Business Statistic

Week 14

Linear Regression

Learning Objectives

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 plot can be used to show the relationship between two variables
- Correlation analysis is used to measure the 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
  - Scatter plots were first presented in Ch. 2
  - Correlation was first presented in Ch. 3

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

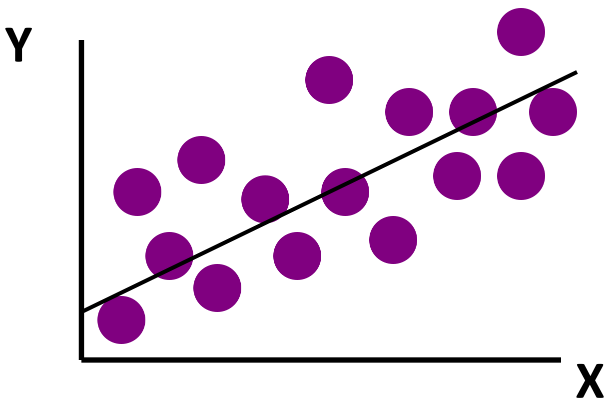
Independent variable: the variable used to predict or explain the dependent variable

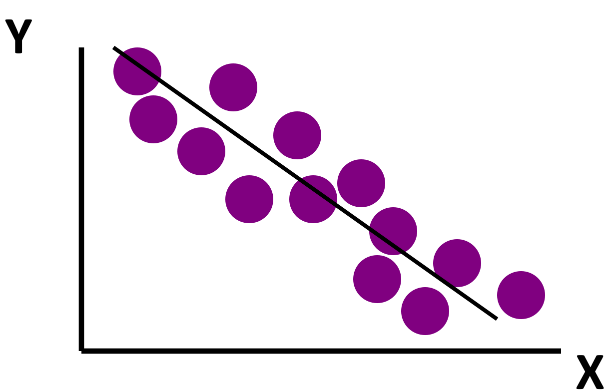
Simple Linear Regression Model

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

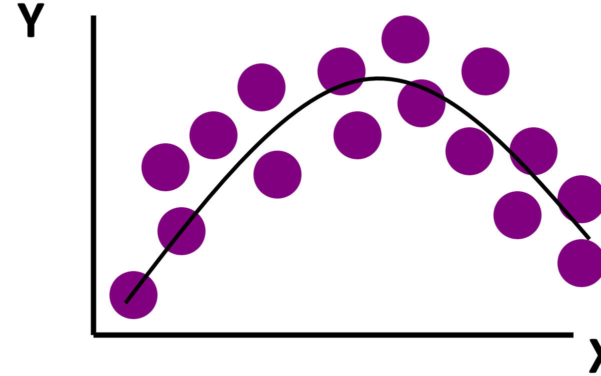
Types of Relationships

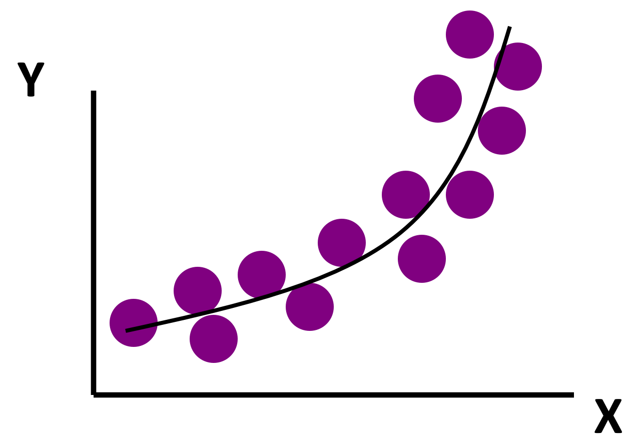
Linear relationships





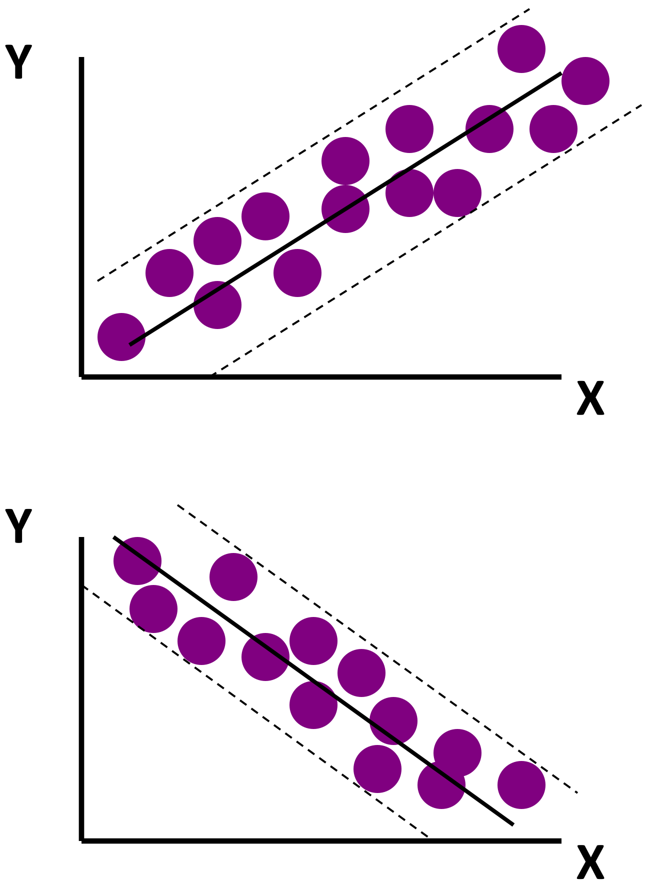
Curvilinear relationships



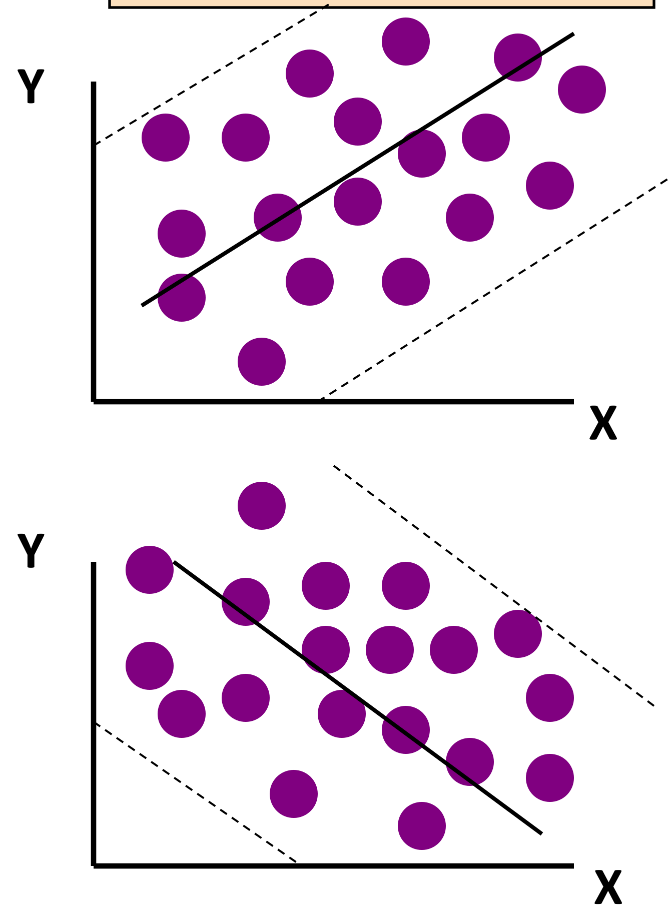


Types of Relationships

Strong relationships

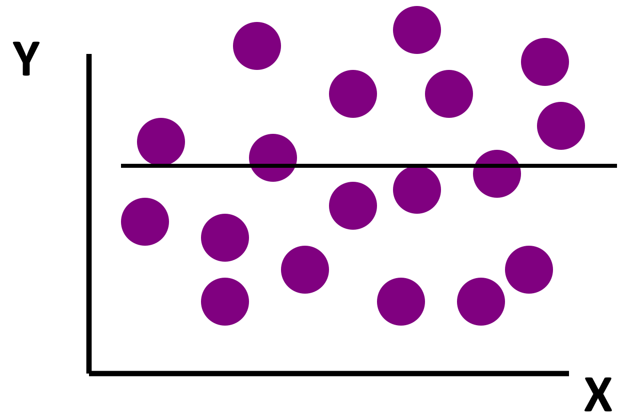


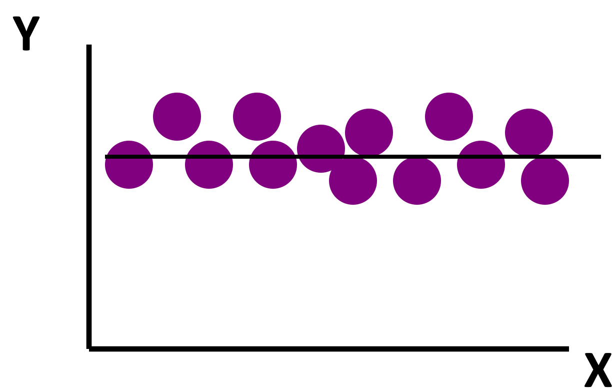
Weak relationships



Types of Relationships

No relationship





Simple Linear Regression Model

Population Y intercept

Population Slope Coefficient

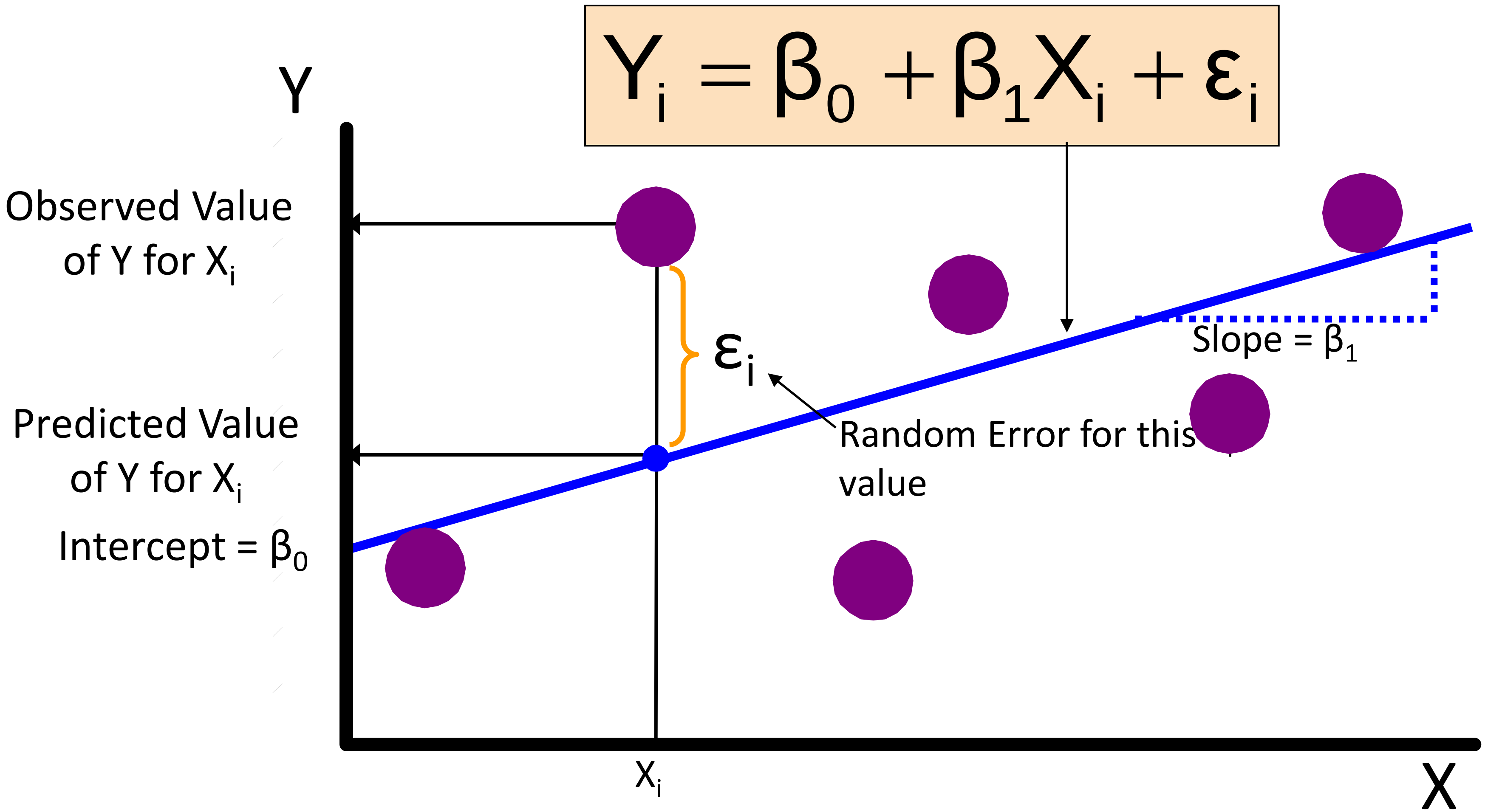
Independent Variable

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



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 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 $\hat{Y}$:

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

Linear Trend Model

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

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^{\,2} - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}$$

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

Interpretation of the Slope and the Intercept

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

- b_1 is the estimated change in the mean 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)

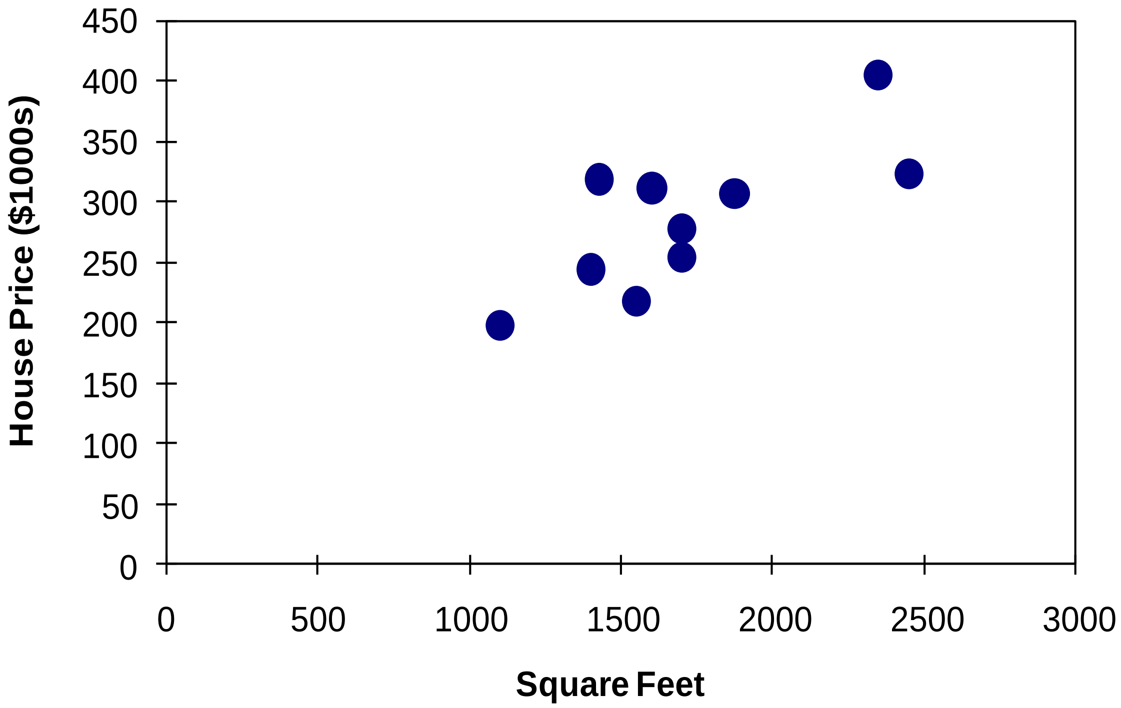
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Simple Linear Regression Example: Data

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

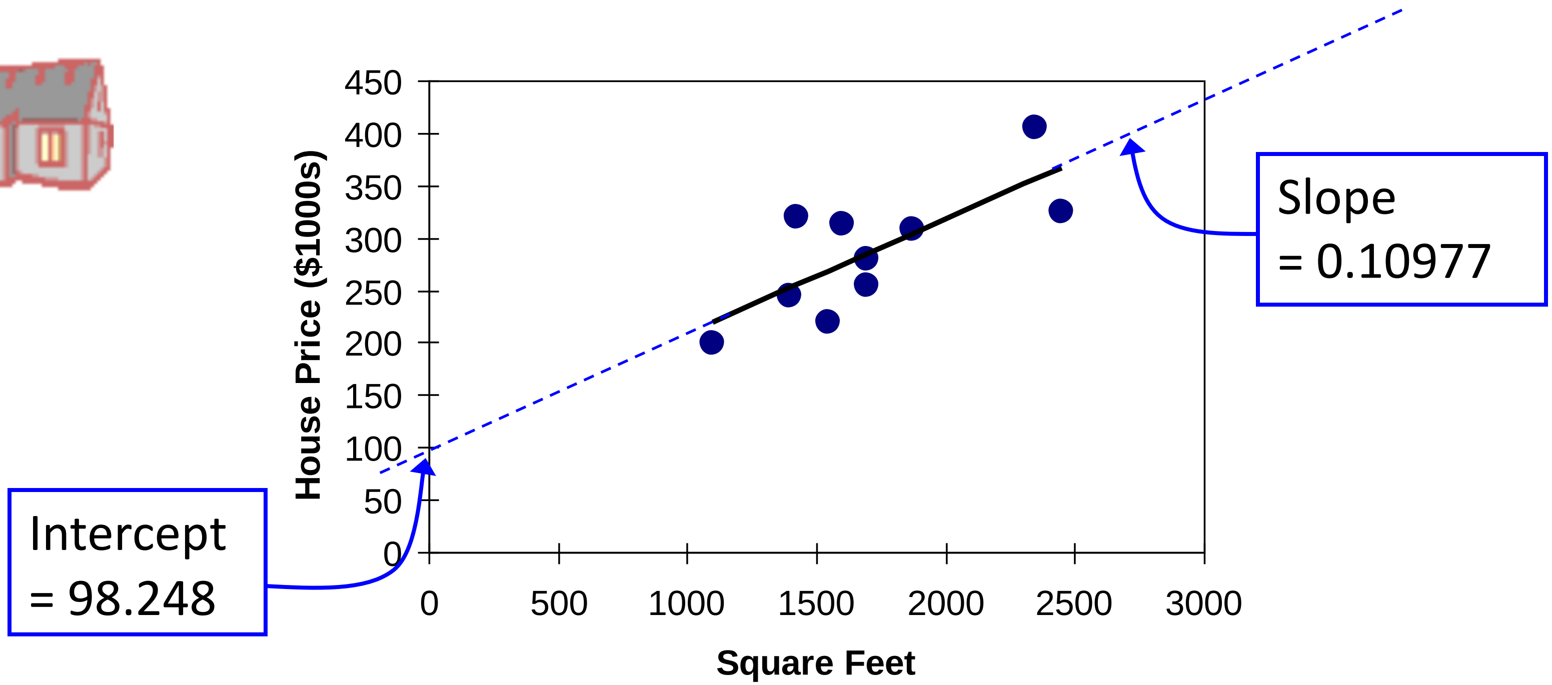
Simple Linear Regression Example: Scatter Plot

House price model: Scatter Plot



Simple Linear Regression Example: Graphical Representation

House price model: Scatter Plot and Prediction Line



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

Simple Linear Regression Example: Interpretation of b_o

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

- b_0 is the estimated mean value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
- Because a house cannot have a square footage of 0, b_0 has no practical application

Simple Linear Regression Example: Interpreting b_1

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

- b_1 estimates the change in the mean value of Y as a result of a one-unit increase in X
  - Here, b_1 = 0.10977 tells us that the mean value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Simple Linear Regression Example: Making Predictions

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{sq.ft.})$$

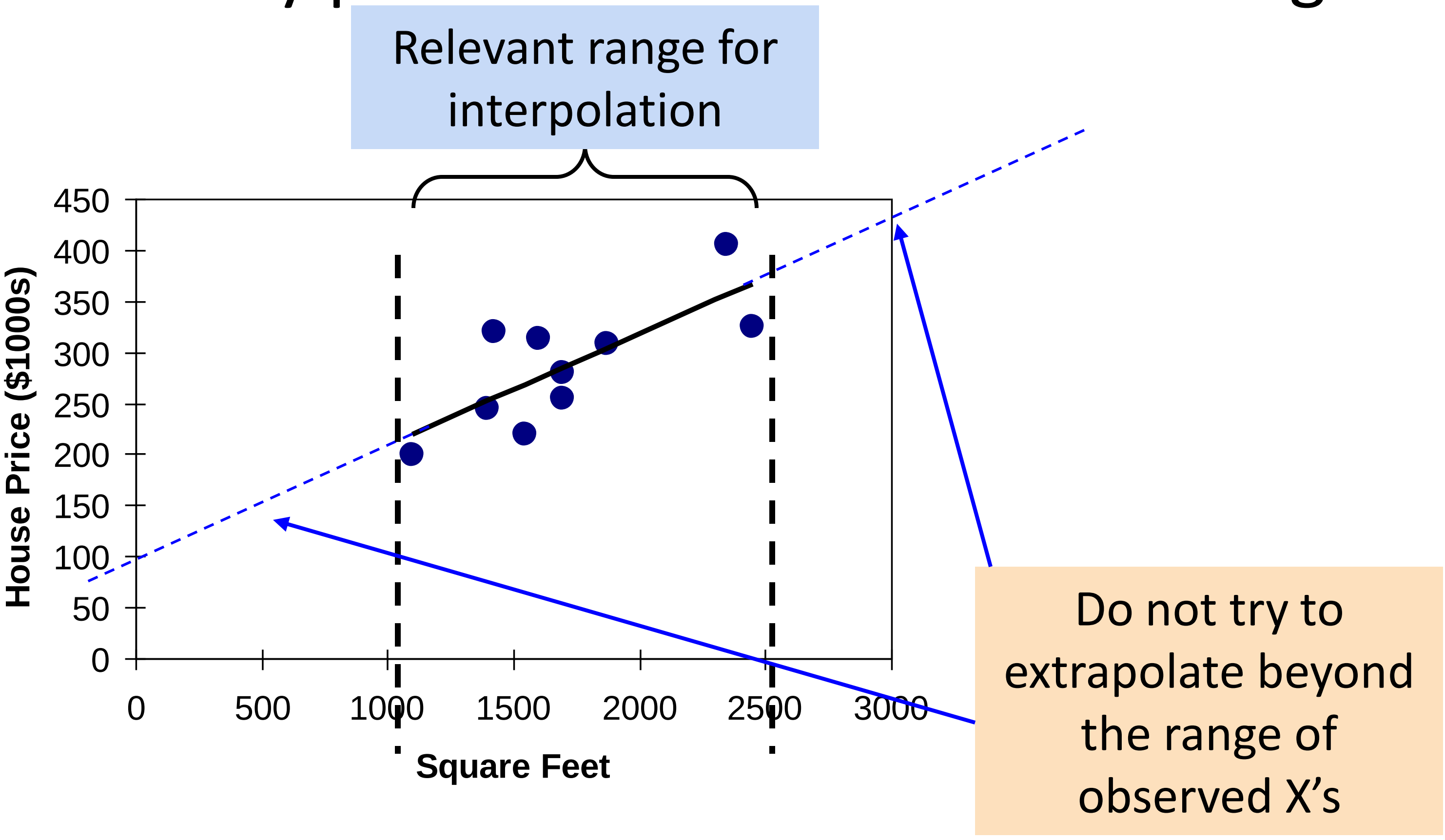
$$= 98.24833 + 0.10977(2000)$$

$$= 317.78$$

The predicted price for a house with 2000 square feet is 317.78($1,000s) = $317,780

Simple Linear Regression Example: Making Predictions

- 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:

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

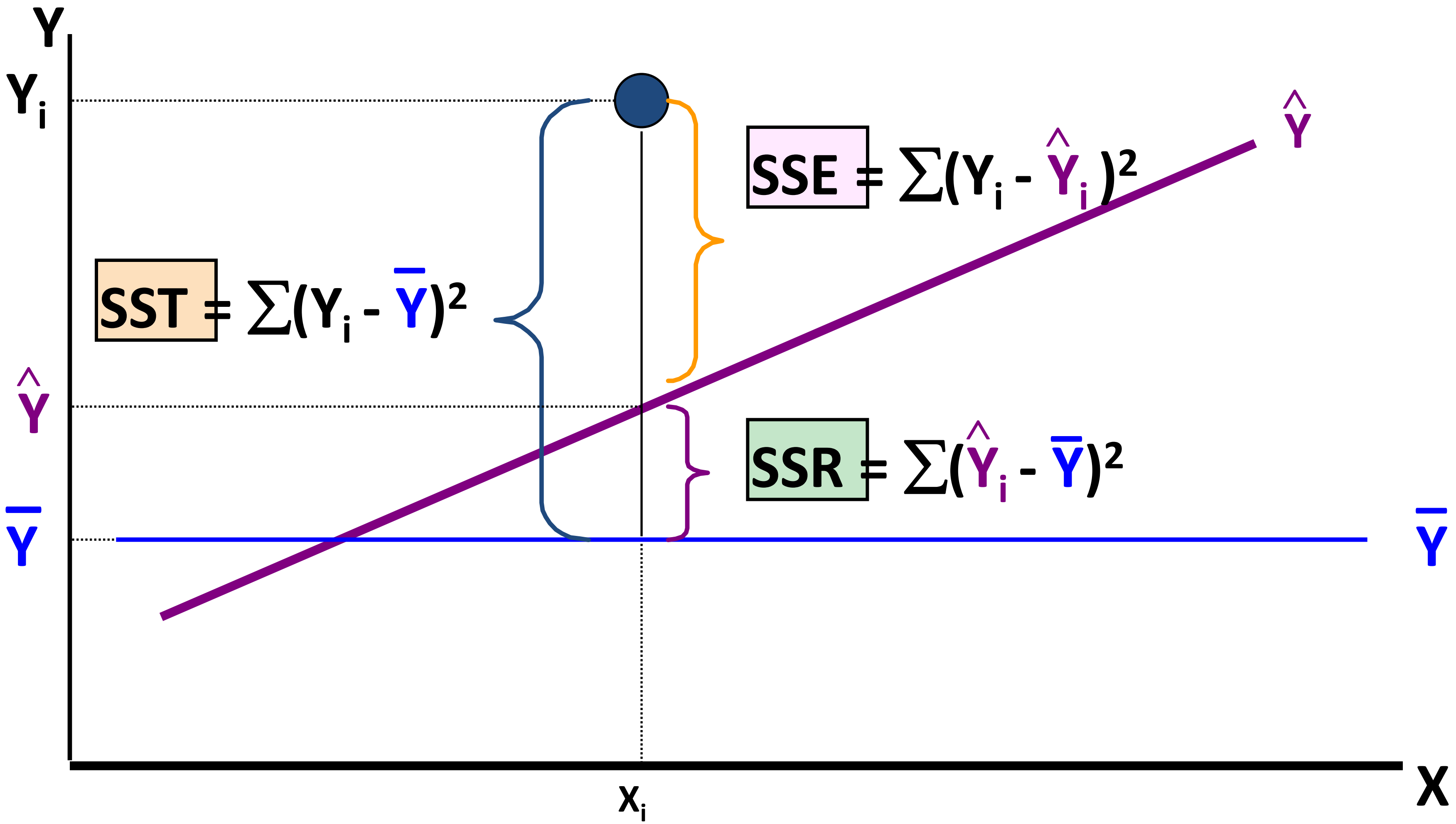
Y_i = Observed value of the dependent variable

= Predicted value of Y for the given X_i value

Measures of Variation

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

Measures of Variation



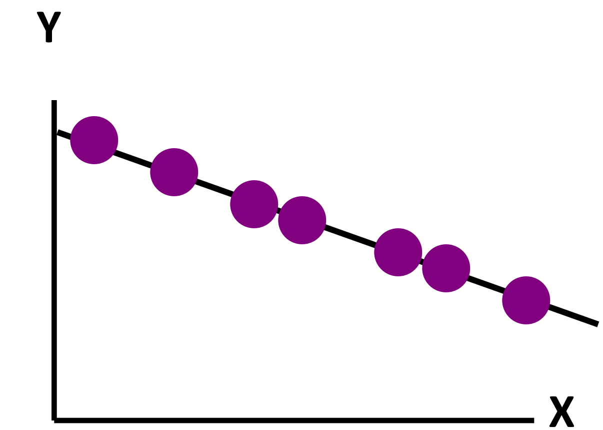
Coefficient of Determination, r^2

- 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

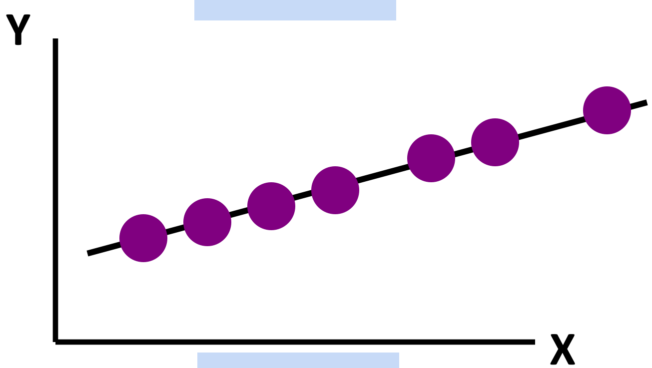
$$\boxed{r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}}$$

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

Examples of Approximate r^2 Values



$r^2 = 1$



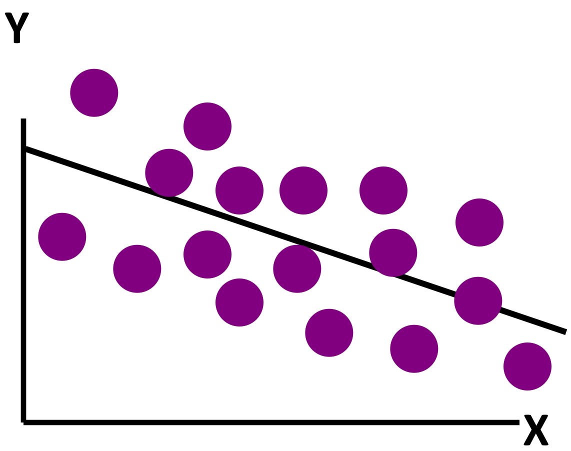
$r^2 = 1$

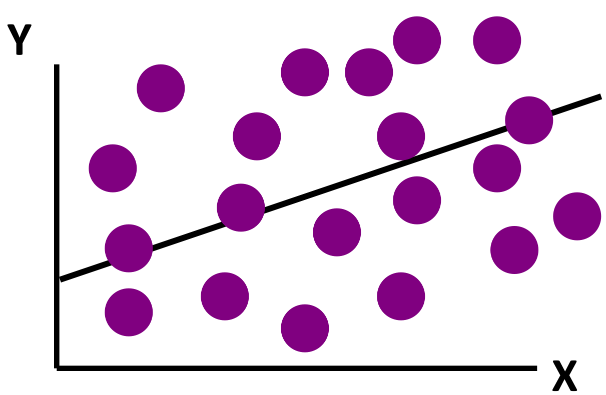
$r^2 = 1$

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



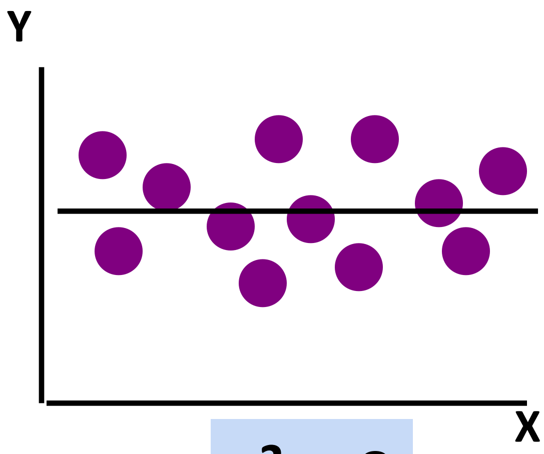


$0 < r^2 < 1$

Weaker linear 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



$r^2 = 0$

$r^2 = 0$

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)

Standard Error of Estimate

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

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

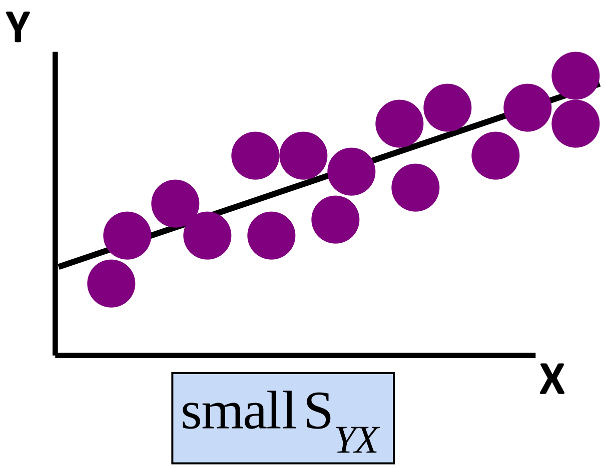
Where

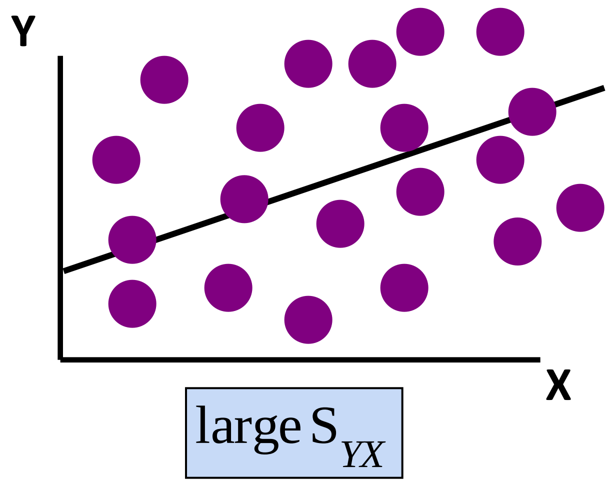
SSE = error sum of squares

n = sample size

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 \$200K - \$400K range

Assumptions of Regression
L.I.N.E

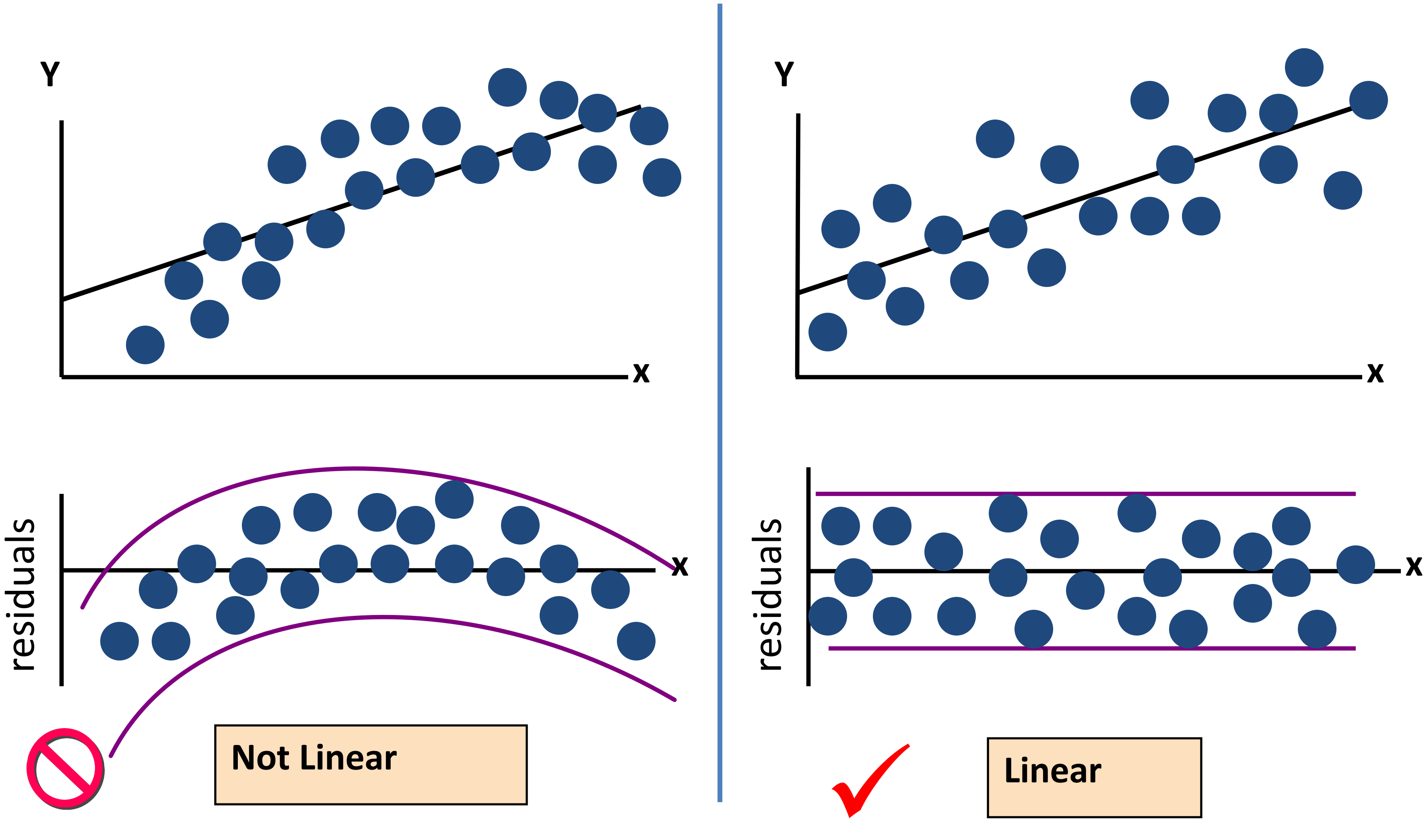
- Linearity
  - The relationship between X and Y is linear
- Independence of Errors
  - Error values are statistically independent
- Normality of Error
  - Error values are normally distributed for any given value of X
- Equal Variance (also called homoscedasticity)
  - The probability distribution of the errors has constant variance

Residual Analysis

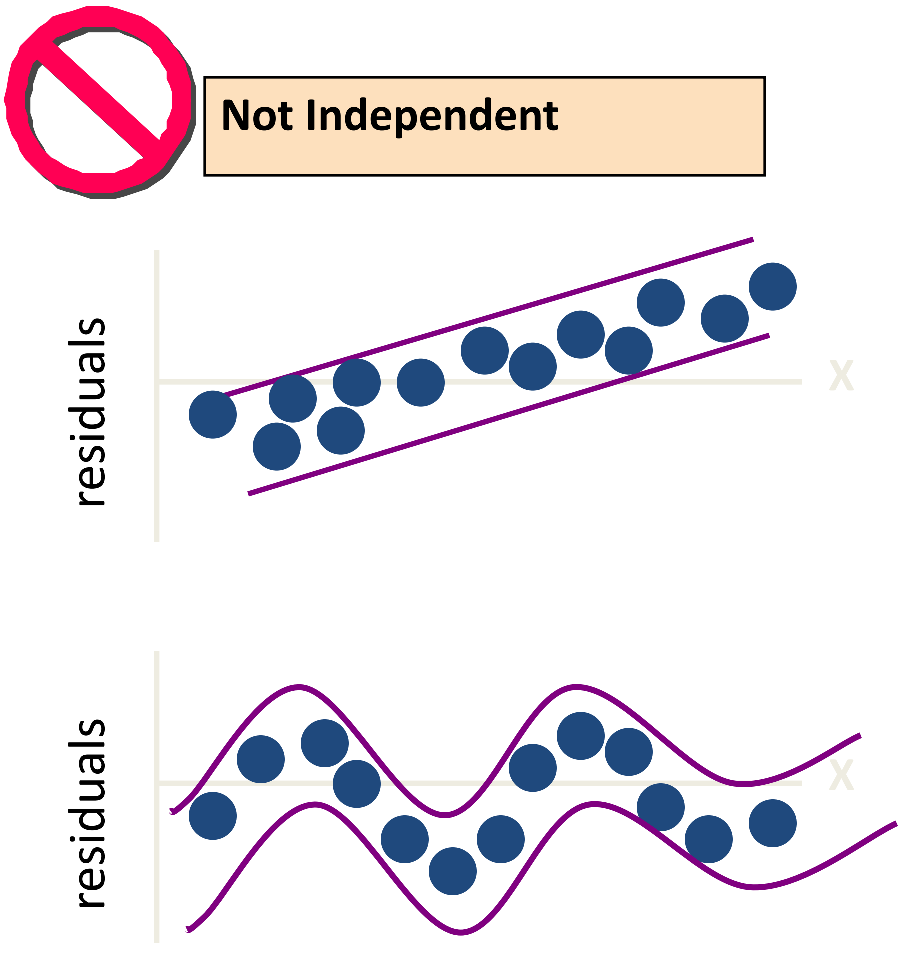
$$\boxed{e_i = Y_i - \hat{Y}_i}$$

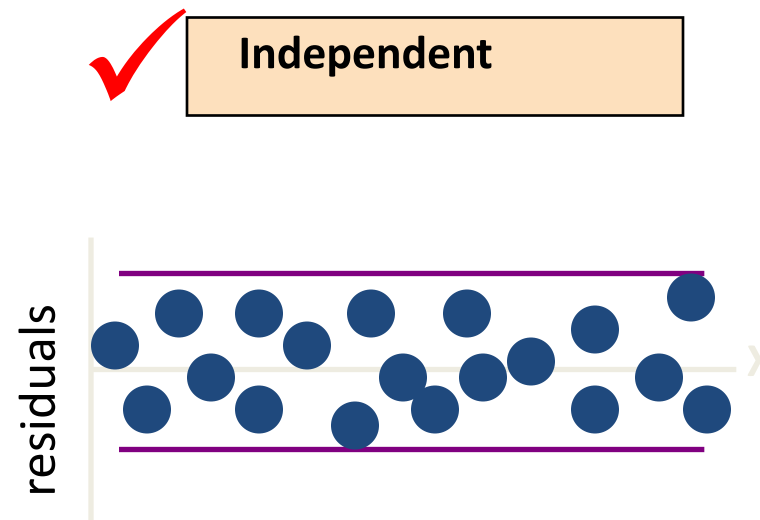
- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant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

Residual Analysis for Linearity



Residual Analysis for Independence



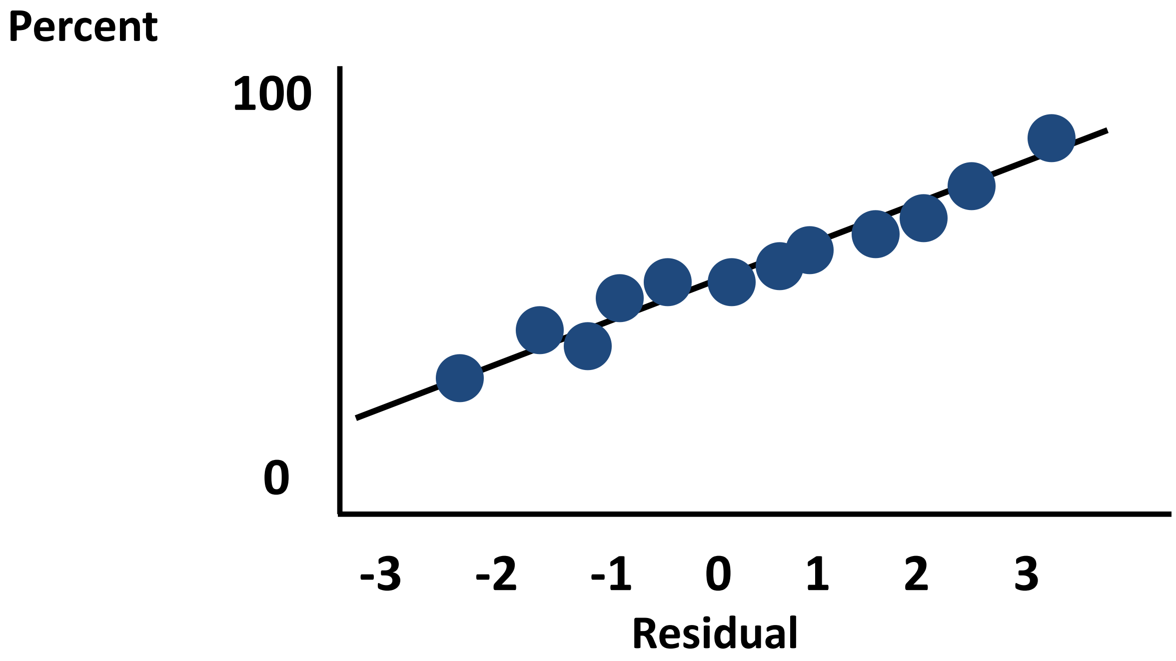


Checking for Normality

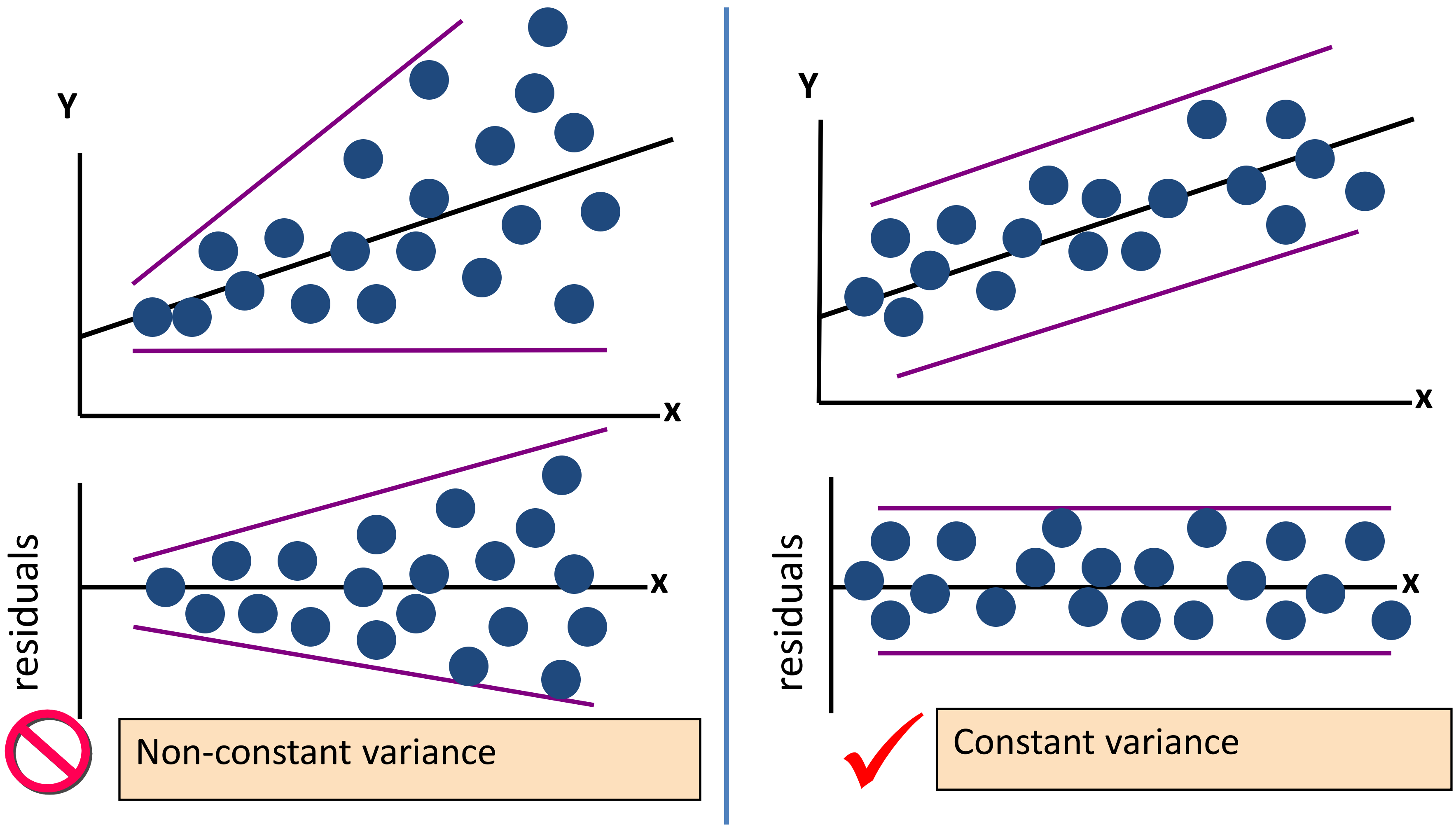
- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Boxplot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals

Residual Analysis for Normality

When using a normal probability plot, normal errors will approximately display in a straight line



Residual Analysis for Equal Variance

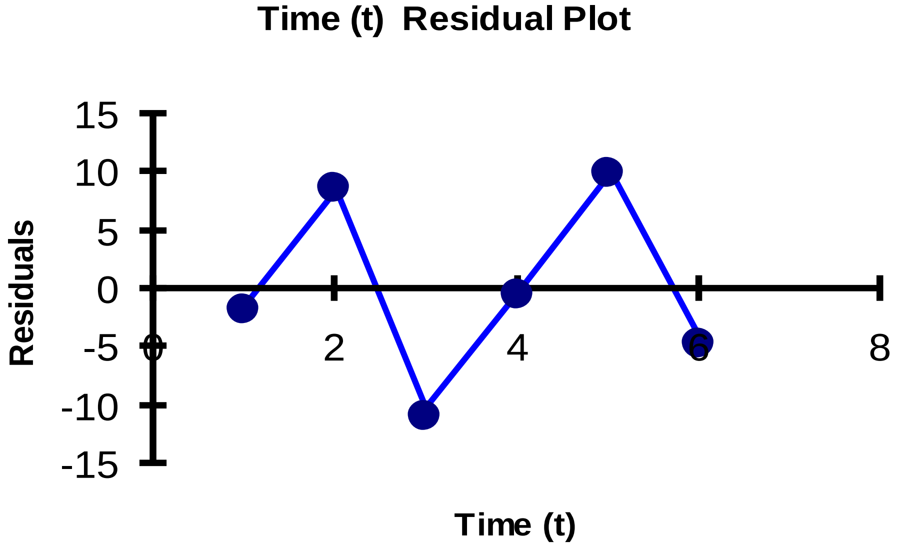


Measuring Autocorrelation: The Durbin-Watson Statistic

- Used when data are collected over time to detect if autocorrelation is present
- Autocorrelation exists if residuals in one time period are related to residuals in another period

Autocorrelation

- Autocorrelation is correlation of the errors (residuals) over time



- Here, residuals show a cyclic pattern (not random.) Cyclical patterns are a sign of positive autocorrelation

- Violates the regression assumption that residuals are random and independent

The Durbin-Watson Statistic

- The Durbin-Watson statistic is used to test for autocorrelation

H_0: residuals are not correlated

H_1: positive autocorrelation is present

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

- The possible range is $0 \le D \le 4$
- D should be close to 2 if H_0 is true
- D less than 2 may signal positive autocorrelation, D greater than 2 may signal negative autocorrelation

Testing for Positive Autocorrelation

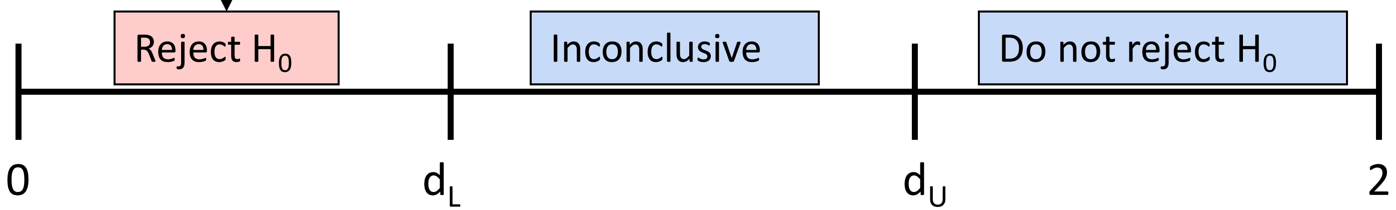
H_0: positive autocorrelation does not exist

H_1: positive autocorrelation is present

- Calculate the Durbin-Watson test statistic = D
  (The Durbin-Watson Statistic can be found using Excel or Minitab)

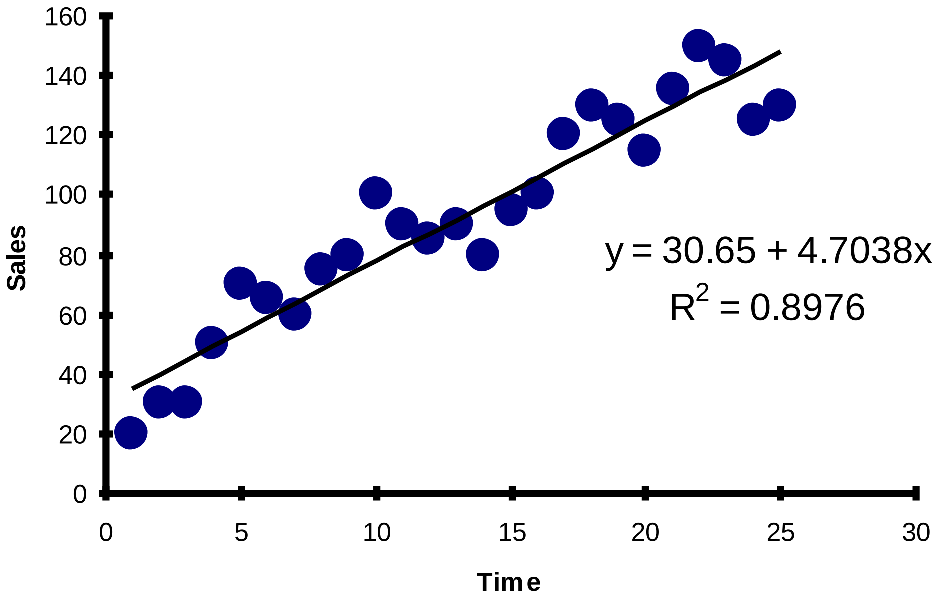
- Find the values d_L and d_U from the Durbin-Watson table
  (for sample size **n** and number of independent variables **k**)

Decision rule: reject H_0 if $D < d_L$



Testing for Positive Autocorrelation

- Suppose we have the following time series data:



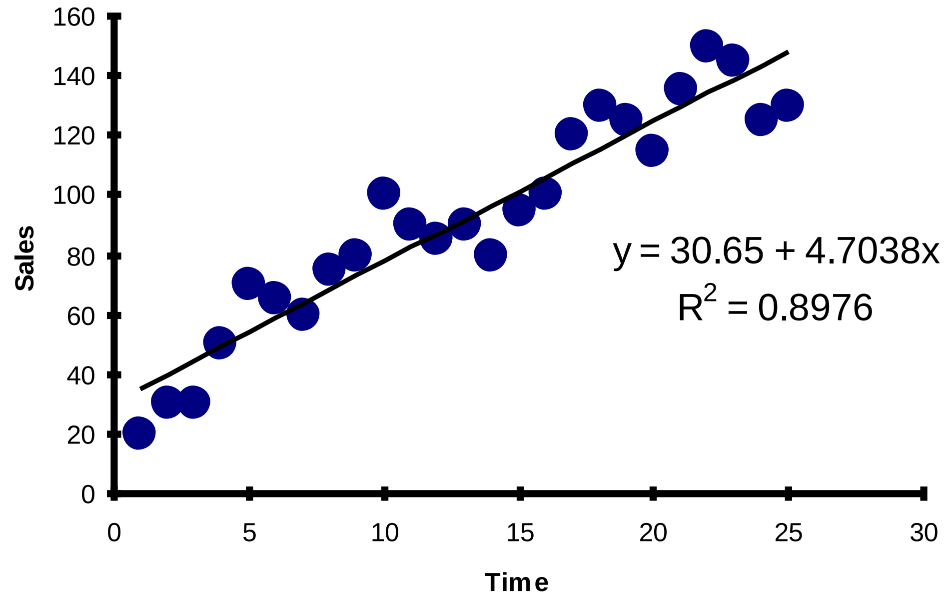
- Is there autocorrelation?

Testing for Positive Autocorrelation

- Example with n = 25:

Excel/PHStat output:

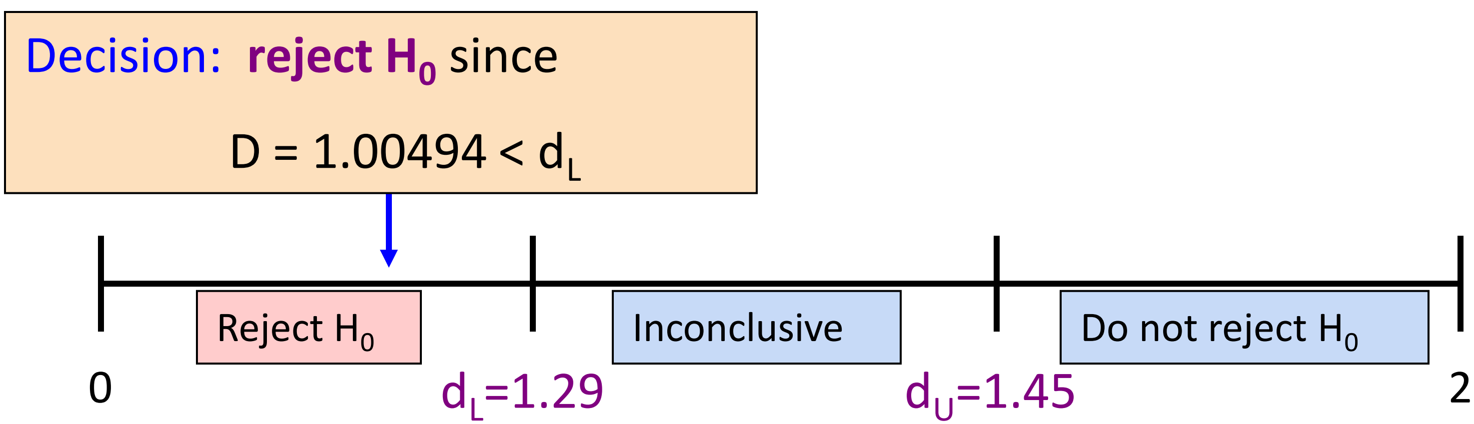
Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	3296.18
Sum of Squared Residuals	3279.98
Durbin-Watson Statistic	**1.00494**



$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{3296.18}{3279.98} = 1.00494$$

Testing for Positive Autocorrelation

- Here, n = 25 and there is k = 1 one independent variable
- Using the Durbin-Watson table, $d_L = 1.29$ and $d_U = 1.45$
- $D = 1.00494 < d_L = 1.29$, so reject H_0 and conclude that significant positive autocorrelation exists



Inferences About the Slope

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

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

where:

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

$S_{YX} = \sqrt{\frac{SSE}{n-2}}$ = Standard error of the estimate

Inferences 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

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

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

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b1} = standard error of the slope

Inferences About the Slope: t Test Example

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:

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

The slope of this model is 0.1098

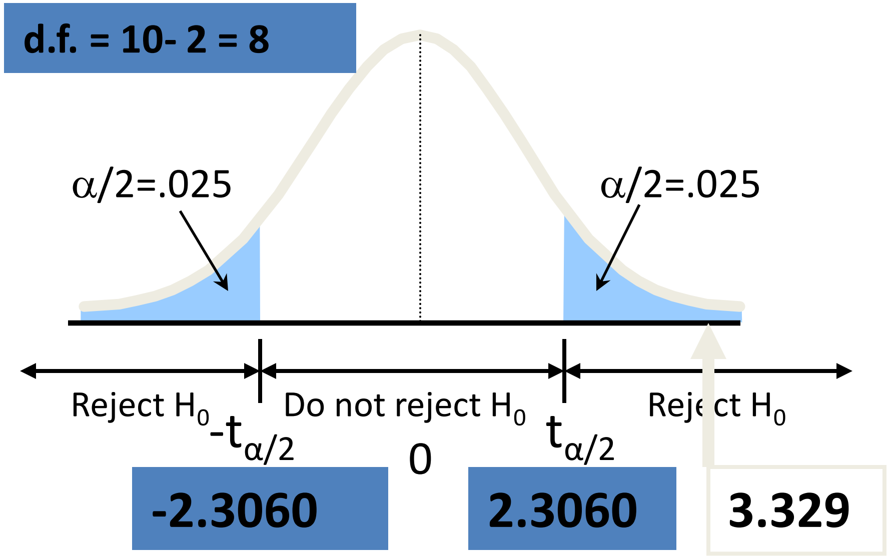
Is there a relationship between the square footage of the house and its sales price?

Inferences About the Slope: t Test Example

Test Statistic: $\mathbf{t_{STAT} = 3.329}$

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$



Decision: Reject H_0

There is sufficient evidence that square footage affects house price

F Test for Significance

- F Test statistic:

where

$$F_{STAT} = \frac{MSR}{MSE}$$

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

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

where F_{STAT} 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)

F Test for Significance

$H_0: \beta_1 = 0$

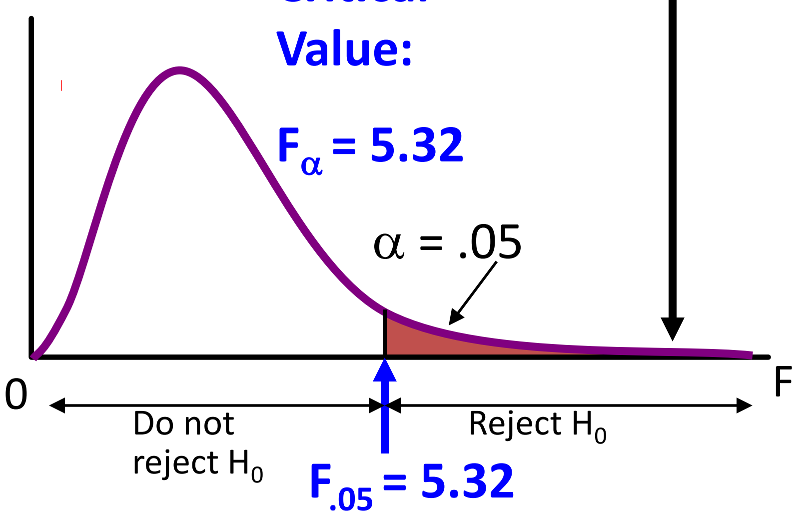
$H_1: \beta_1 \neq 0$

$\alpha = .05$

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

Critical Value:

$F_\alpha = 5.32$



Test Statistic:

$$F_{STAT} = \frac{MSR}{MSE} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

t Test for a Correlation Coefficient

- Hypotheses

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

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

- Test statistic

$$t_{STAT} = \frac{r - \rho}{\sqrt{\dfrac{1 - r^2}{n - 2}}}$$

(with n – 2 degrees of freedom)

where

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

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

t-test For A Correlation Coefficient

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)

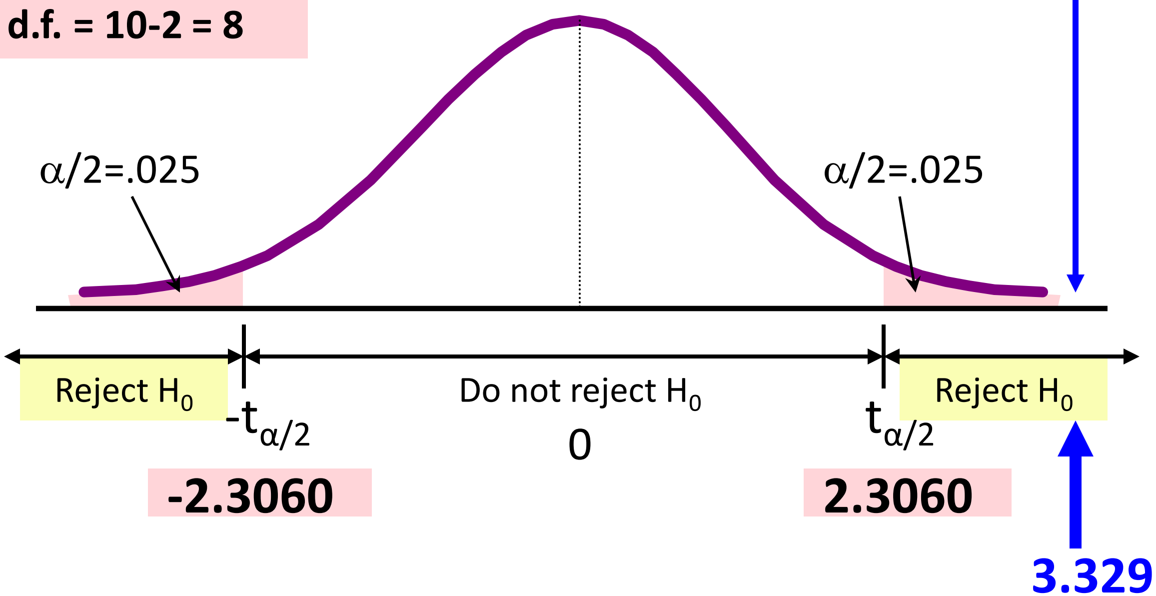
α =.05 , df = 10 - 2 = 8

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

t-test For A Correlation Coefficient

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

d.f. = 10-2 = 8



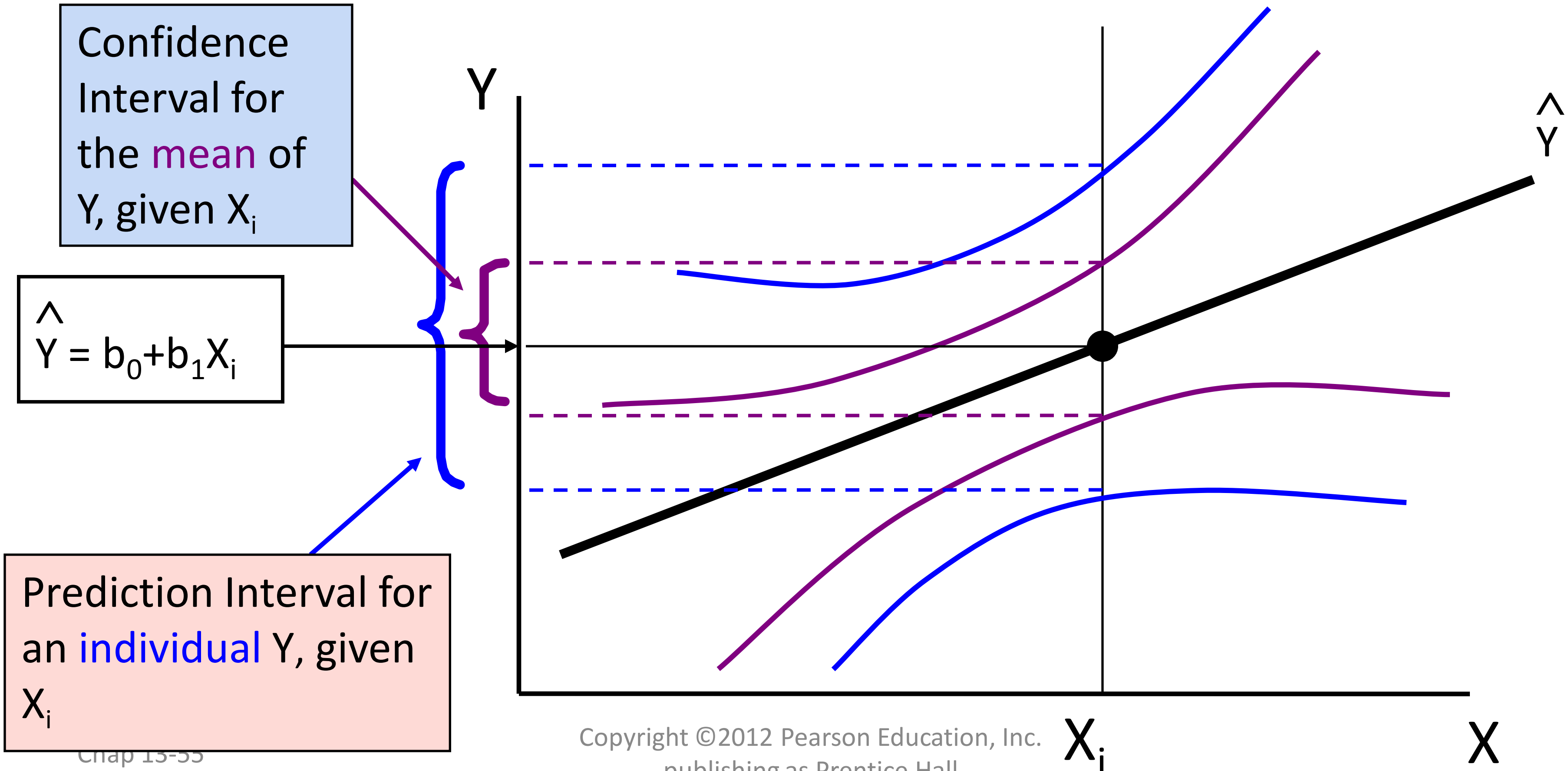
Decision:
Reject H_0

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

Estimating Mean Values and Predicting Individual Values

DCOVA

Goal: Form intervals around Y to express uncertainty about the value of Y for a given 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 plot 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, boxplot, or normal probability plot of the residuals to uncover possible non-normality

Strategies for Avoiding the Pitfalls of Regression

- 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

EXERCISE

13.4 (cont’d)

The marketing manager of a large supermarket chain would like to use shelf space to predict the sales of pet food. A random sample of 12 equal-sized stores is selected, with the following results

13.4 (cont’d)

Store	Shelf Space (X) (Feet)	Weekly Sales (Y) ($)
1	5	160
2	5	220
3	5	140
4	10	190
5	10	240
6	10	260
7	15	230
8	15	270
9	15	280
10	20	260
11	20	290
12	20	310

13.4

a. Assuming a linear relationship, use the least-squares method to compute the regression coefficients b_0 and b_1
b. Interpret the meaning of the *Y* intercept, b_0 and the slope, b_1 in this problem
c. Predict the weekly sales of pet food for stores with 8 feet of shelf space for pet food.

Example

Hours Spent Studying (X)	Math SAT Score (Y)
4	390
9	580
10	650
14	730
4	410
7	530
12	600
22	790
1	350
3	400
8	590
11	640
5	450
6	520
10	690

THANK YOU