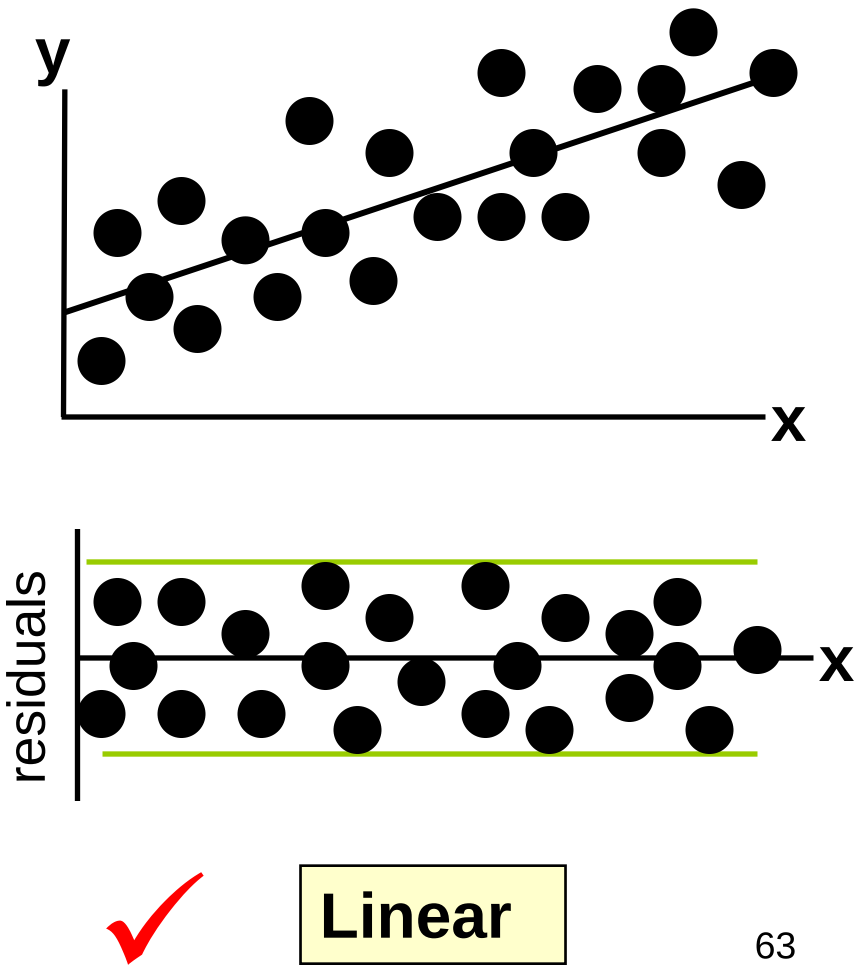
The prediction interval endpoints are 215.50 -- 420.07, or from $215,500 -- $420,070

Residual Analysis

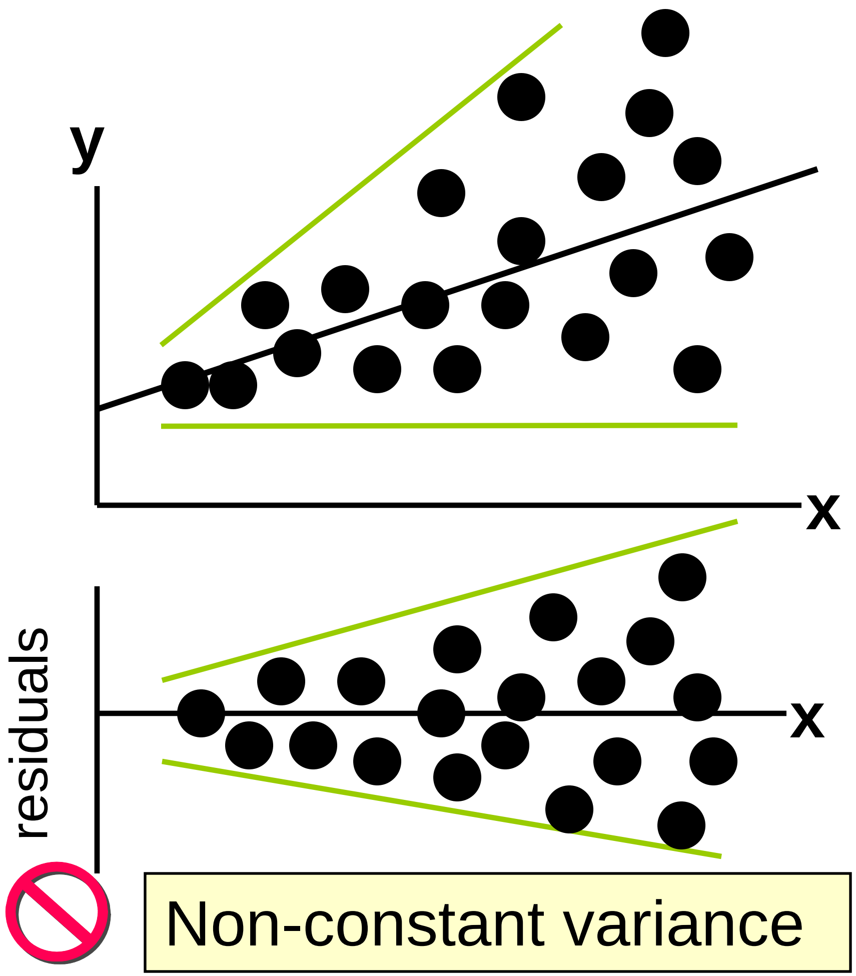
- Purposes
  - Examine for linearity assumption
  - Examine for constant variance for all levels of x
  - Evaluate normal distribution assumption
- Graphical Analysis of Residuals
  - Can plot residuals vs. x
  - Can create histogram of residuals to check for normality

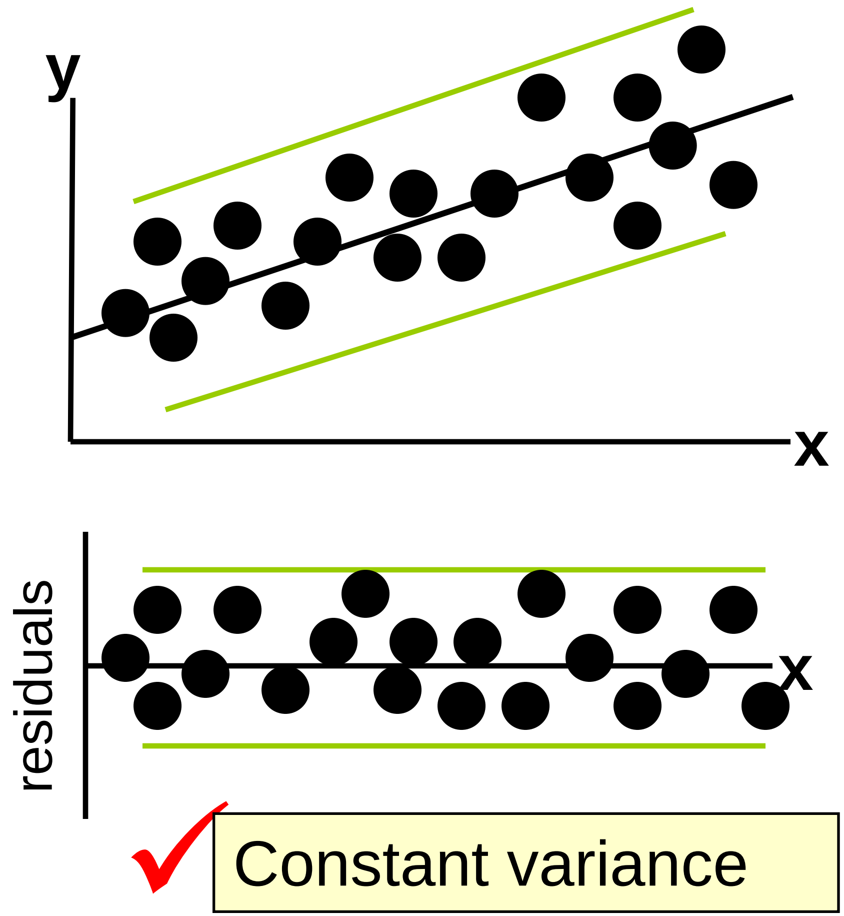
Residual Analysis for Linearity





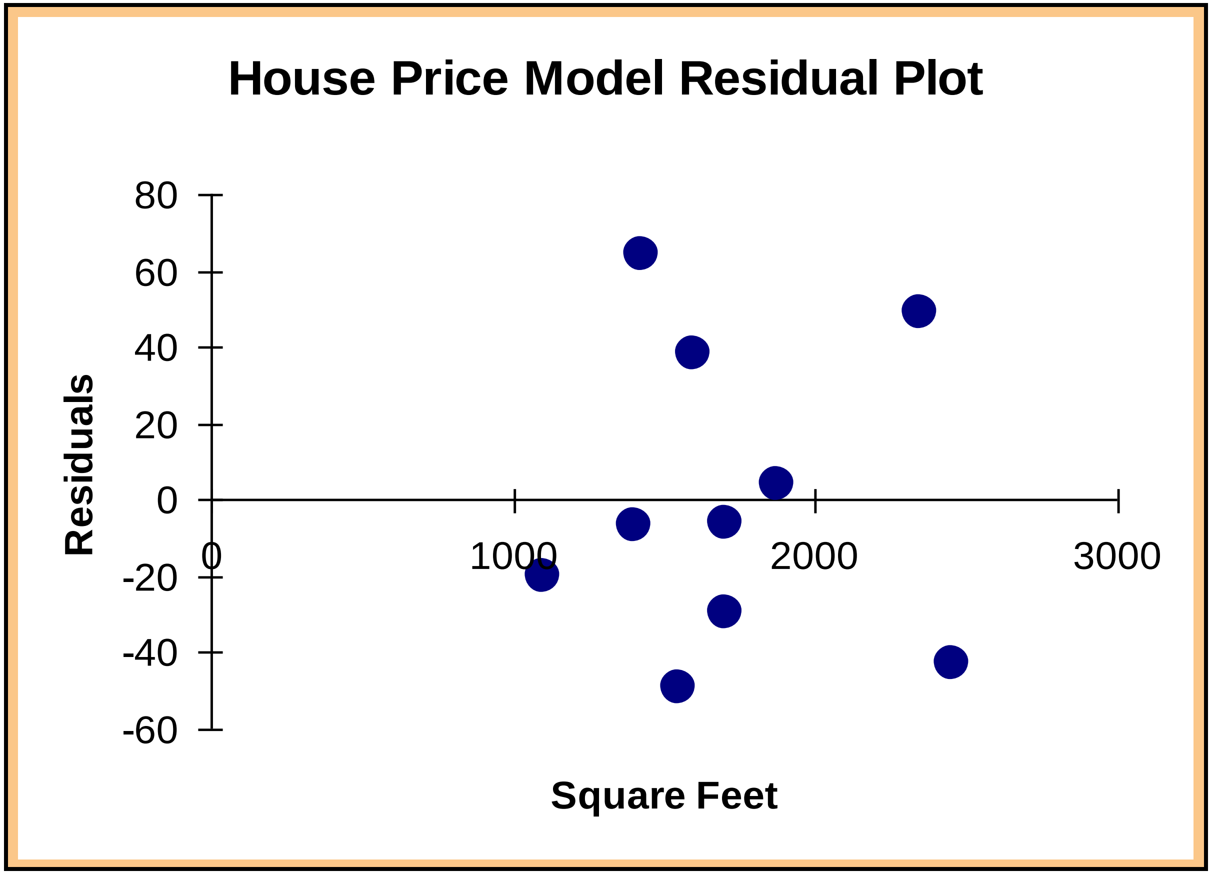
Residual Analysis for Constant Variance





Excel Output

RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



- Thank you.