Simple Linear Regression and Correlation

In this chapter, you learn:

- How to use regression analysis to predict the value of a dependent variable based on an independent variable
- The meaning of the regression coefficients b_0 and b_1
- How to evaluate the assumptions of regression analysis and know what to do if the assumptions are violated
- To make inferences about the slope and correlation coefficient
- To estimate mean values and predict individual values

Correlation vs. Regression

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

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 predict or explain (i.e. runoff)

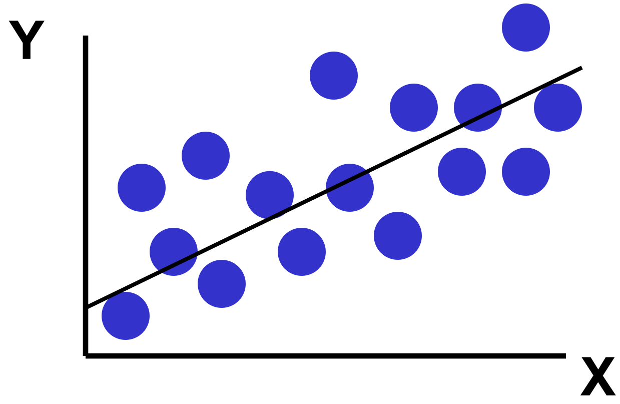
Independent variable: the variable used to explain the dependent variable (i.e. rainfall)

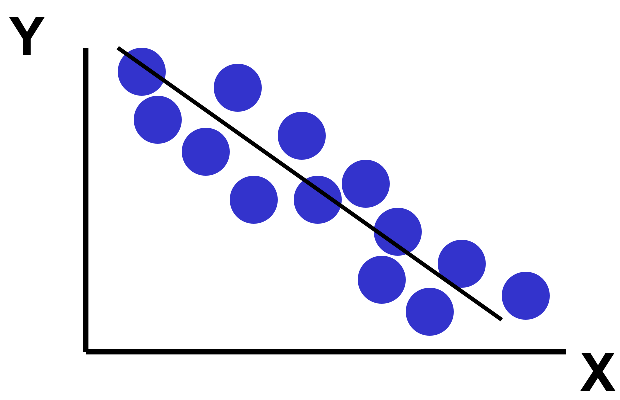
Simple Linear Regression Model

- 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

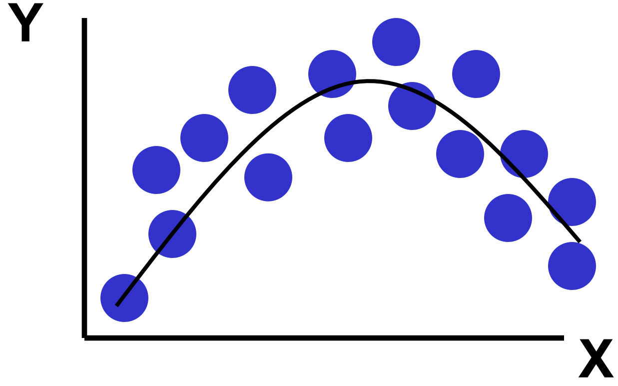
Types of Relationships

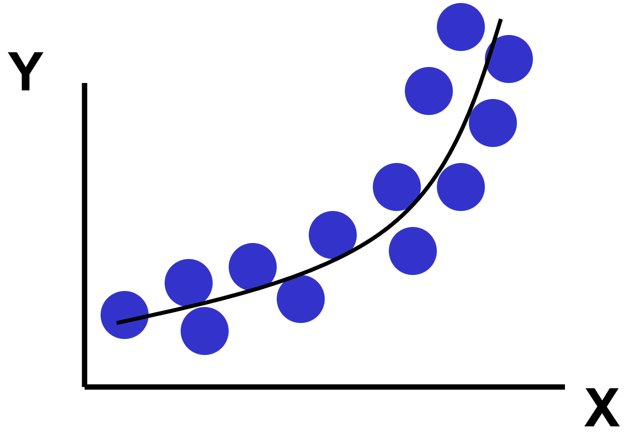
Linear relationships





Curvilinear relationships

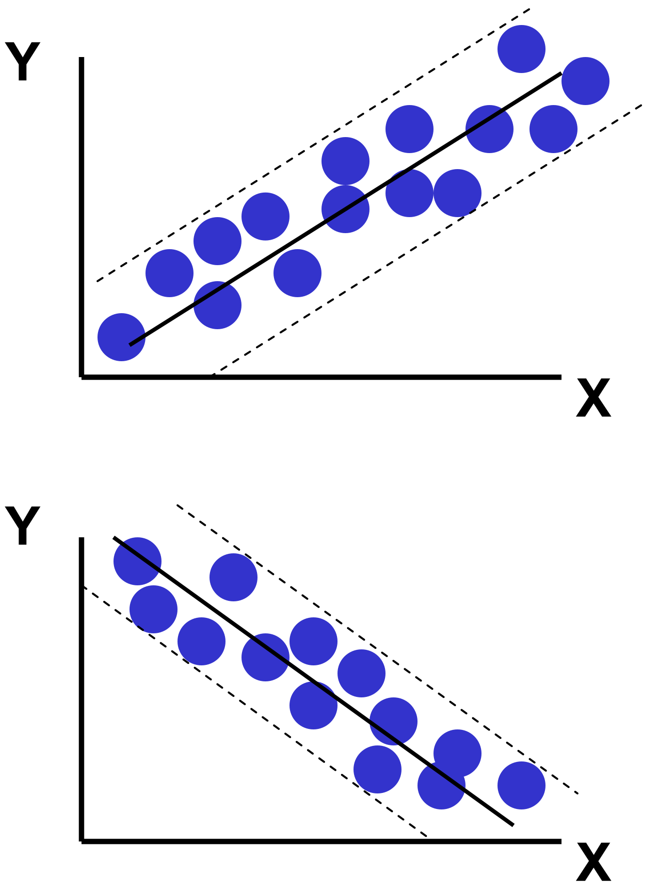




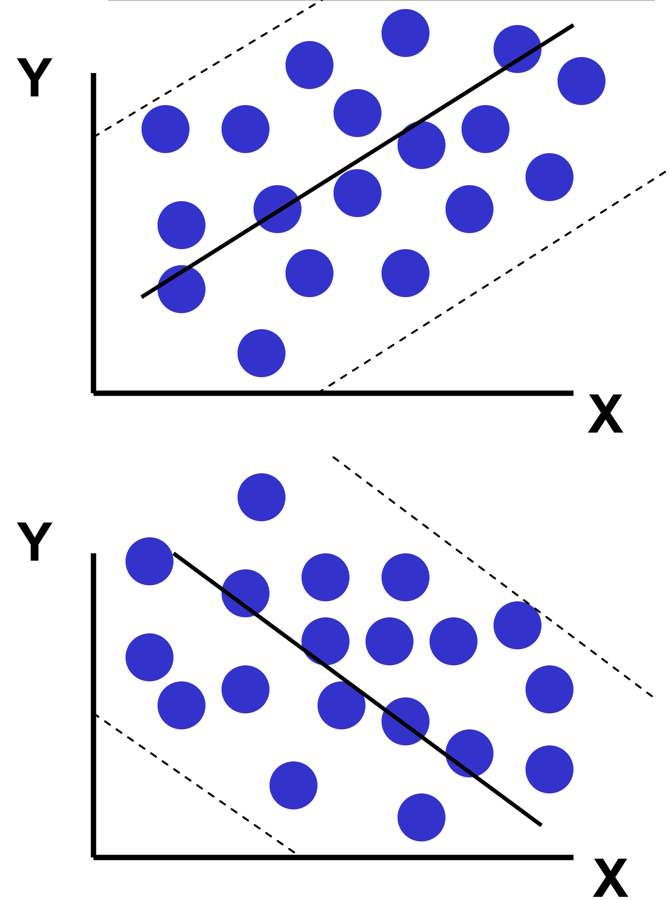
Types of Relationships

(continued)

Strong relationships



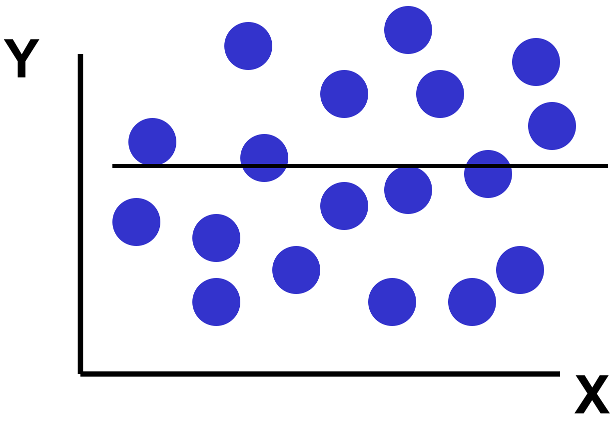
Weak relationships

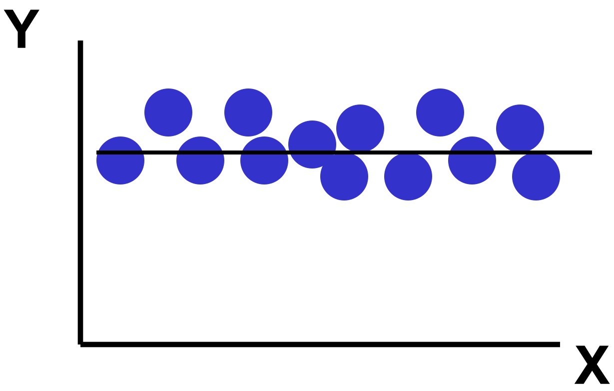


Types of Relationships

(continued)

No relationship





Simple Linear Regression Model

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

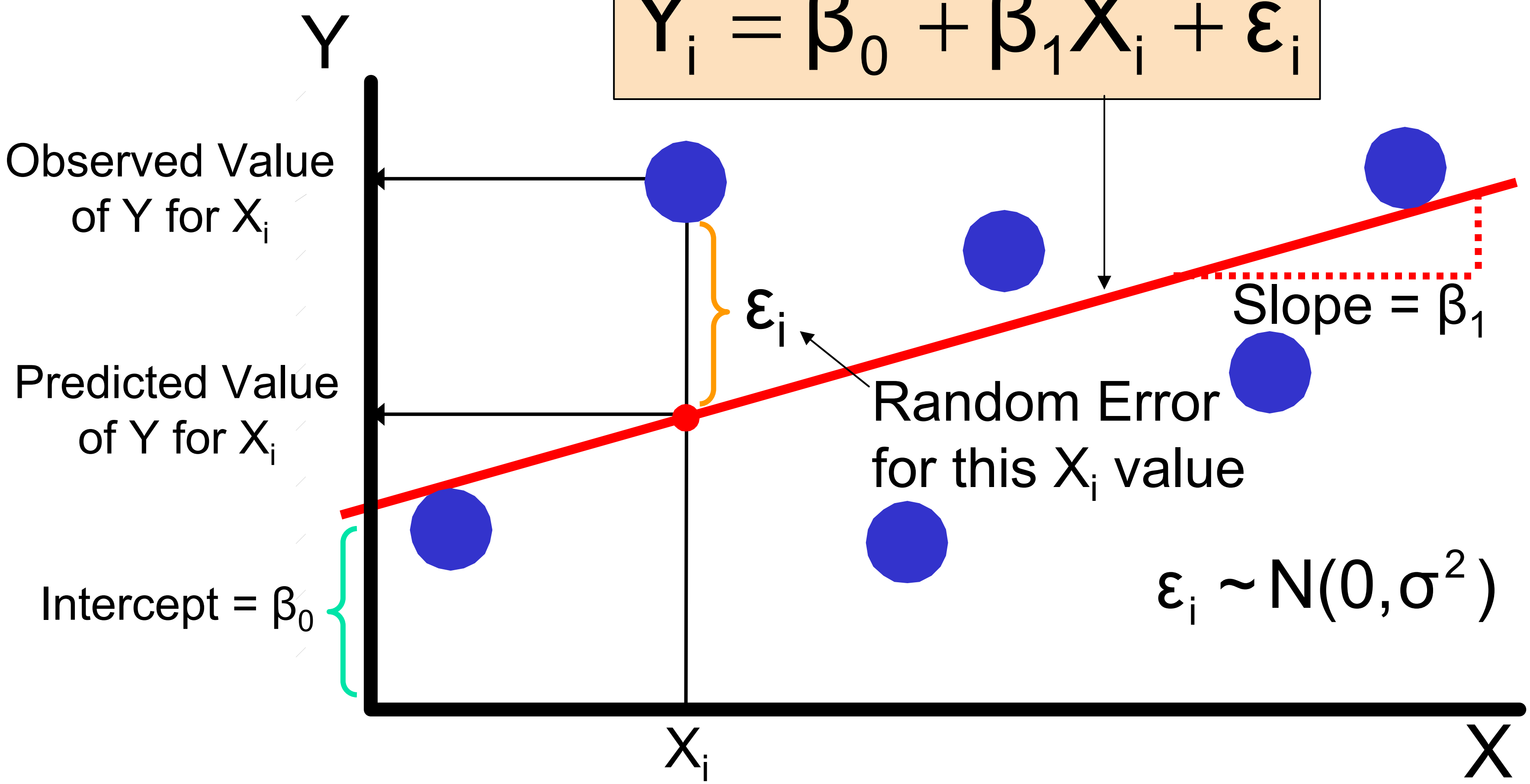
Dependent Variable

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Simple Linear Regression Model

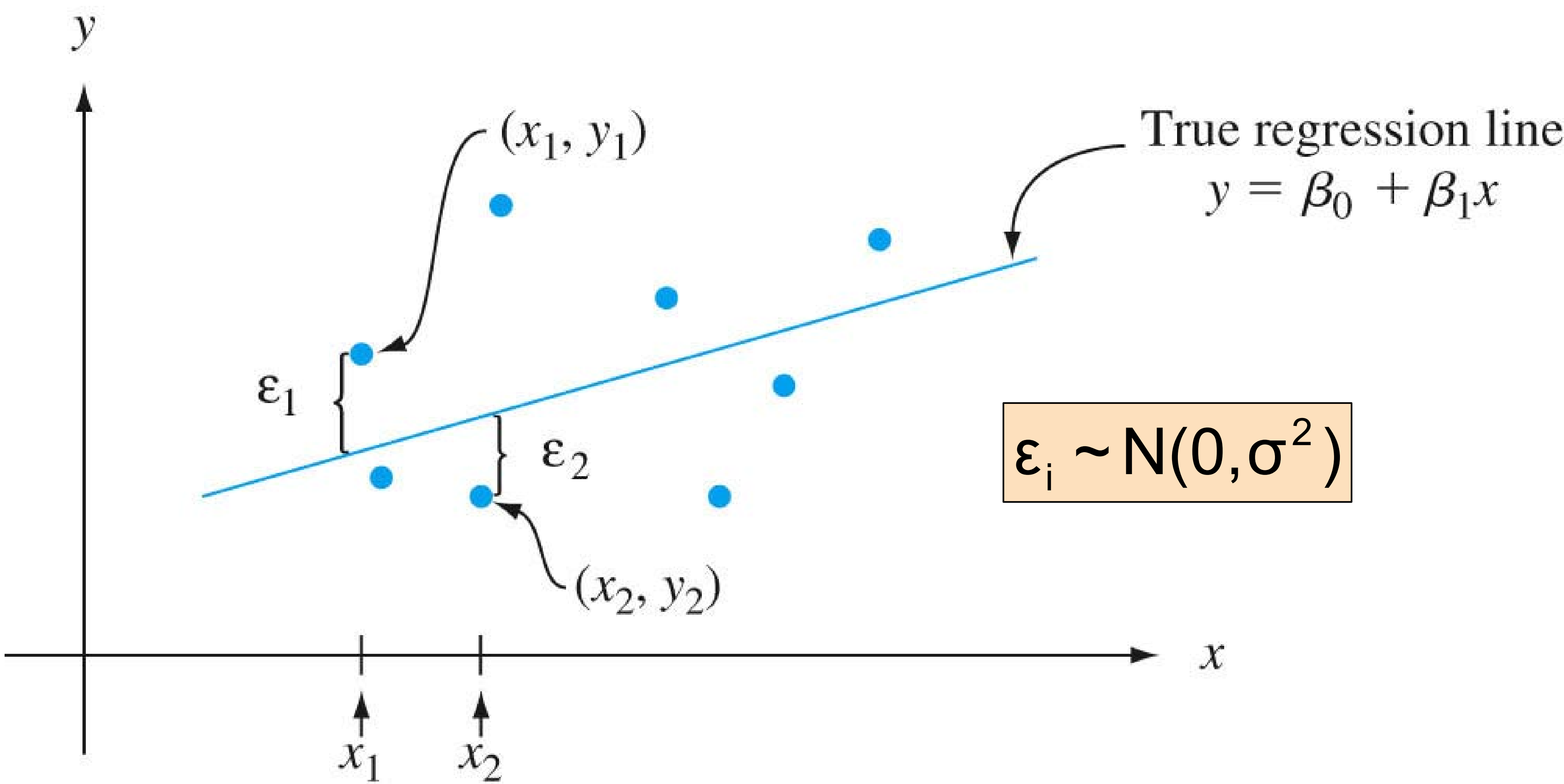
(continued)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Simple Linear Regression Model

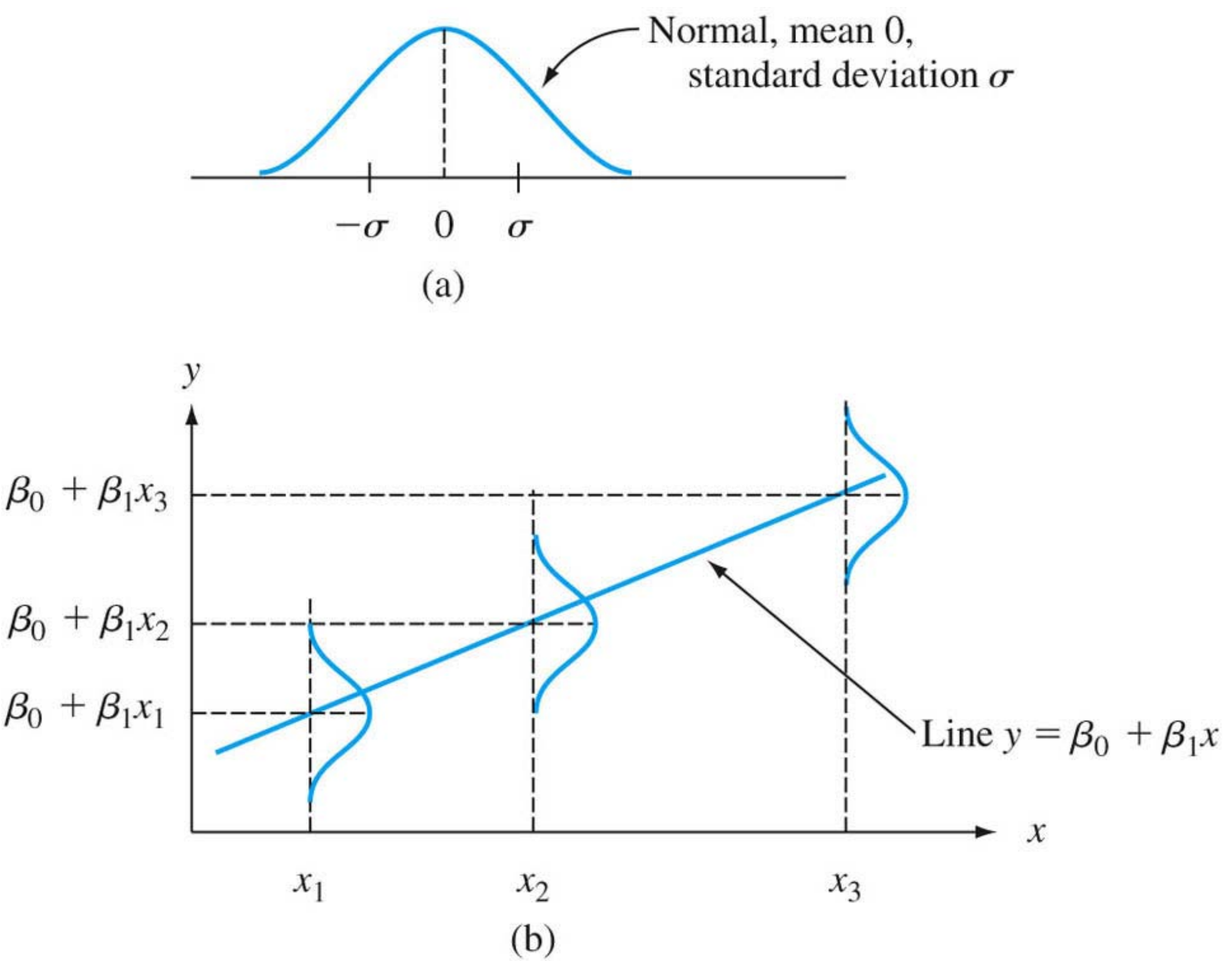
(continued)



$$\varepsilon_i \sim N(0, \sigma^2)$$

Simple Linear Regression Model

(continued)



$$\varepsilon_i \sim N(0, \sigma^2)$$

Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

The individual random error terms e_i have a mean of zero

Least Squares Method

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

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

Finding the Least Squares Equation

- Computational formula for the slope b_1:

$$\boxed{b_1 = \frac{SS_{XY}}{SS_{XX}}}$$

$$b_1 \Rightarrow \hat{\beta}_1 \text{ in the text book}$$

where

$$SS_{XY} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

$$SS_{XX} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

Finding the Least Squares Equation

- Computational formula for the Y intercept b_0:

$$\boxed{b_0 = \overline{Y} - b_1 \overline{X}}$$

$b_0 \Rightarrow \hat{\beta}_0$ in the text book

where

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Finding the Least Squares Equation

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

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

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)

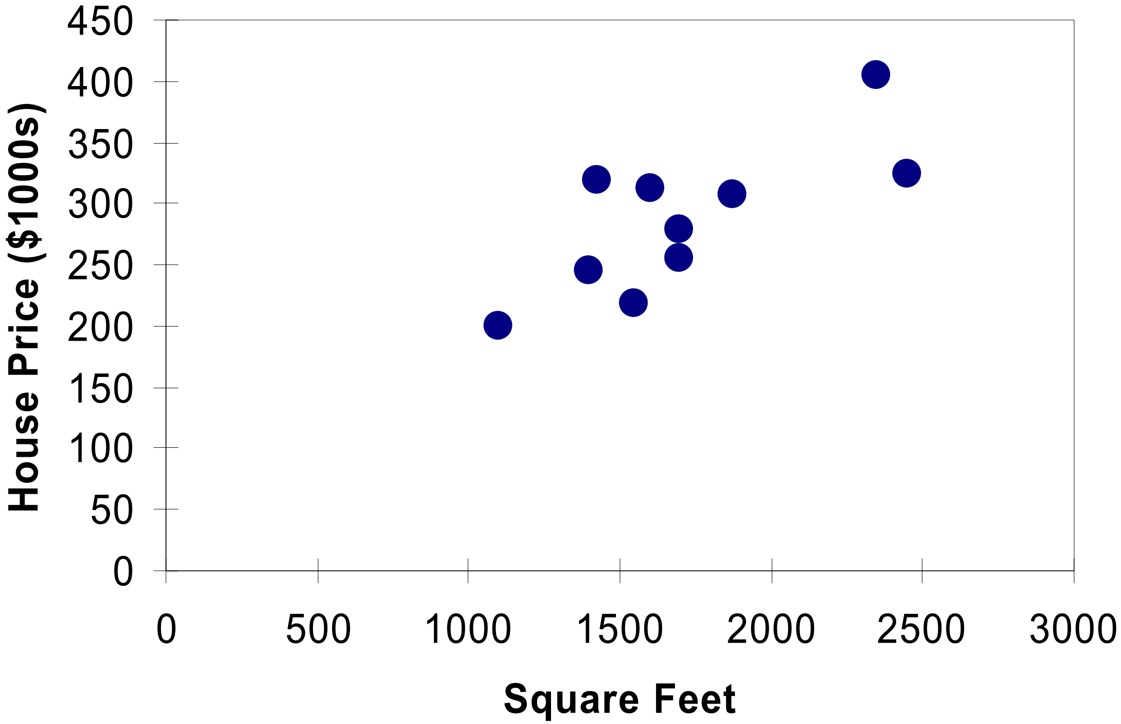
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

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

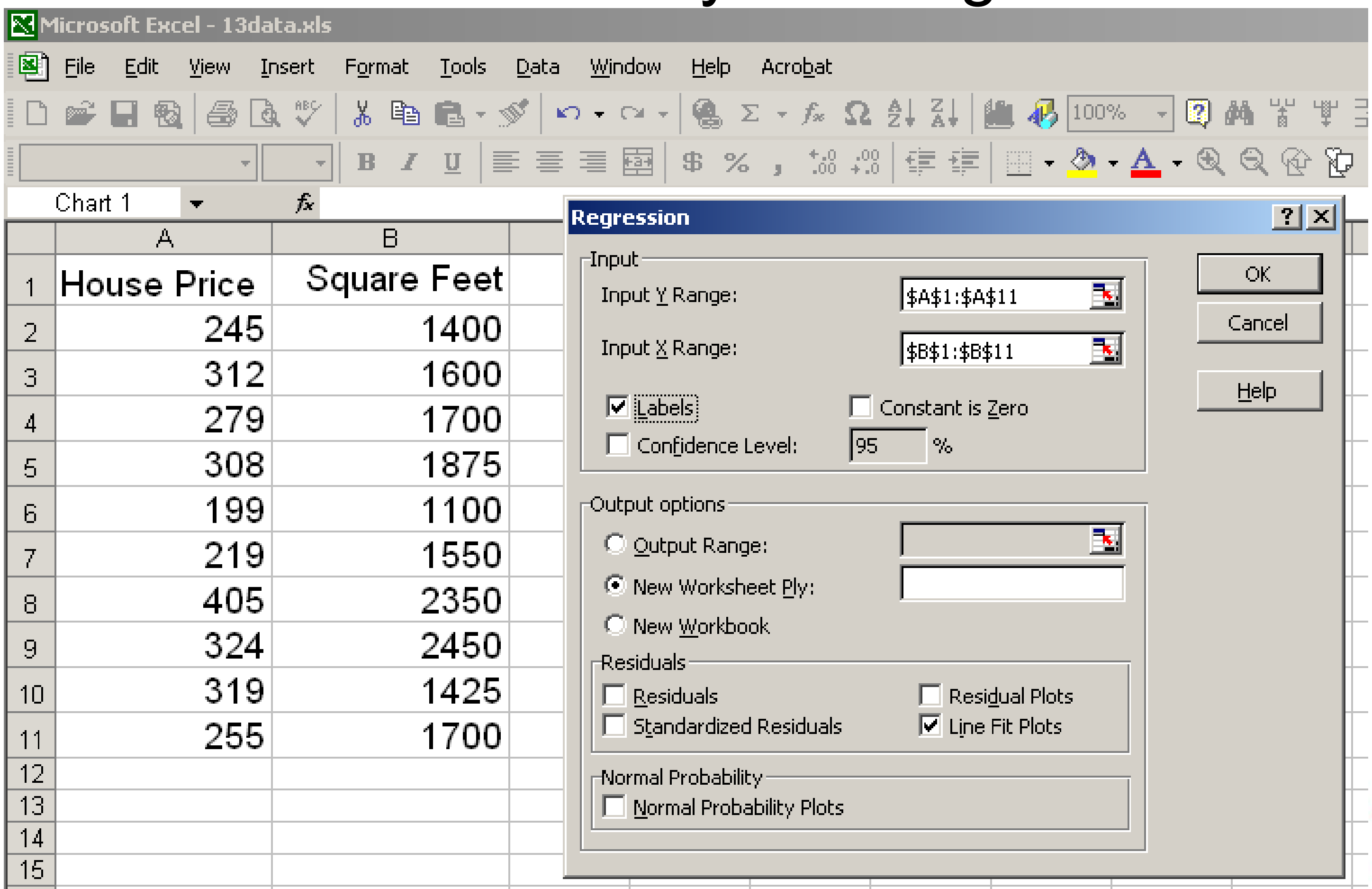
Graphical Presentation

- House price model: scatter plot



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:

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

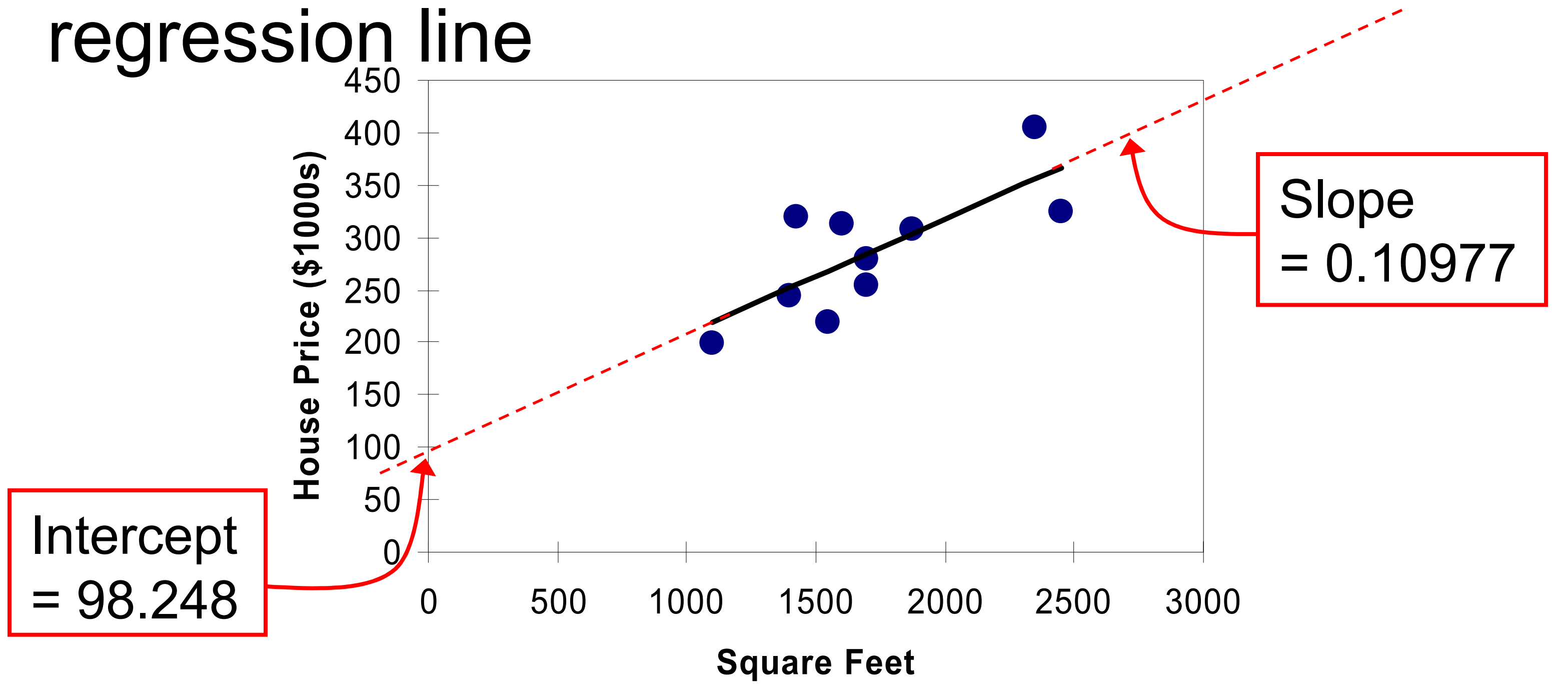
ANOVA

	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



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

Interpretation of the Intercept, b_0

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

- 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 $\boxed{b_0 = 98.24833}$ 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

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

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $\boxed{b_1 = .10977}$ 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

Predictions using Regression Analysis

Predict the price for a house with 2000 square feet:

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

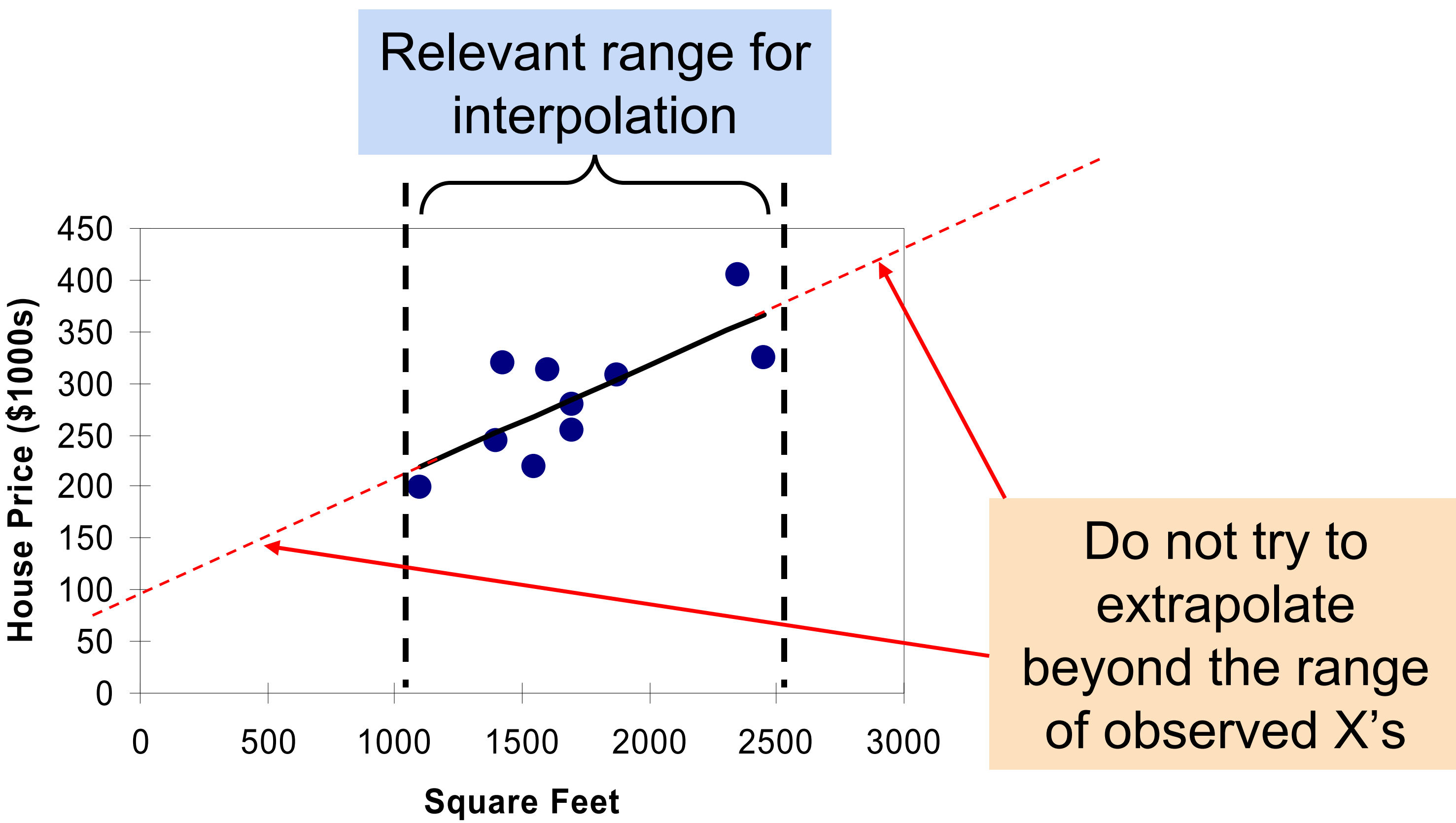
$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

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

Interpolation vs. Extrapolation

- When using a regression model for prediction, only predict within the relevant range of data



Measures of Variation

- Total variation is made up of two parts:

$$SST \quad = \quad SSR \quad + \quad SSE$$

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

where:

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

Y_i = Observed values of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value

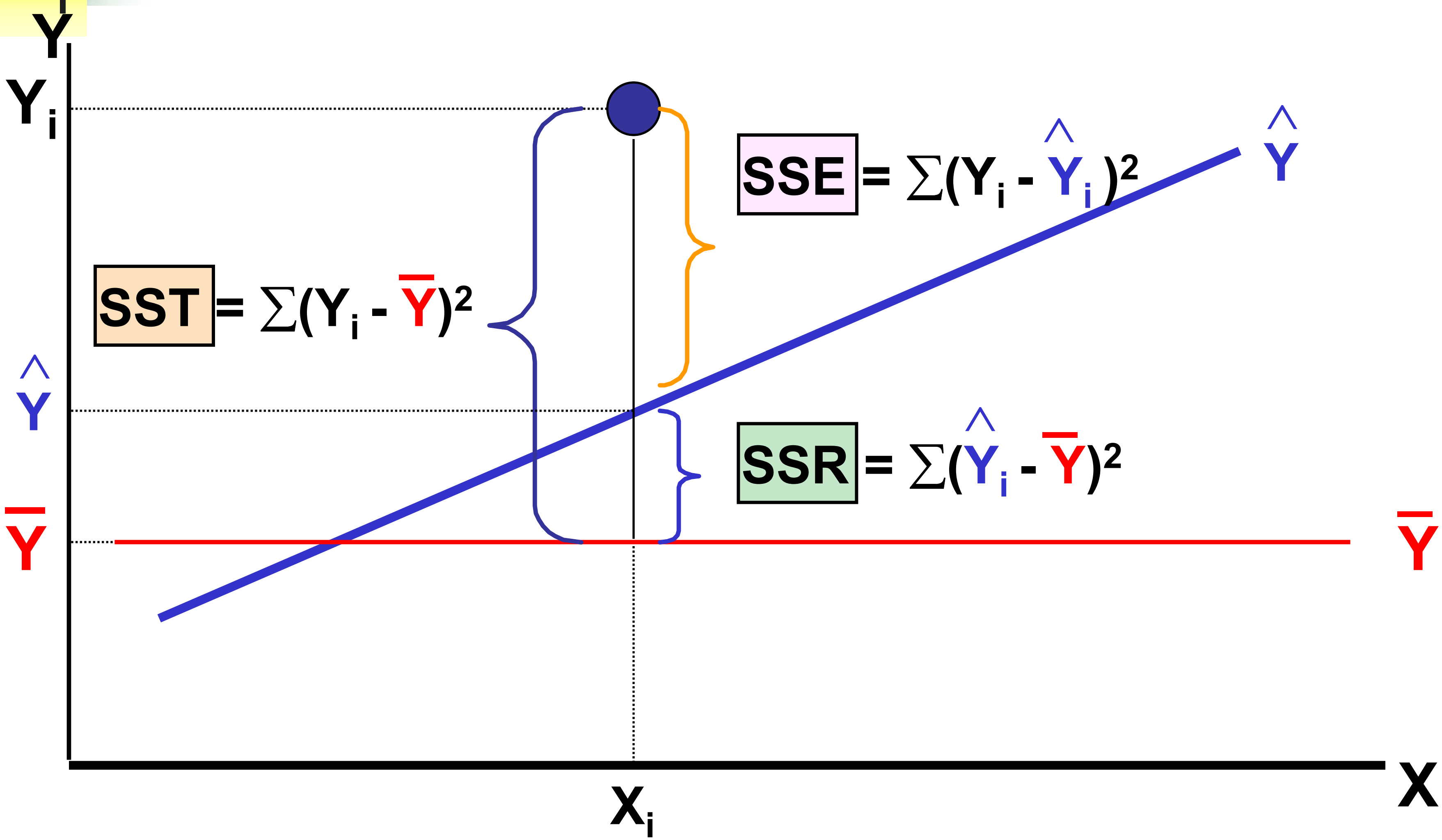
Measures of Variation

(continued)

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

Measures of Variation

(continued)



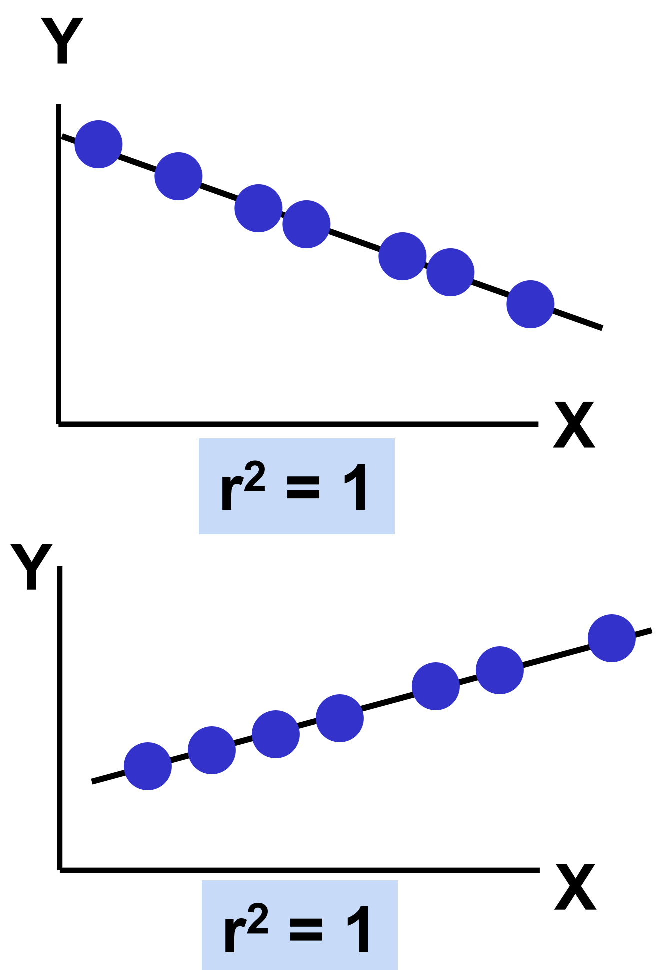
Coefficient of Determination, r^2

- 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

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

note: $\boxed{0 \le r^2 \le 1}$

Examples of Approximate r^2 Values

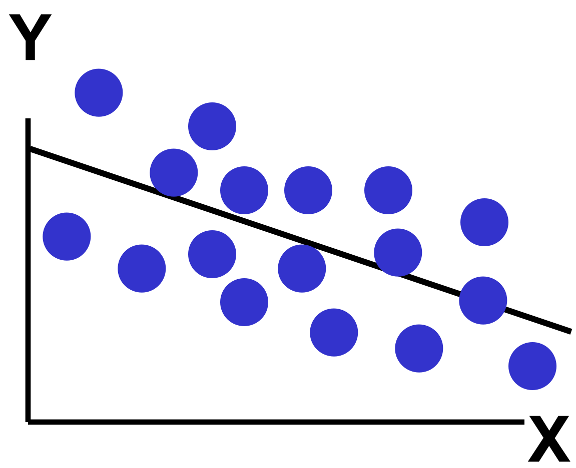


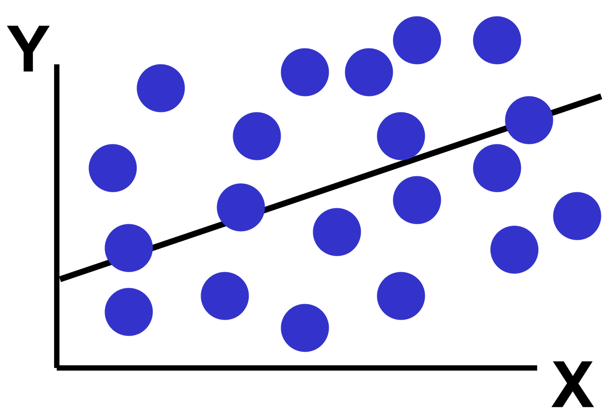
$r^2 = 1$

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



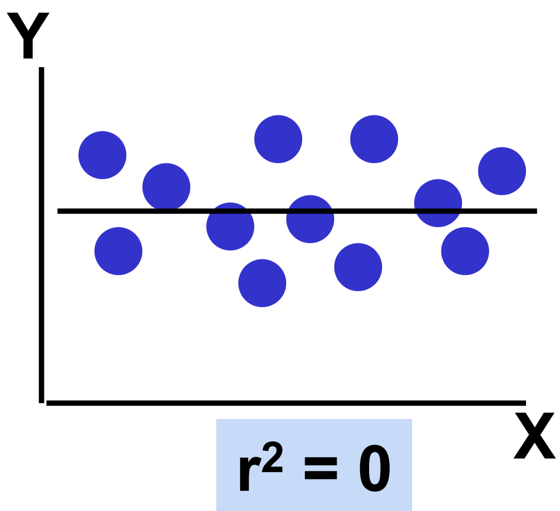


$0 < r^2 < 1$

Weaker linear relationships 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



$r^2 = 0$

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

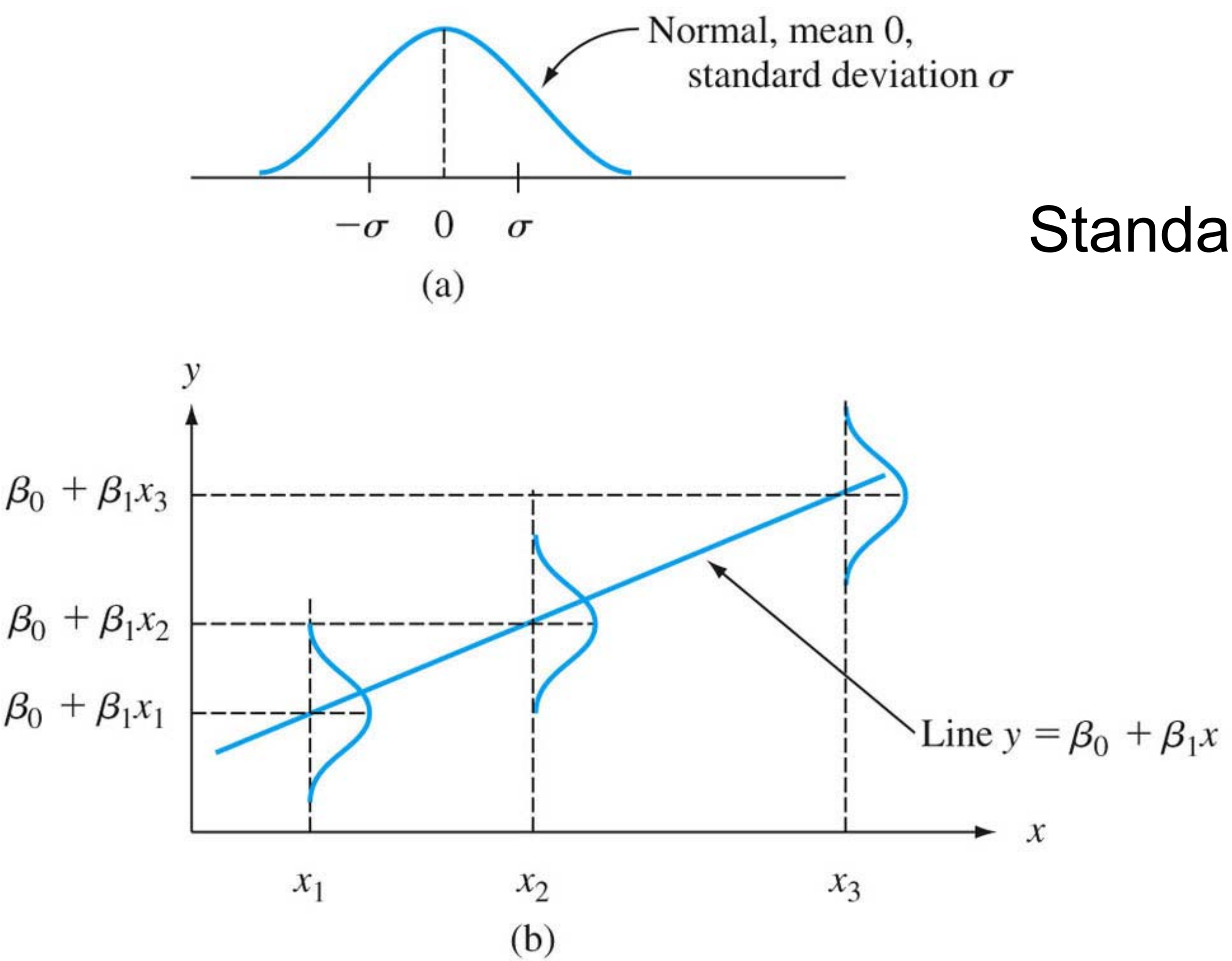
ANOVA

	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

(continued)



Standard Error Variance

$$\varepsilon_i \sim N(0, \sigma^2)$$

Standard Error of Estimate

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

$$\hat{\sigma} \longrightarrow \boxed{S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}}$$

Where

SSE = error sum of squares

n = sample size

Excel Output

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

$$S_{YX} = 41.33032$$

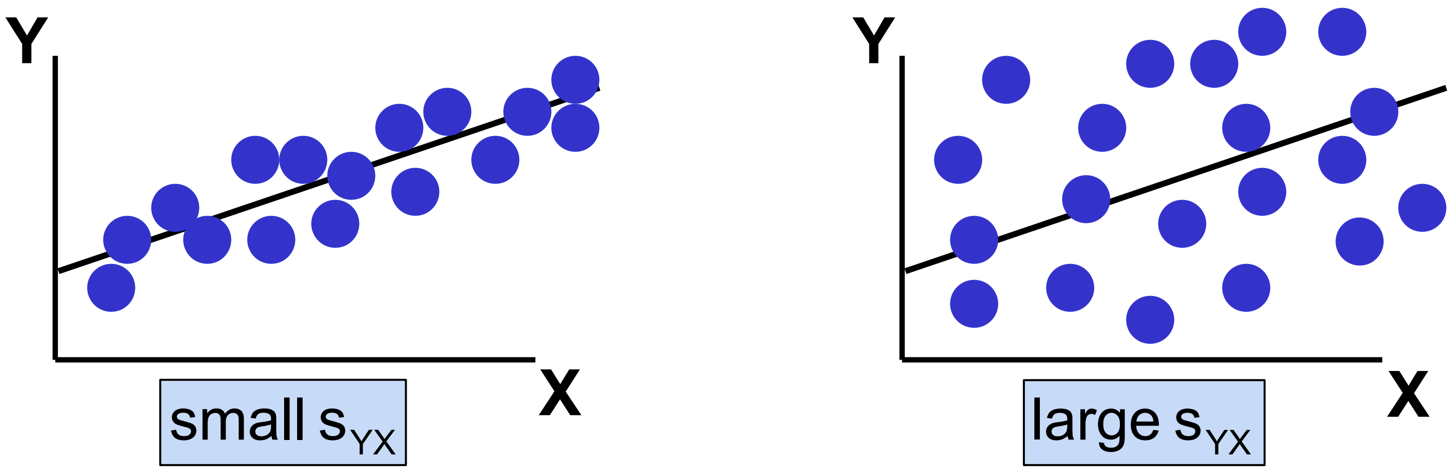
ANOVA

	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

S_{YX} is a measure of the variation of observed Y values from the regression line



The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data

i.e., S_{YX} = \$41.33K is moderately small relative to house prices in the \$200 - \$300K range

Assumptions of Regression

Use the acronym LINE:

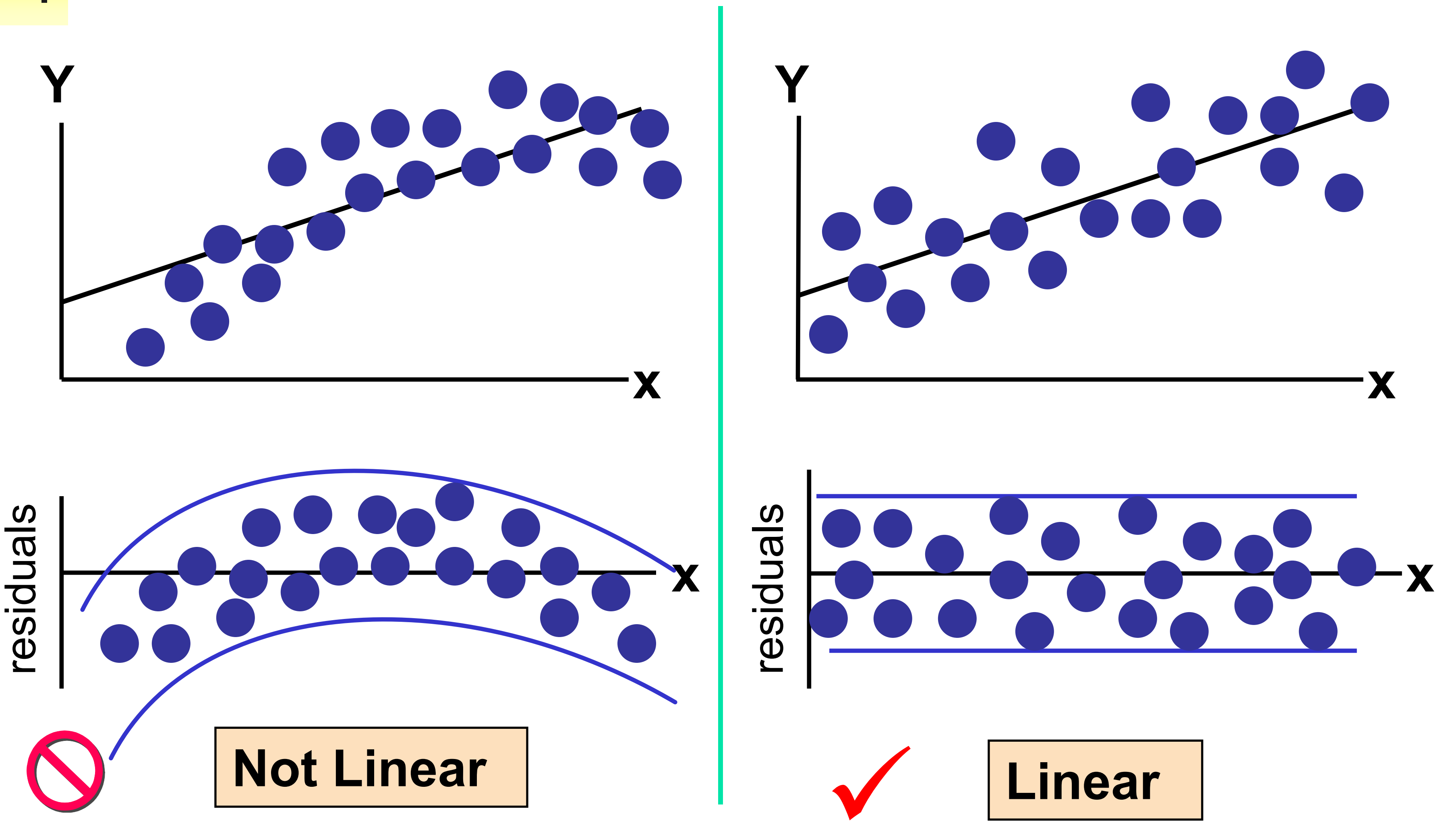
- **L**inearity
  - The underlying relationship between X and Y is linear
- **I**ndependence of Errors
  - Error values are statistically independent
- **N**ormality of Error
  - Error values (ε) are normally distributed for any given value of X
- **E**qual Variance (Homoscedasticity)
  - The probability distribution of the errors has constant variance

Residual Analysis

$$e_i = Y_i - \hat{Y}_i$$

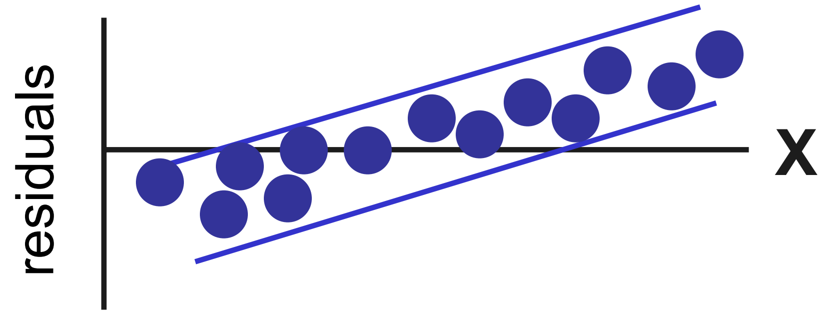
- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
- Graphical Analysis of Residuals
  - Can plot residuals vs. X

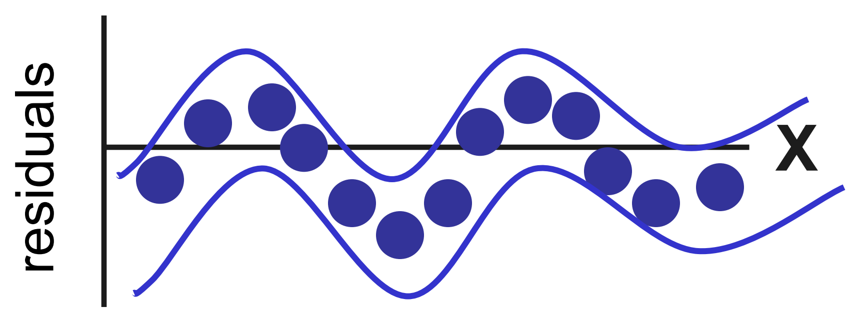
Residual Analysis for Linearity



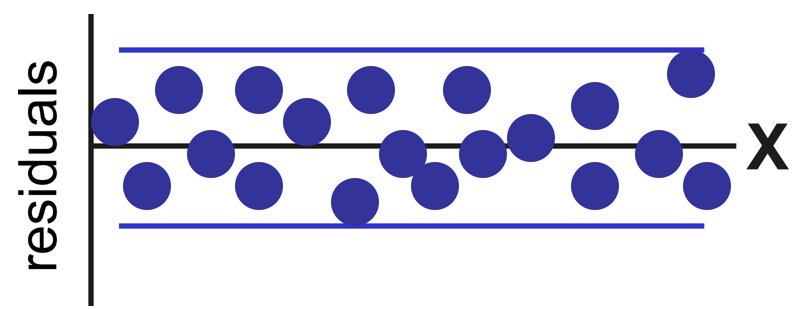
Residual Analysis for Independence

Not Independent



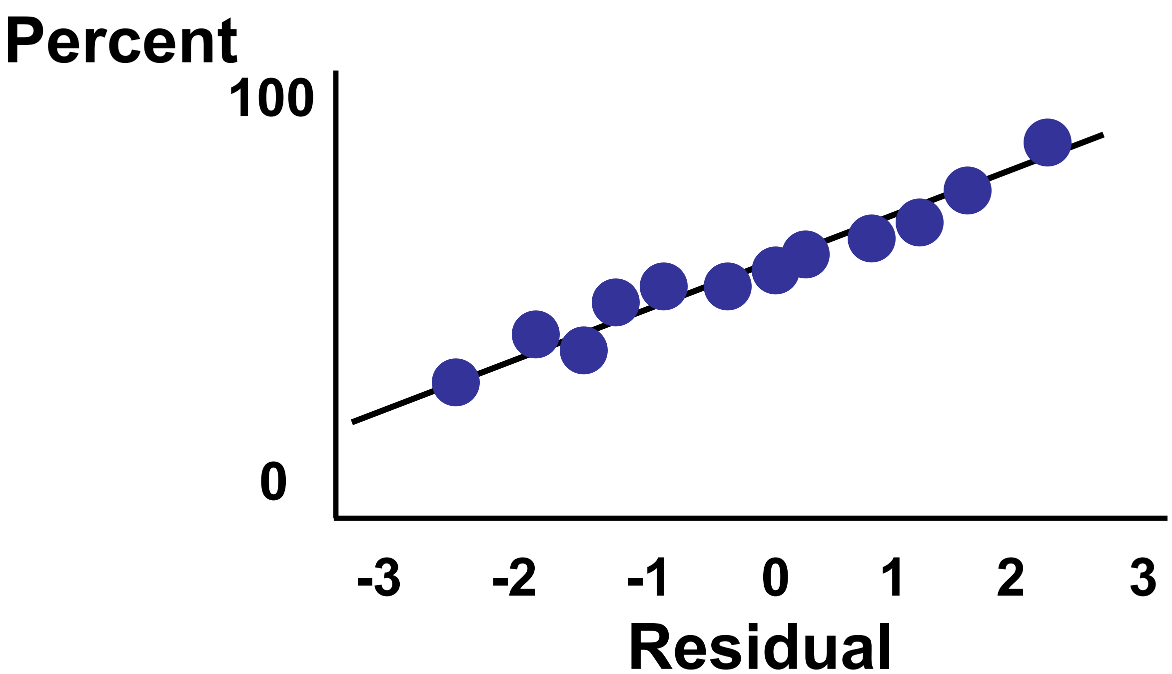


Independent

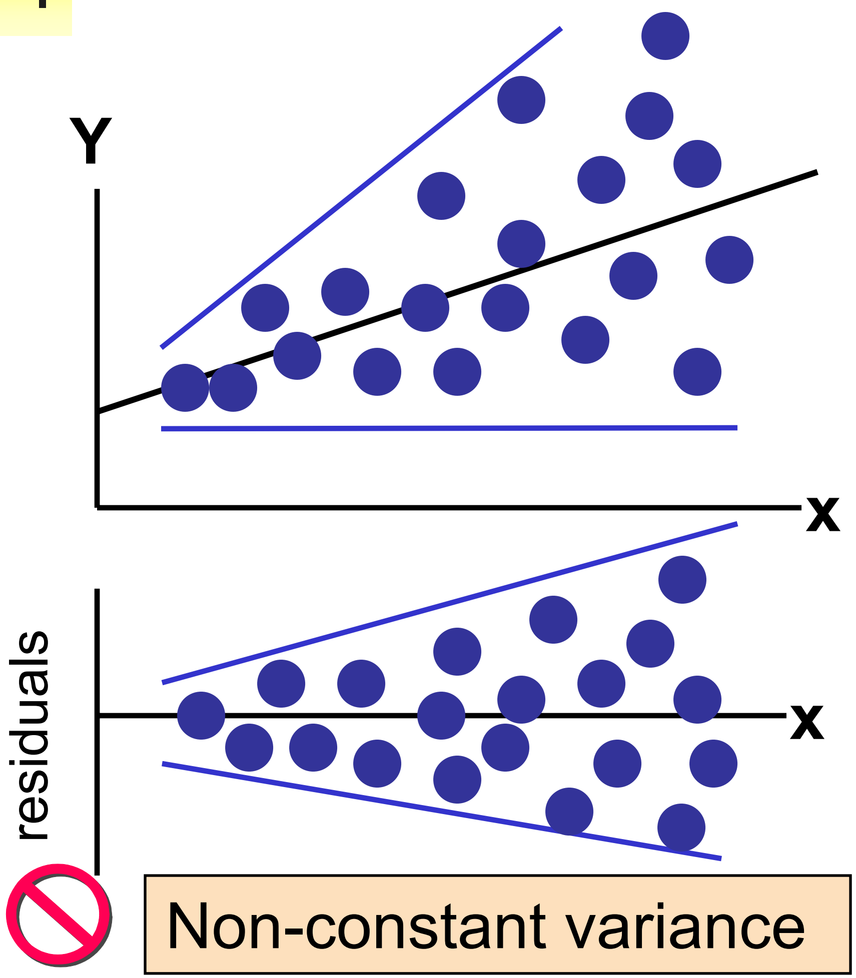


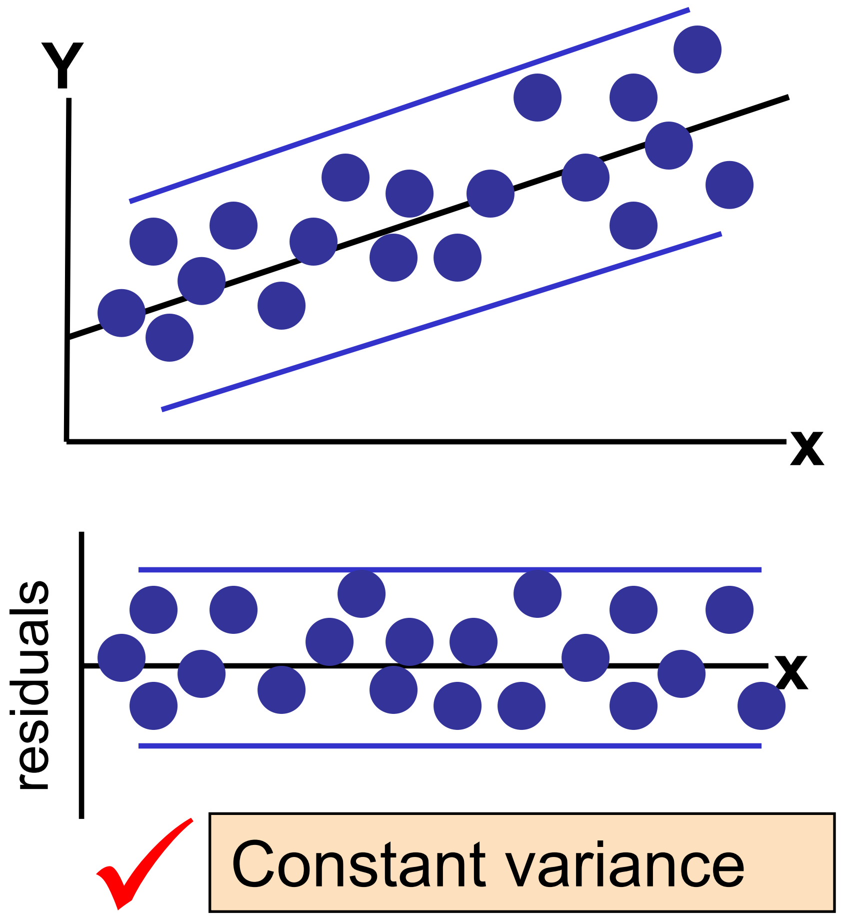
Residual Analysis for Normality

- A normal probability plot of the residuals can be used to check for normality:



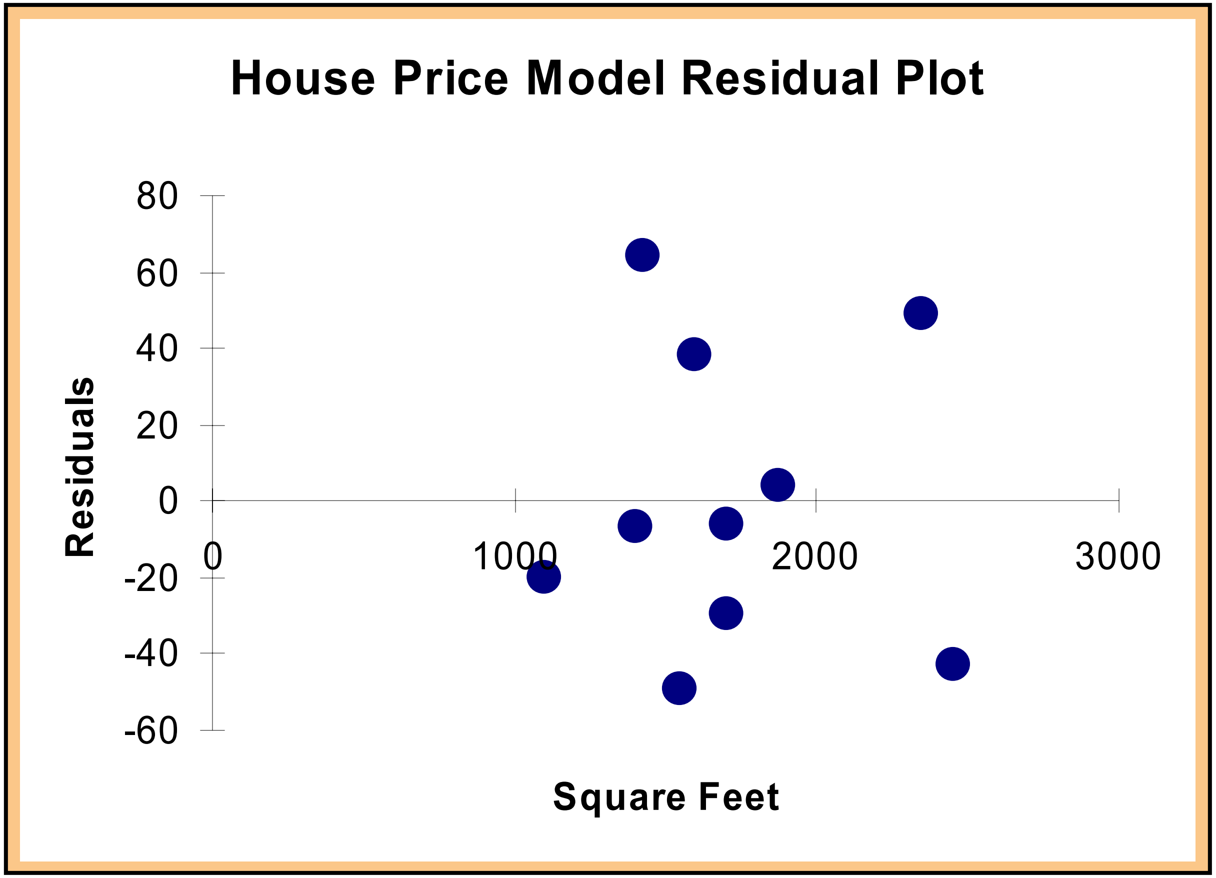
Residual Analysis for Equal Variance





Excel Residual 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



Does not appear to violate any regression assumptions

Inferences About the Slope

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

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SS_{XX}}} = \frac{S_{YX}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

where:

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

$S_{YX} = \sqrt{\frac{SSE}{n-2}}$ = 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**

$$S_{b_1} = 0.03297$$

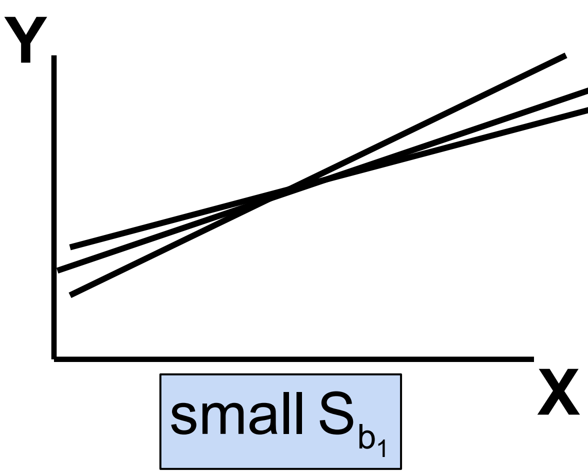
ANOVA

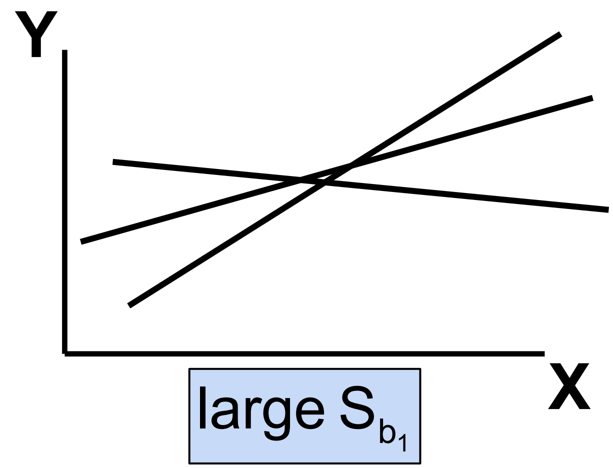
	df	*SS*	*MS*	*F*	*Significance F*
Regression	**1**	**18934.9348**	**18934.9348**	**11.0848**	**0.01039**
Residual	**8**	**13665.5652**	**1708.1957**		
Total	**9**	**32600.5000**			

	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 of the Slope

S_{b_1} is a measure of the 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

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

$$d.f. = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b_1} = 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

Simple Linear 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

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

From Excel output:

b_1 S_{b_1}

	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**

$$t = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferences about the Slope: t Test Example

(continued)

Test Statistic: t = 3.329

H_0: $\beta_1 = 0$

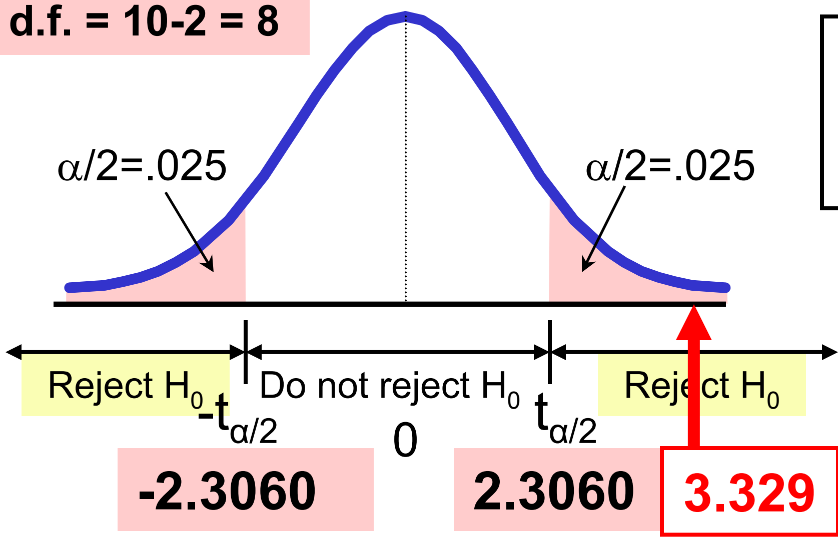
H_1: $\beta_1 \neq 0$

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

Inferences about the Slope: t Test Example

(continued)

P-value = **0.01039**

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

From Excel output:

P-value

	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

This is a two-tail test, so the p-value is

$P(t > 3.329) + P(t < -3.329) = 0.01039$

(for 8 d.f.)

Decision: P-value < α so Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

F Test for Significance

- F Test statistic: $$\boxed{F = \frac{MSR}{MSE}}$$

where

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n - k - 1}$$

where F follows an F distribution with k numerator and (n – k - 1) denominator degrees of freedom

(k = the number of independent variables in the regression model)

Excel Output

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

$$F = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$

With 1 and 8 degrees of freedom

P-value for the F Test

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**

F Test for Significance

(continued)

$H_0: \beta_1 = 0$

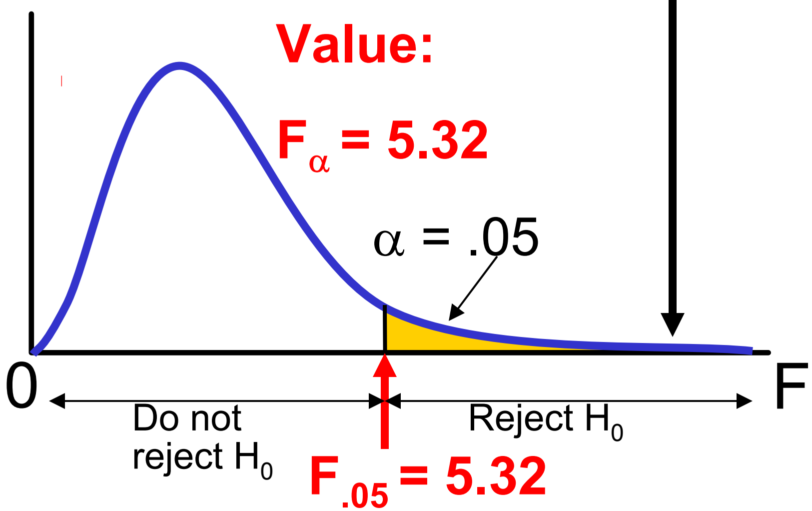
$H_1: \beta_1 \neq 0$

$\alpha = .05$

$$\underbrace{df_1 = 1 \qquad df_2 = 8}$$

Critical Value:

$\mathbf{F_\alpha = 5.32}$



Test Statistic:

$$F = \frac{MSR}{MSE} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{n-2} S_{b_1}$$

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)

Confidence Interval Estimate for the Slope

(continued)

	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

The Sample Covariance

- The sample covariance measures the strength of the linear relationship between **two variables** (called bivariate data)

- The sample covariance:

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

  - Only concerned with the strength of the relationship
  - No causal effect is implied

Interpreting Covariance

- **Covariance** between two random variables:

cov(X,Y) > 0 $\longrightarrow$ X and Y tend to move in the same direction

cov(X,Y) < 0 $\longrightarrow$ X and Y tend to move in opposite directions

cov(X,Y) = 0 $\longrightarrow$ X and Y are independent

Coefficient of Correlation

- Measures the relative strength of the linear relationship between two variables
- Sample coefficient of correlation:

$$r = \frac{\operatorname{cov}(X,Y)}{S_X S_Y}$$

where

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

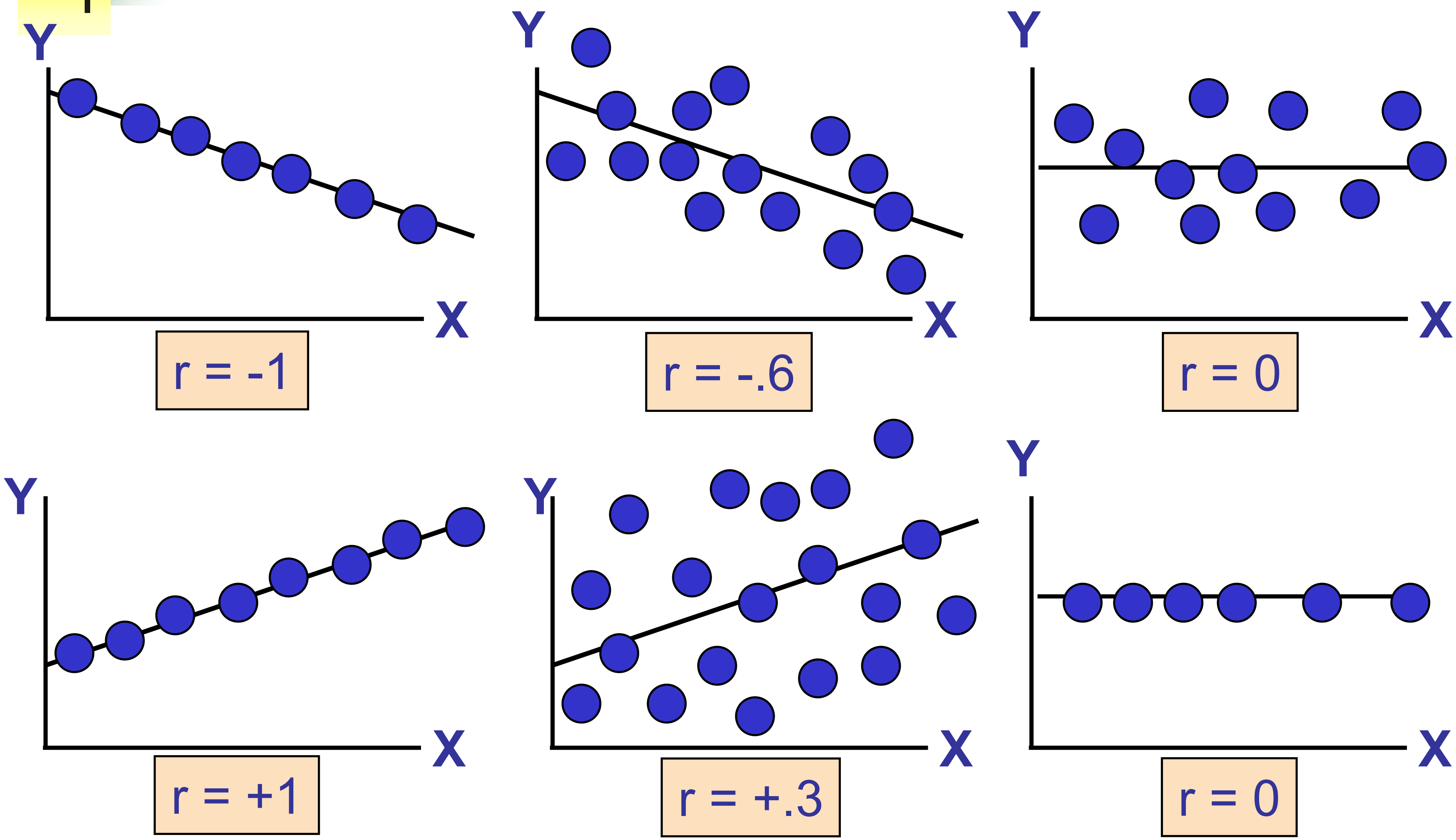
$$S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}$$

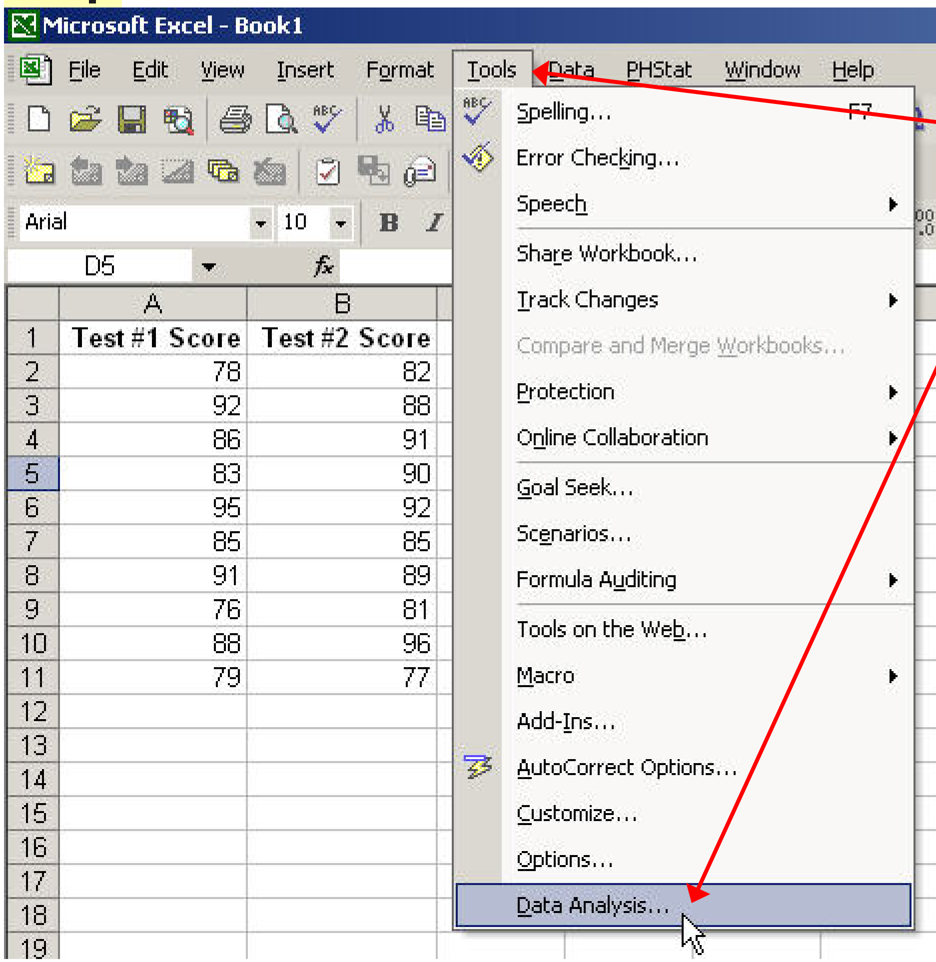
Features of Correlation Coefficient, r

- Unit free
- Ranges 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

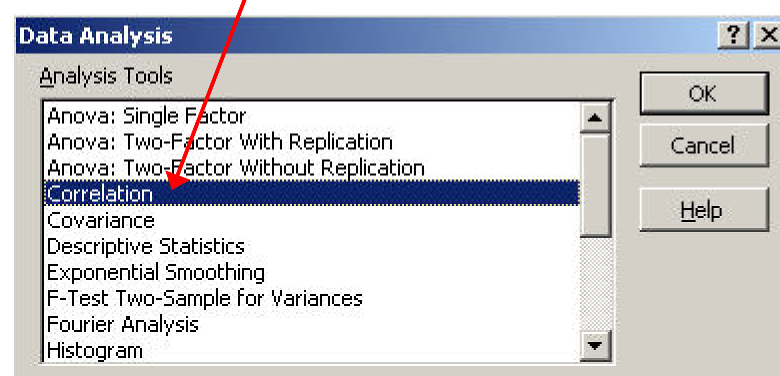
Scatter Plots of Data with Various Correlation Coefficients



Using Excel to Find the Correlation Coefficient

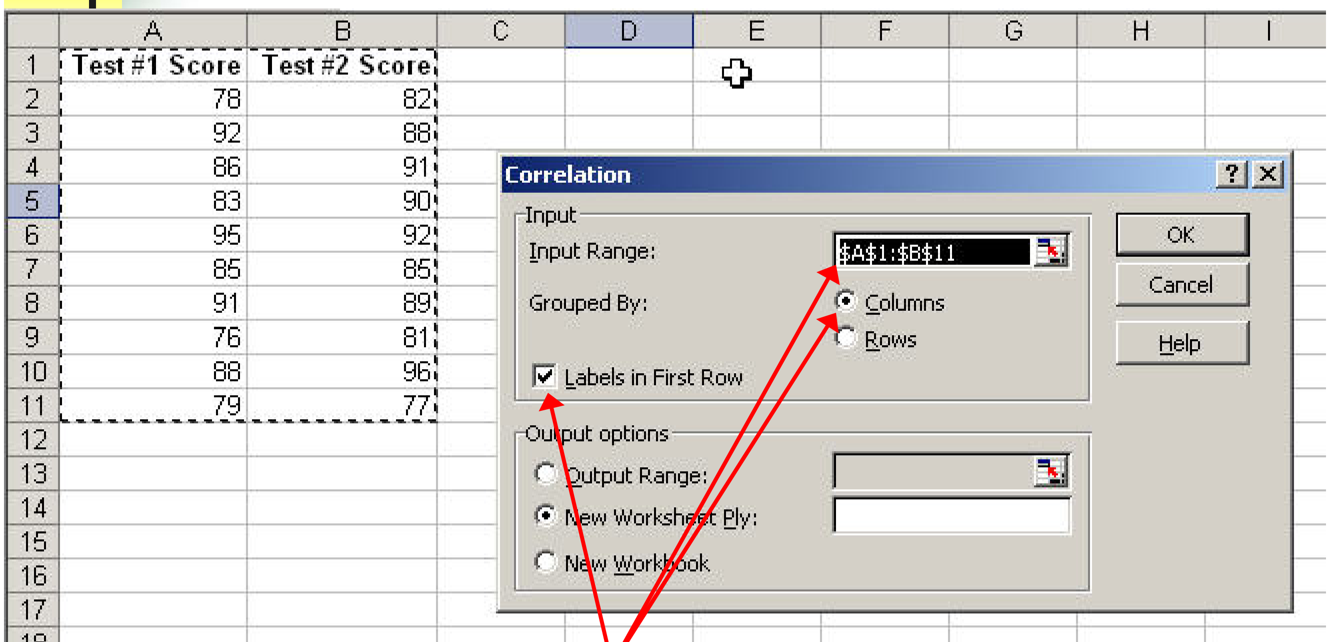


- Select **Tools/Data Analysis**
- Choose **Correlation** from the selection menu
- Click OK . . .

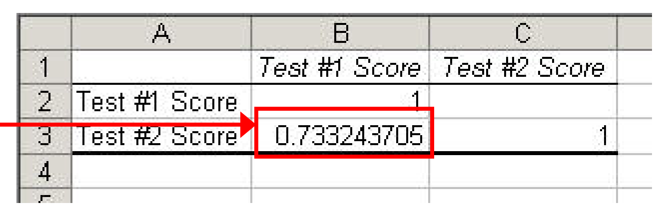


Using Excel to Find the Correlation Coefficient

(continued)



- Input data range and select appropriate options
- Click OK to get output

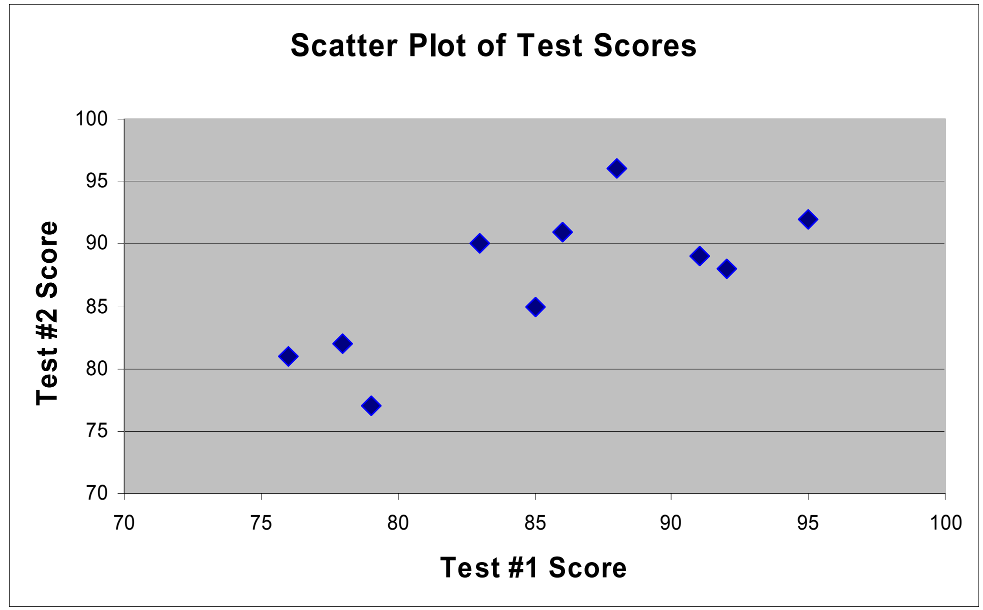


	A	B	C
1		*Test #1 Score*	*Test #2 Score*
2	Test #1 Score	1	
3	Test #2 Score	0.733243705	1
4			

Interpreting the Result

- r = .733

- There is a relatively strong positive linear relationship between test score #1 and test score #2



- Students who scored high on the first test tended to score high on second test, and students who scored low on the first test tended to score low on the second test

t Test for a Correlation Coefficient

- Hypotheses

$H_0: \rho = 0$	(no correlation between X and Y)
$H_A: \rho \neq 0$	(correlation exists)

- Test statistic
  - $$t = \frac{r - \rho}{\sqrt{\dfrac{1 - r^2}{n - 2}}}$$

(with n – 2 degrees of freedom)

where

$r = +\sqrt{r^2}$ if $b_1 > 0$

$r = -\sqrt{r^2}$ if $b_1 < 0$

Example: House Prices

Is there evidence of a linear relationship between square feet and house price at the .05 level of significance?

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

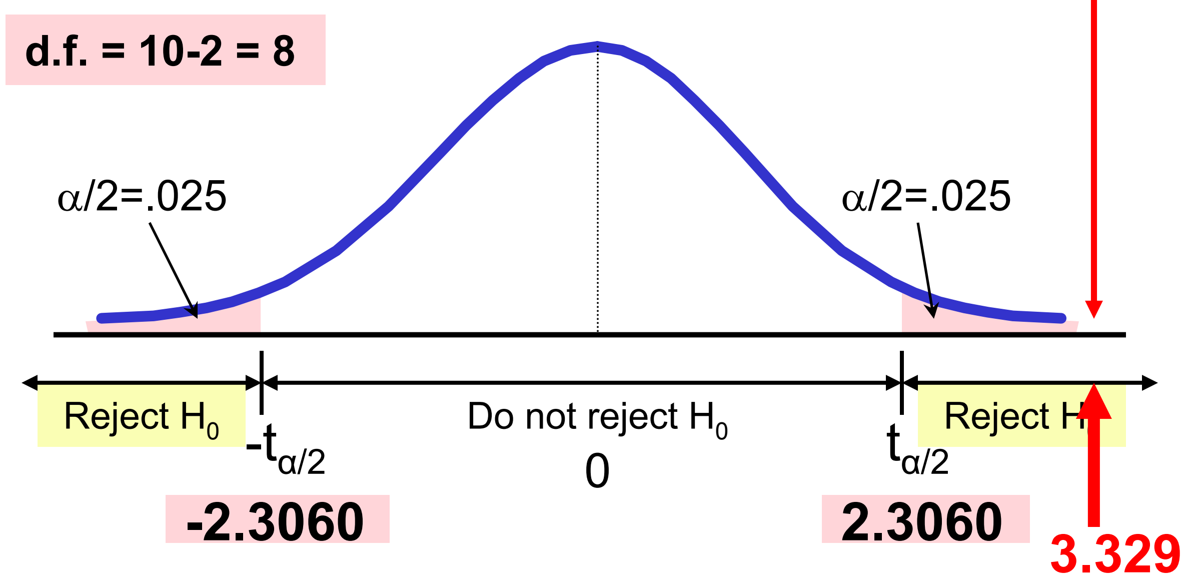
H_1: $\rho \neq 0$ (correlation exists)

α =.05 , df = 10 - 2 = 8

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.762 - 0}{\sqrt{\frac{1 - .762^2}{10 - 2}}} = 3.329$$

Example: Test Solution

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.762 - 0}{\sqrt{\frac{1 - .762^2}{10 - 2}}} = 3.329$$

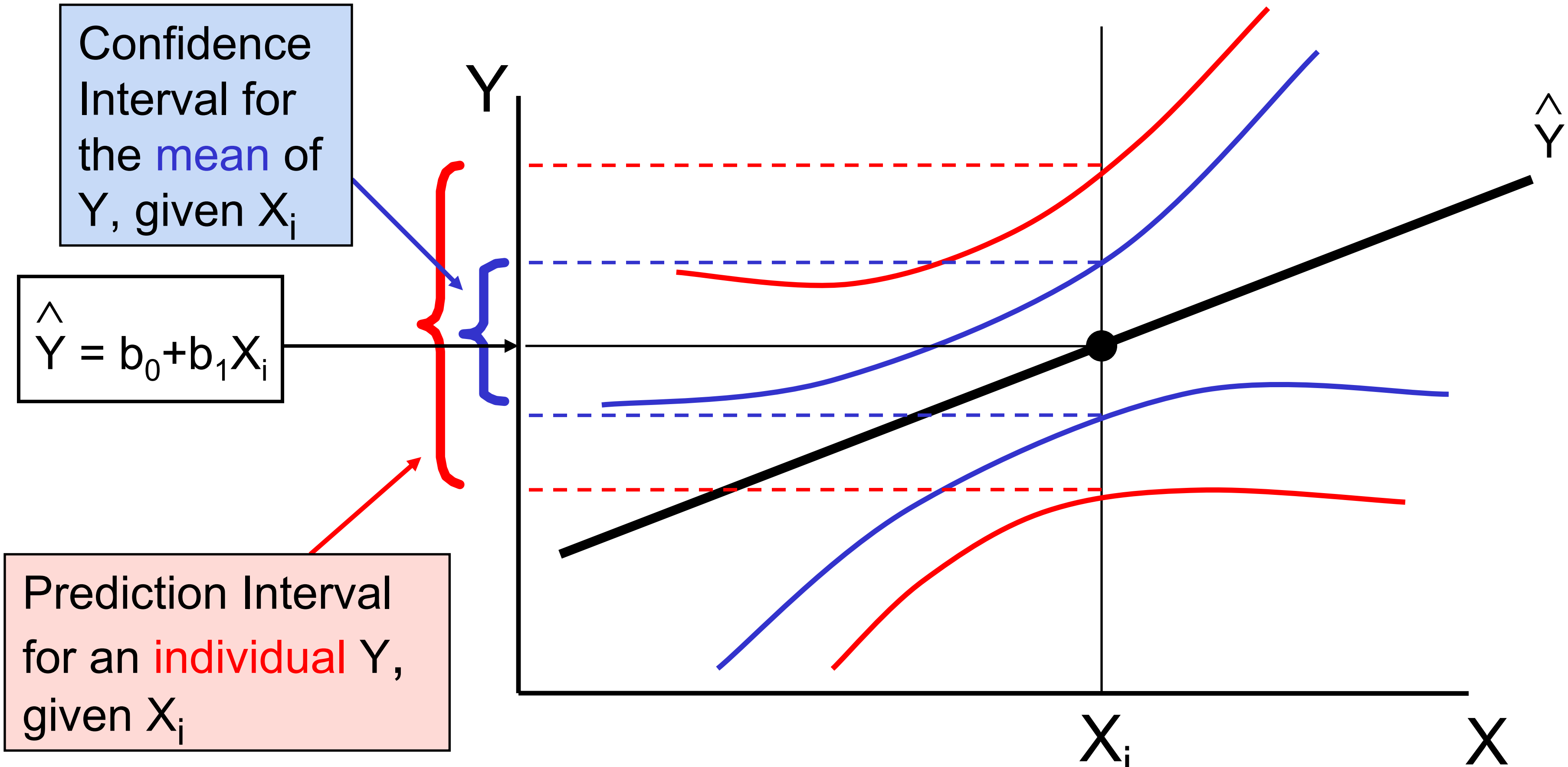


Decision:
Reject H_0

Conclusion:
There **is evidence** of a linear association at the 5% level of significance

Estimating Mean Values and Predicting Individual Values

Goal: Form intervals around Y to express uncertainty about the value of Y for a given X_i



Confidence Interval for the Average Y, Given X

Confidence interval estimate for the **mean value of Y** given a particular X_i

Confidence interval for $\mu_{Y|X=X_i}$:

$$\hat{Y} \pm t_{n-2} S_{YX} \sqrt{h_i}$$

Size of interval varies according to distance away from mean, $\overline{X}$

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{SSX} = \frac{1}{n} + \underbrace{\frac{(X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2}}$$

Prediction Interval for an Individual Y, Given X

Confidence interval estimate for an **Individual value of Y** given a particular X_i

Confidence interval for $Y_{X=X_i}$:

$$\hat{Y} \pm t_{n-2} S_{YX} \sqrt{1 + h_i}$$

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

Estimation of Mean Values: Example

Confidence Interval Estimate for $\mu_{Y|X=X_i}$

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

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

$$\hat{Y} \pm t_{n-2} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 37.12$$

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

Estimation of Individual Values: Example

Prediction Interval Estimate for $Y_{X=X_i}$

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

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

$$\hat{Y} \pm t_{n-1} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2}} = 317.85 \pm 102.28$$

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

Finding Confidence and Prediction Intervals in Excel

- In Excel, use

  PHStat | regression | simple linear regression …

  - Check the
    “confidence and prediction interval for X=”
    box and enter the X-value and confidence level desired

Finding Confidence and Prediction Intervals in Excel

(continued)

	A	B
1	**Confidence Interval Estimate**	
2		
3	**Data**	
4	**X Value**	**2000**
5	**Confidence Level**	**95%**
6		
7	Intermediate Calculations	
8	Sample Size	10
9	Degrees of Freedom	8
10	t Value	2.306006
11	Sample Mean	1715
12	Sum of Squared Difference	1571500
13	Standard Error of the Estimate	41.33032
14	h Statistic	0.151686
15	Average Predicted Y (YHat)	317.7838
16		
17	**For Average Predicted Y (YHat)**	
18	Interval Half Width	37.11952
19	**Confidence Interval Lower Limit**	**280.6643**
20	**Confidence Interval Upper Limit**	**354.9033**
21		
22	**For Individual Response Y**	
23	Interval Half Width	102.2813
24	**Prediction Interval Lower Limit**	**215.5025**
25	**Prediction Interval Upper Limit**	**420.0651**

Input values

$\hat{Y}$

Confidence Interval Estimate for $\mu_{Y|X=X_i}$

Prediction Interval Estimate for $Y_{X=X_i}$

Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions underlying least-squares regression
- Not knowing how to evaluate the assumptions
- Not knowing the alternatives to least-squares regression if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range

Strategies for Avoiding the Pitfalls of Regression

- Start with a scatter diagram of X vs. Y to observe possible relationship
- Perform residual analysis to check the assumptions
  - Plot the residuals vs. X to check for violations of assumptions such as homoscedasticity
  - Use a histogram, stem-and-leaf display, box-and-whisker plot, or normal probability plot of the residuals to uncover possible non-normality

Strategies for Avoiding the Pitfalls of Regression

(continued)

- If there is violation of any assumption, use alternative methods or models
- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals
- Avoid making predictions or forecasts outside the relevant range