7CPAPS_Statistics

2023-2024

Module Outline

Content of this Module

- How to use regression analysis to predict the value of a dependent variable based on a value of an independent variable.
- Understanding the meaning of the regression coefficients b_0 and b_1.
- Evaluating the assumptions of regression analysis and know what to do if the assumptions are violated.
- Making inferences about the slope and correlation coefficient.
- Estimating mean values and predicting individual values.

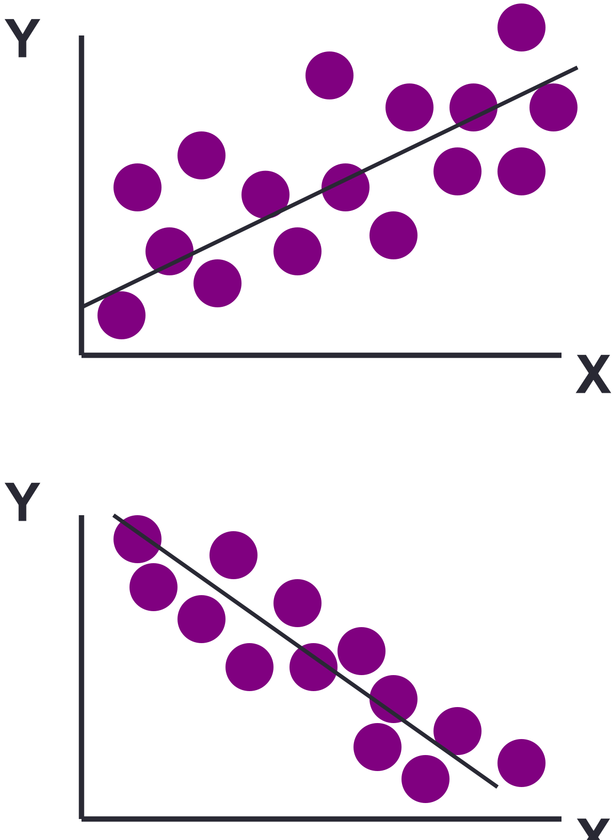
Correlation vs. Regression

- A scatter plot can be used to show the relationship between two variables.

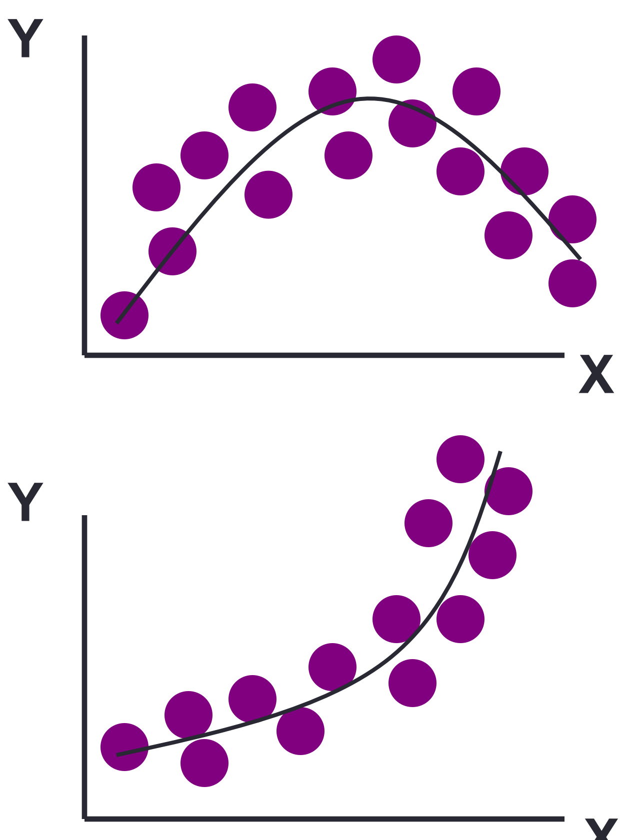
- Correlation analysis is used to measure the strength of the association (linear relationship) between two variables.
  - Correlation is only concerned with strength of the relationship.
  - No causal effect is implied with correlation.

Types of Relationships

Linear relationships

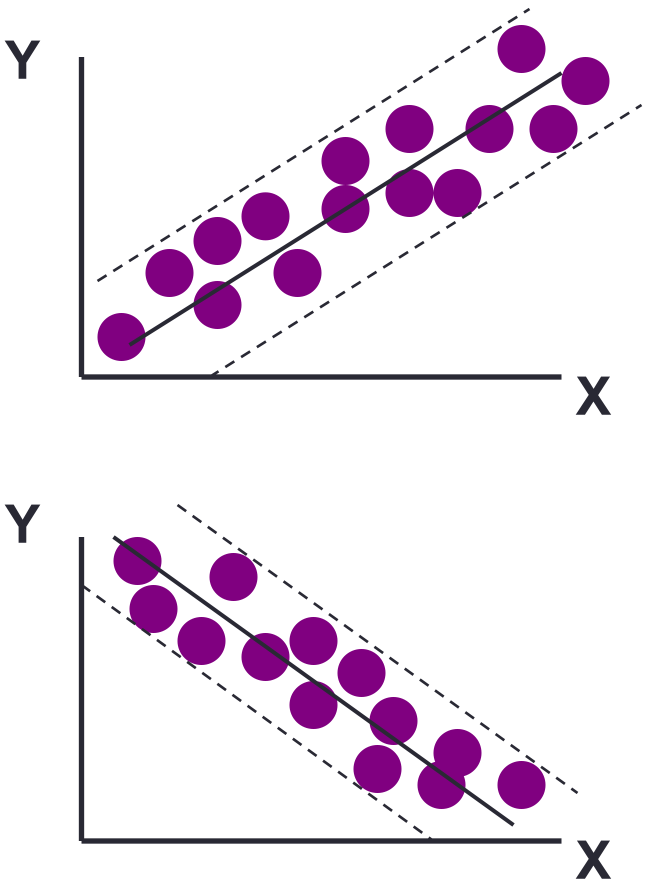


Curvilinear relationships

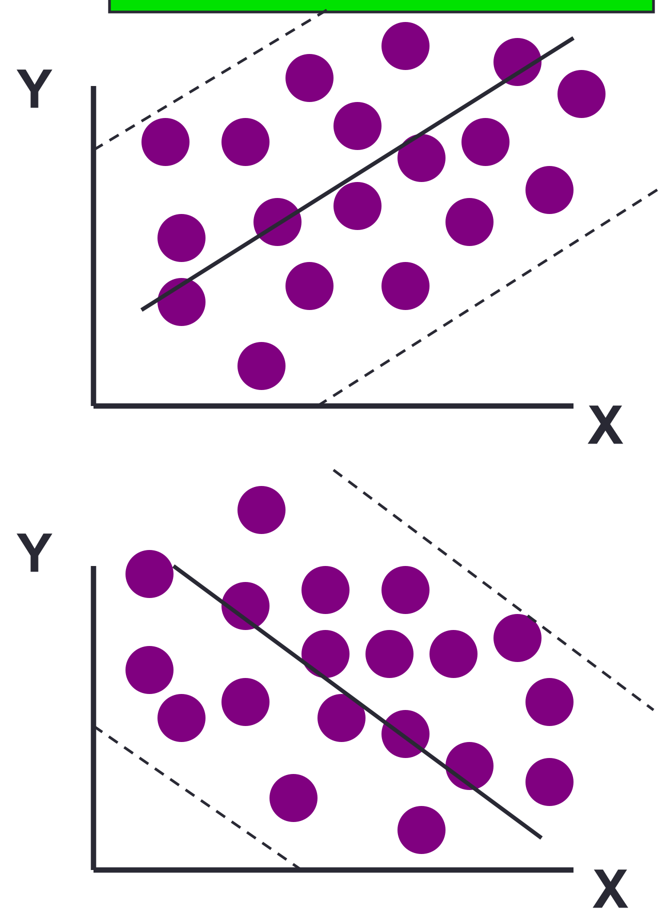


Types of Relationships

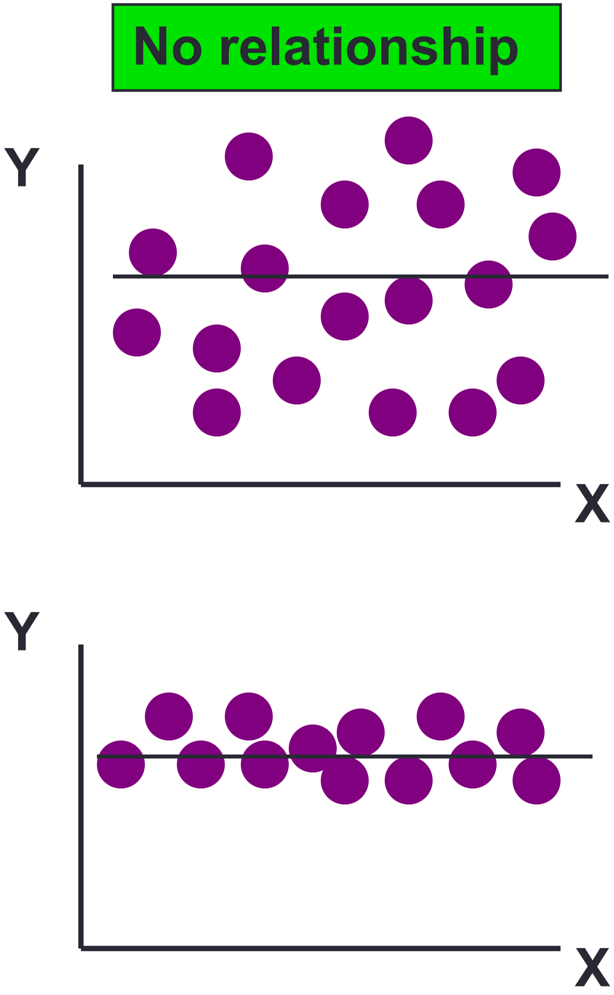
Strong relationships



Weak relationships



Types of Relationships



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.

Simple Linear Regression Model

Dependent Variable

Population Y intercept

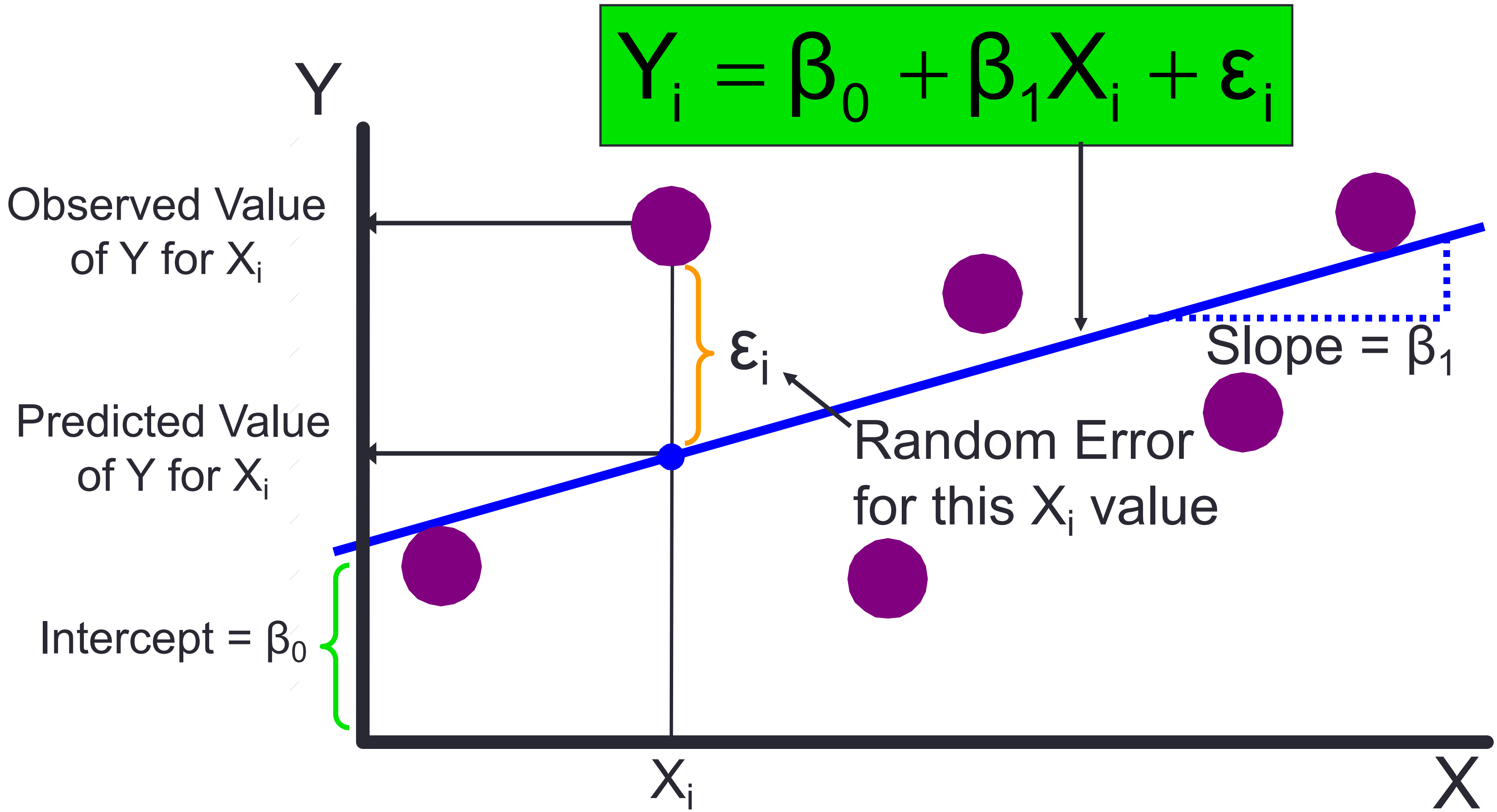
Population Slope Coefficient

Independent Variable

Random Error term

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

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 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, will be found using Excel.
- The calculations for b_0 and b_1 are not shown here but are available in the textbook section 13.2.

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).

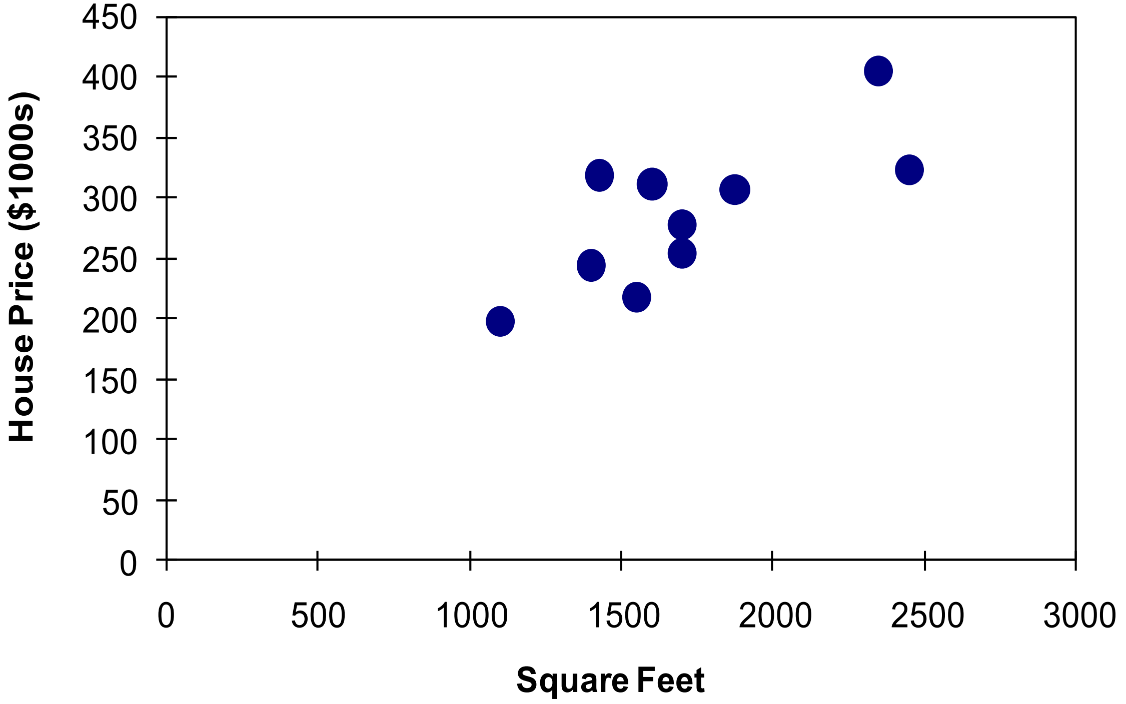
- A random sample of 10 houses is selected.
  - Dependent variable (Y) = house price in $1,000s.
  - Independent variable (X) = square feet.

Simple Linear Regression Example: Data

House Price in $1000s (Y)	Square Feet (X)
245	1,400
312	1,600
279	1,700
308	1,875
199	1,100
219	1,550
405	2,350
324	2,450
319	1,425
255	1,700

Simple Linear Regression Example: Scatter Plot

House price model: Scatter Plot



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:

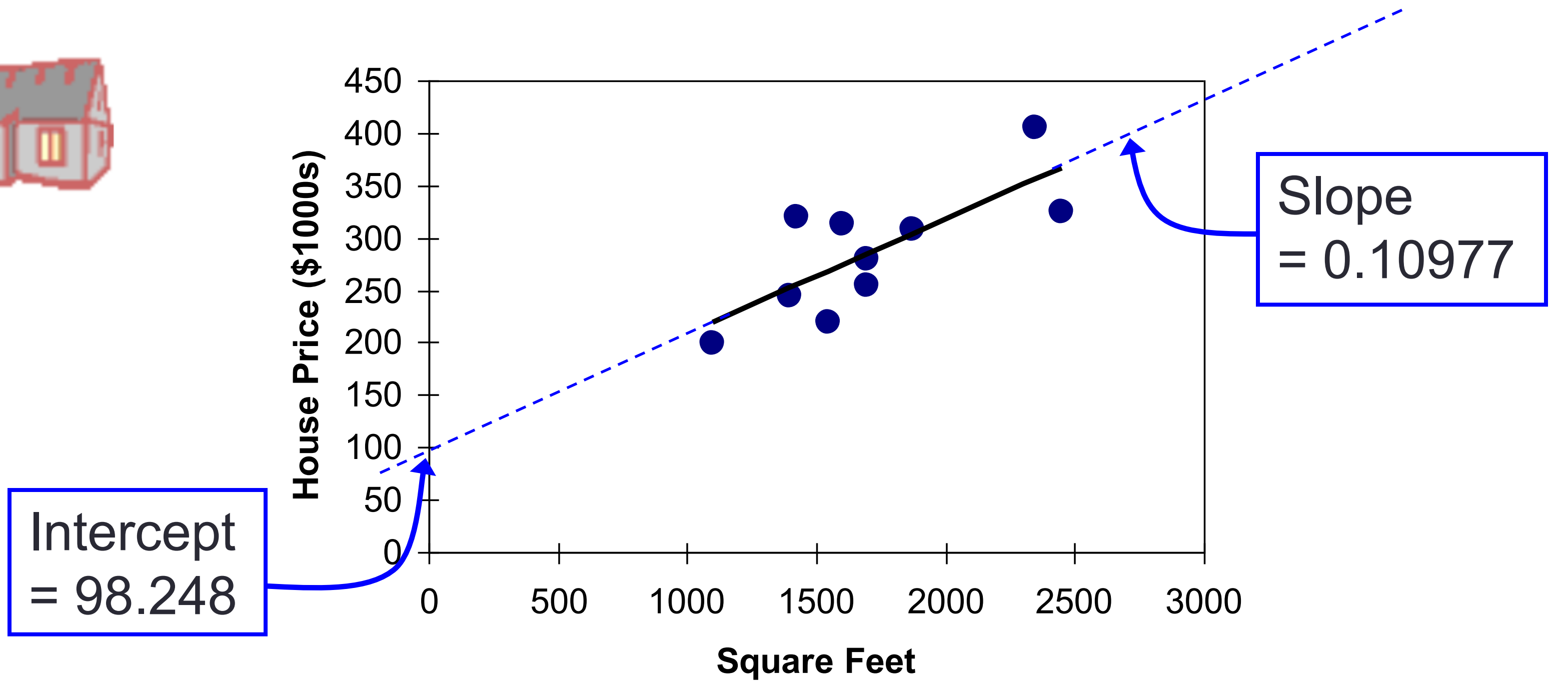
$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

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



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

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, $\boxed{b_1 = 0.10977}$ tells us that the mean value of a house increases by 0.10977($1,000) = $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 2,000 square feet:

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

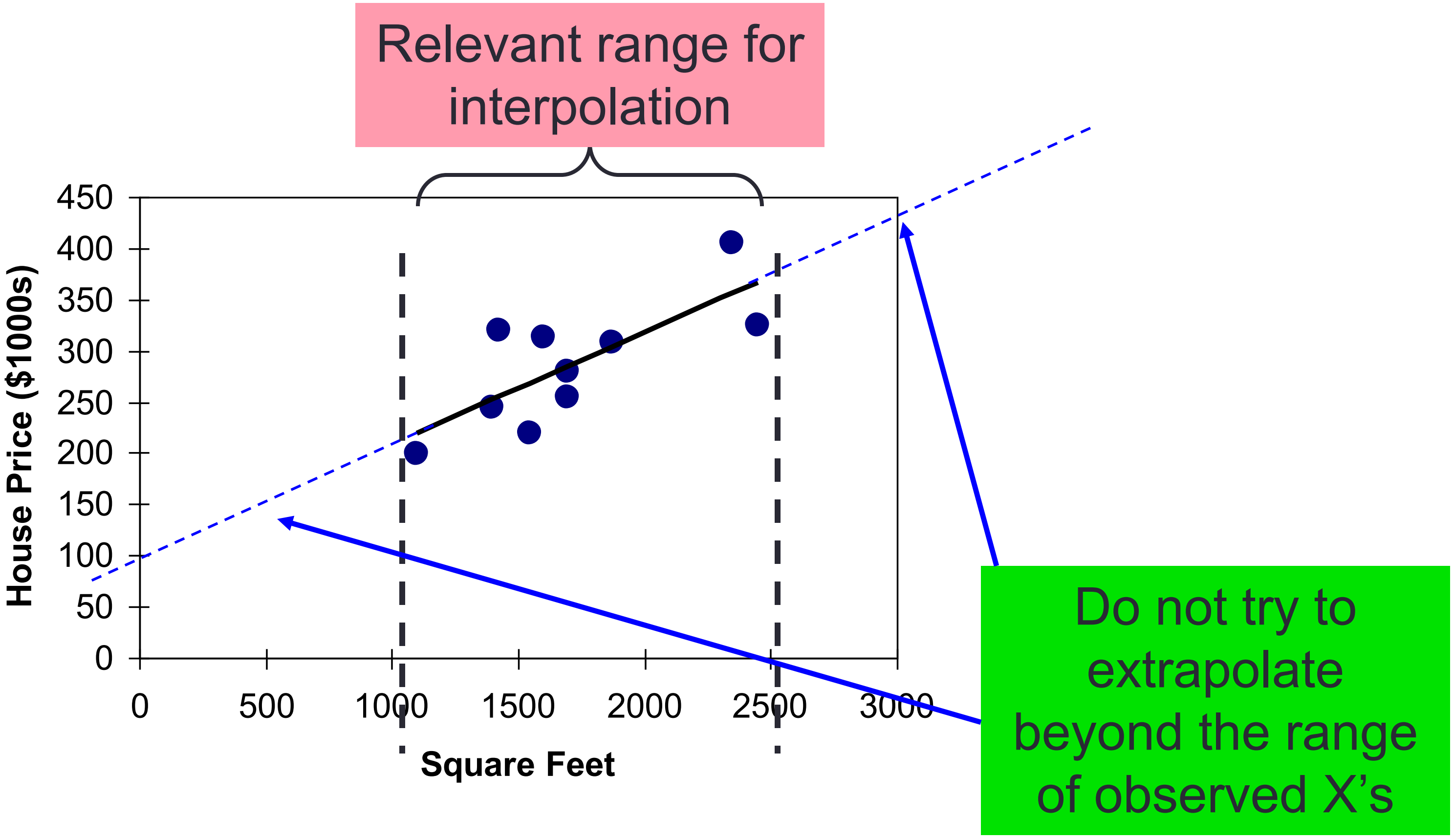
$$= 98.25 + 0.1098(2{,}000)$$

$$= 317.85$$

The predicted price for a house with 2,000 square feet is 317.85($1,000s) = $317,850

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:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares

$$SST = \sum (Y_i - \bar{Y})^2 \qquad SSR = \sum (\hat{Y}_i - \bar{Y})^2 \qquad SSE = \sum (Y_i - \hat{Y}_i)^2$$

where:

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

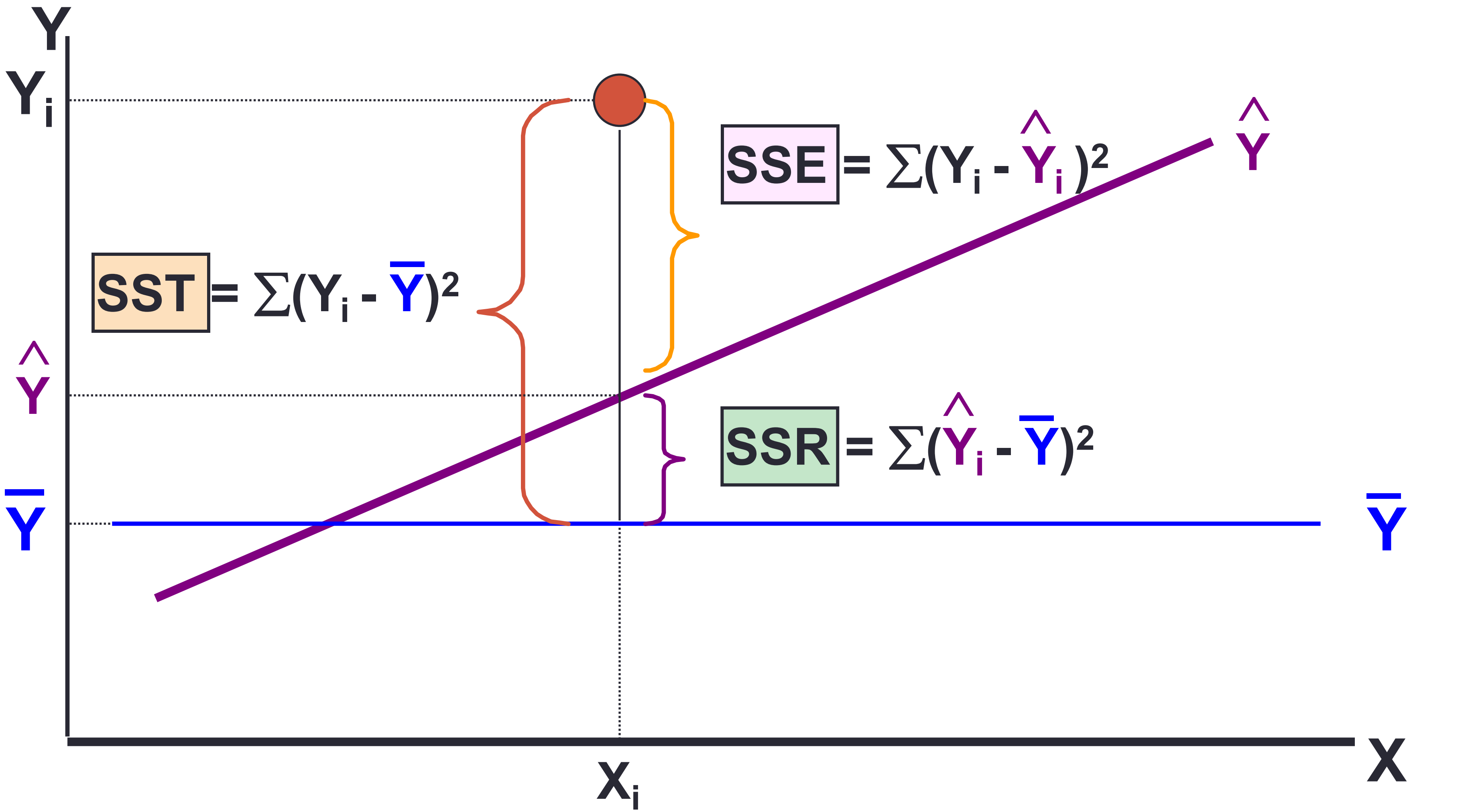
Y_i = Observed value of the dependent variable

$\hat{Y}_i$ = 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



Excel Output Of The Measures Of Variation

10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1	18934.9348	18934.9348	11.0848	0.0104
13	Residual	8	13665.5652	1708.1957		
14	Total	9	32600.5000			

SST = SSR + SSE

32,600.5000 = 18,934.9348 + 13,665.5652

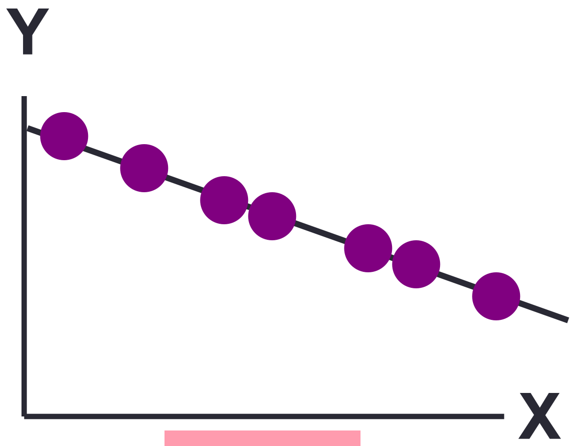
Coefficient of Determination, r^2

- 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-square and is denoted as r^2.

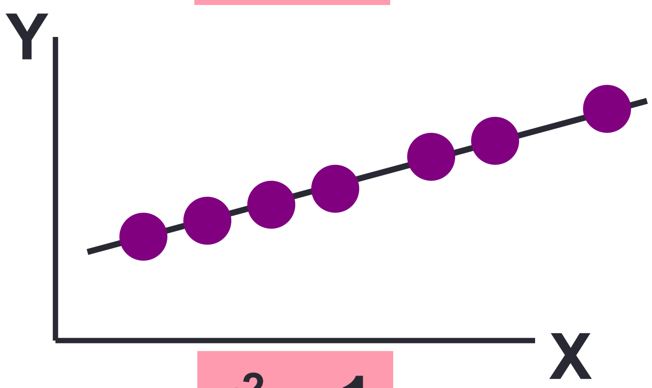
$$\boxed{r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}}$$

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

Examples of Approximate r^2 Values



$r^2 = 1$

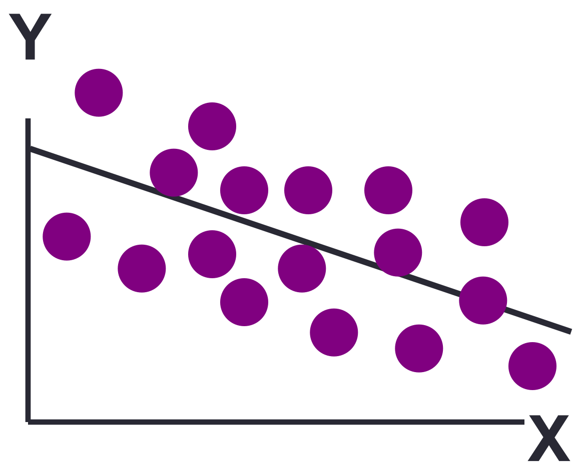


$r^2 = 1$

Perfect linear relationship between X and Y.

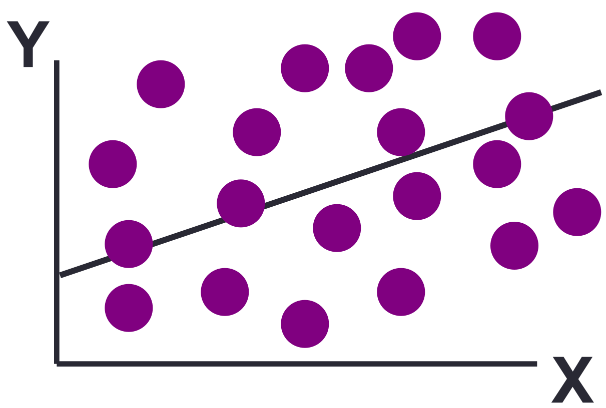
100% of the variation in Y is explained by variation in X.

Examples of Approximate r^2 Values



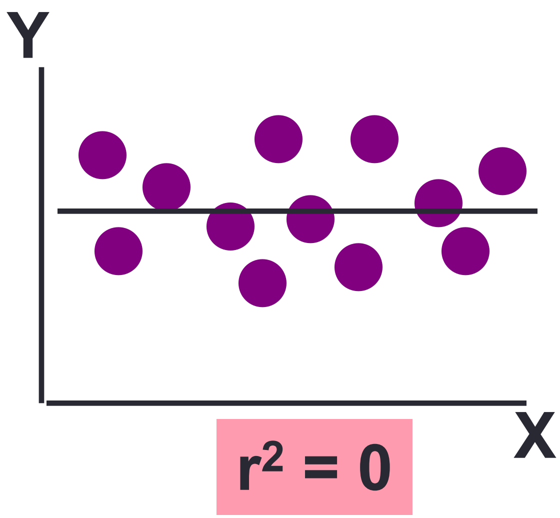
$0 < r^2 < 1$

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.)

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

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

$$r^2 = \frac{SSR}{SST} = \frac{18{,}934.9348}{32{,}600.5000} = 0.58082$$

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:

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

Where

SSE = error sum of squares

n = sample size

Simple Linear Regression Example: Standard Error of Estimate in Excel

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

$$S_{YX} = 41.33032$$

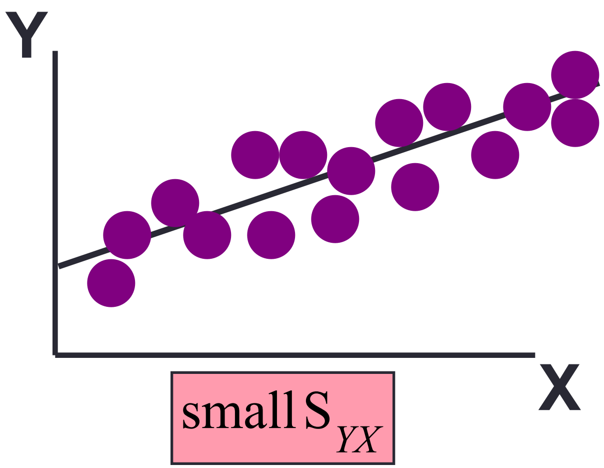
ANOVA

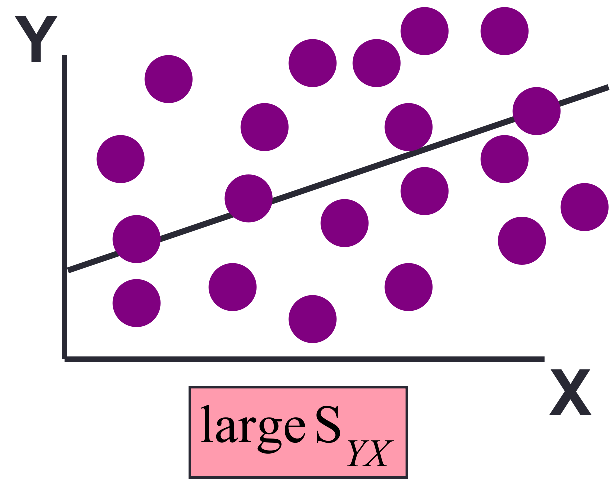
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	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 \$200K - \$400K range

Assumptions of Regression: L.I.N.E

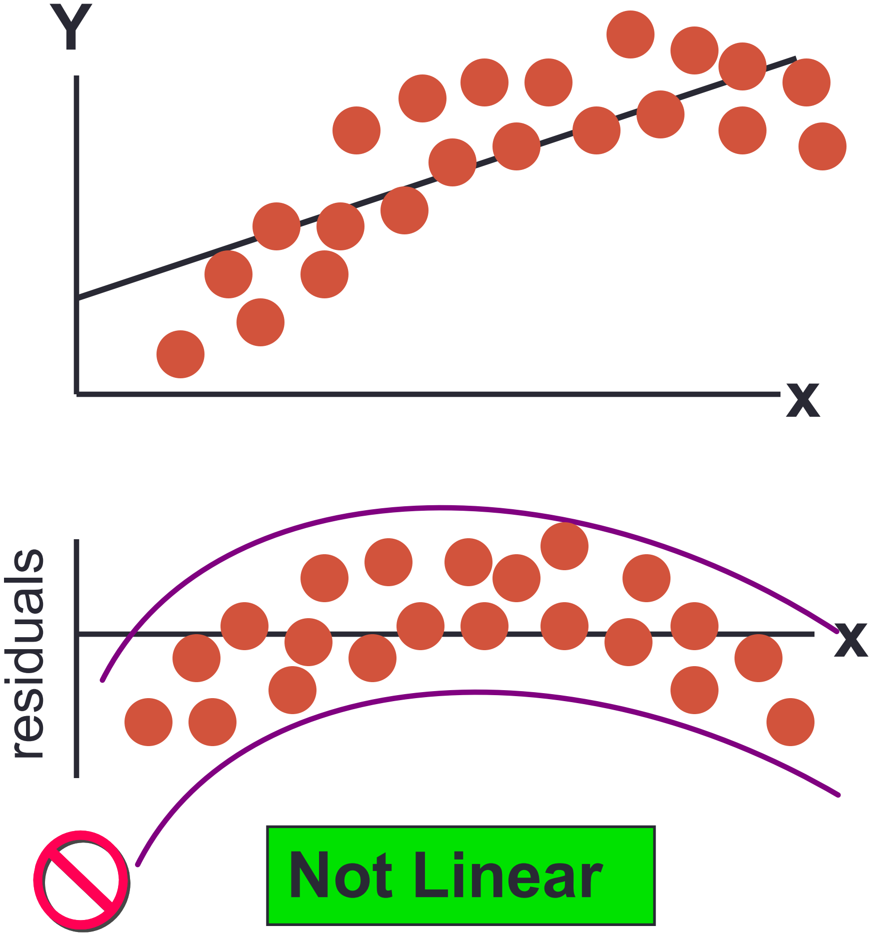
- Linearity:
  - The relationship between X and Y is linear.
- Independence of Errors:
  - Error values are statistically independent.
  - Particularly important when data are collected over a period of time.
- Normality of Error:
  - Error values are normally distributed for any given value of X.
- Equal Variance (also called homoscedasticity):
  - The probability distribution of the errors has constant variance.

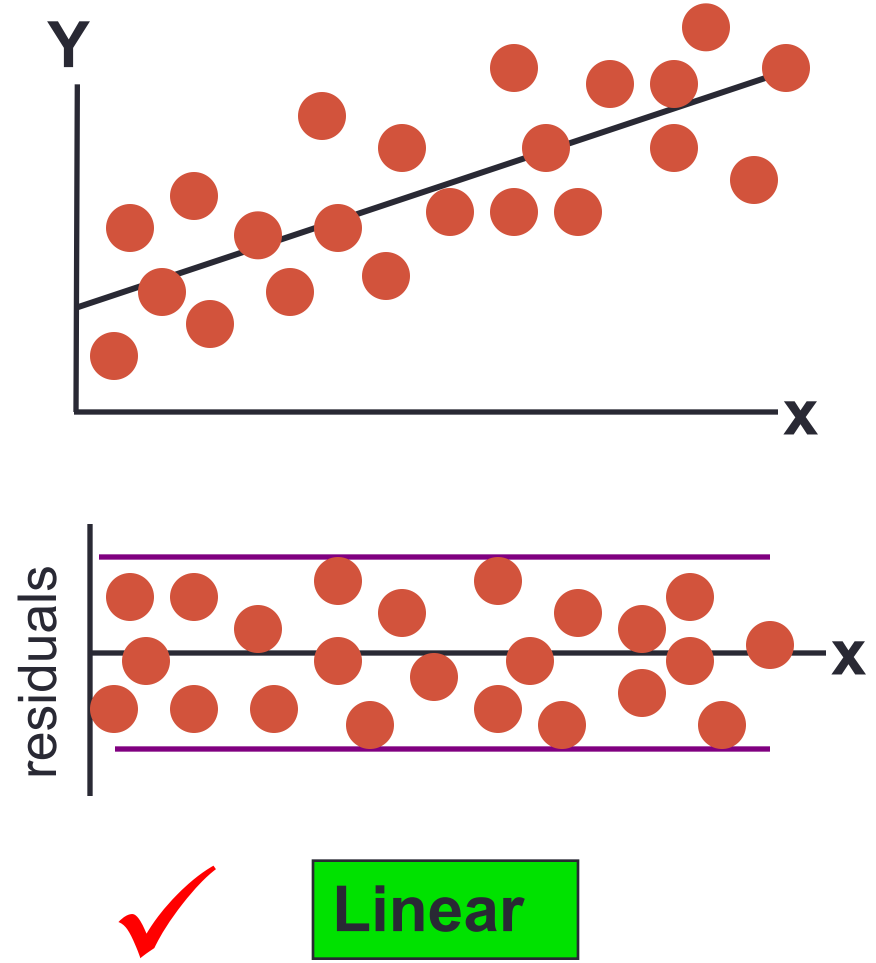
Residual Analysis

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

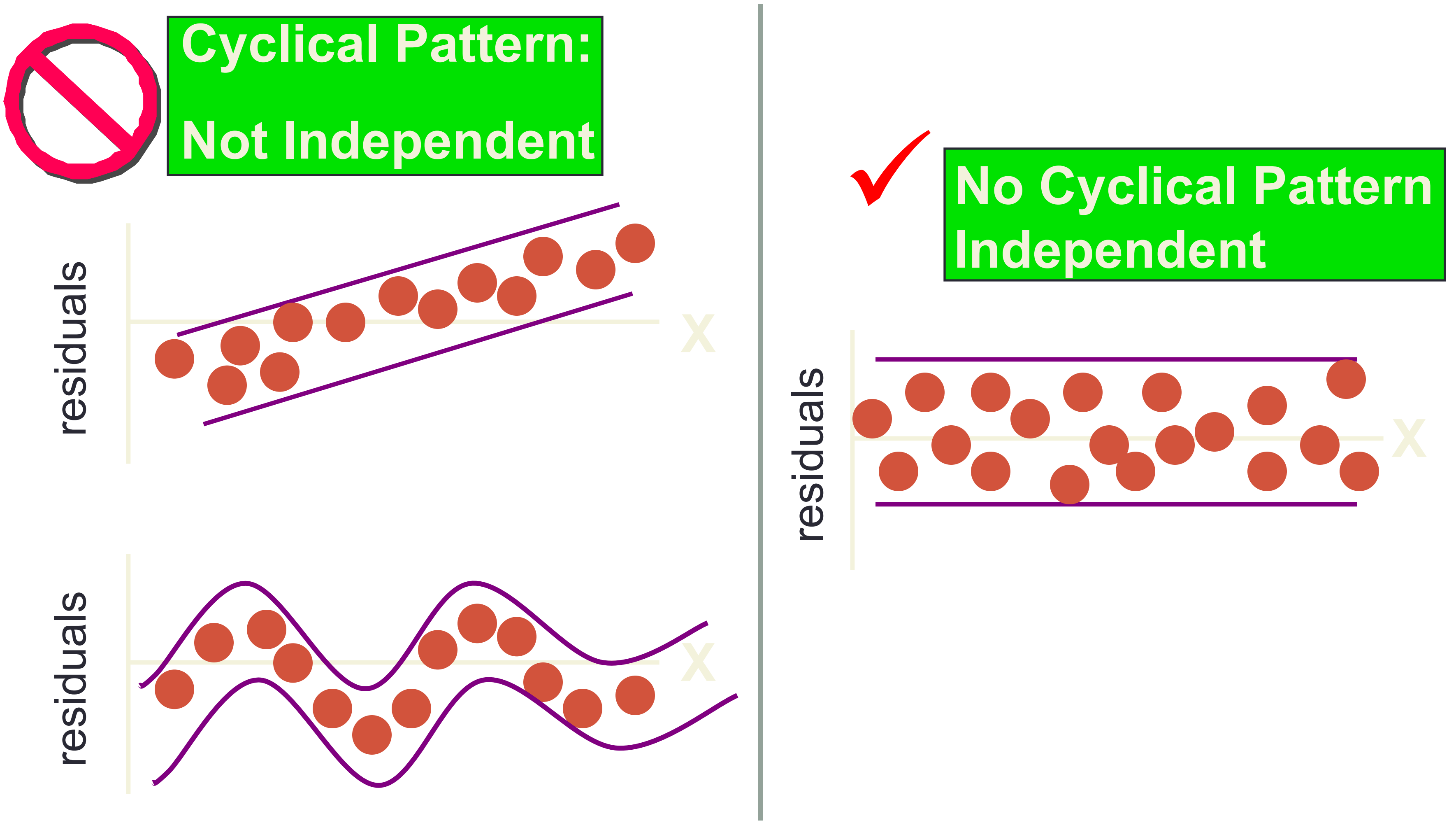
- The residual for observation i, e_i, is the difference between its observed and predicted value.
- Check the assumptions of regression by examining the residuals:
  - Examine for linearity assumption.
  - Evaluate independence assumption.
  - Evaluate normal distribution assumption.
  - Examine for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

Residual Analysis for Linearity





Residual Analysis for Independence

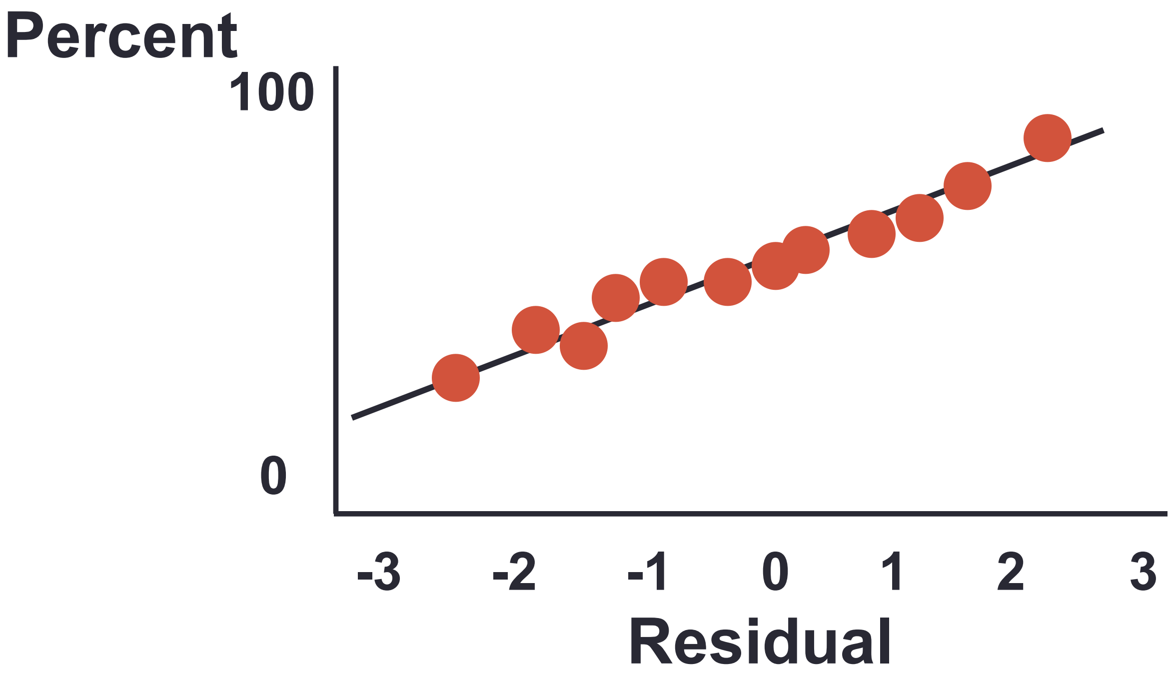


Checking for Normality

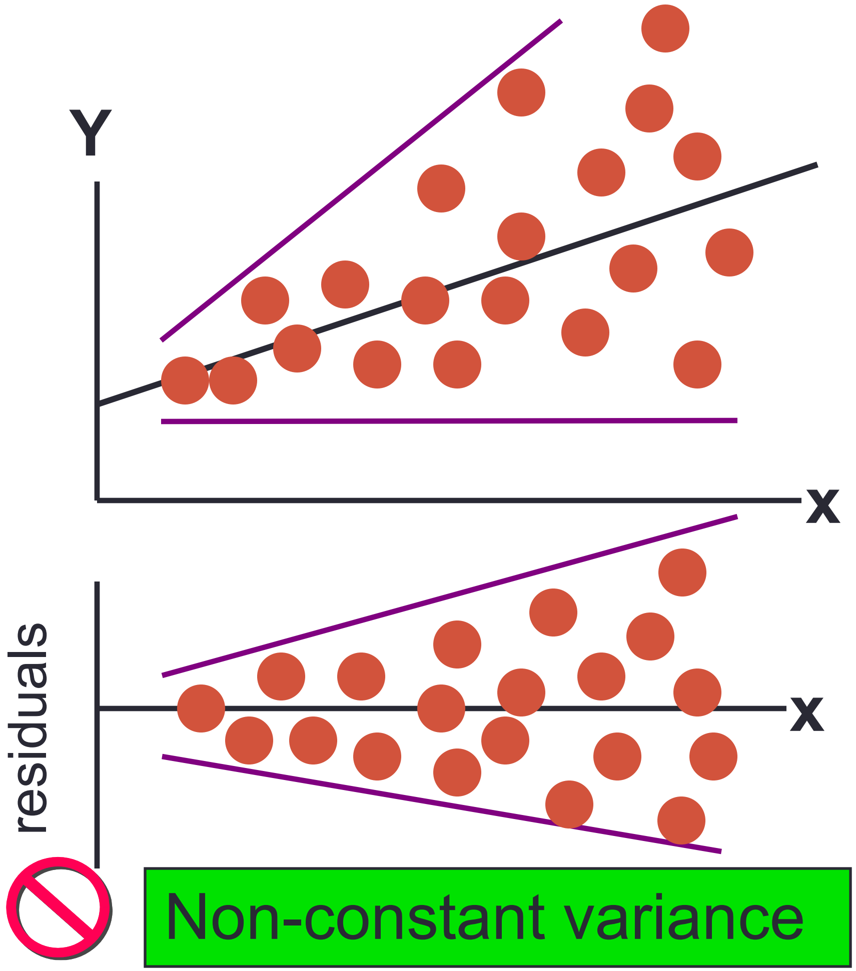
- Examine the Stem-and-Leaf Display of the Residuals.
- Examine the Boxplot of the Residuals.
- Examine the Histogram of the Residuals.
- Construct a Normal Probability Plot of the Residuals.

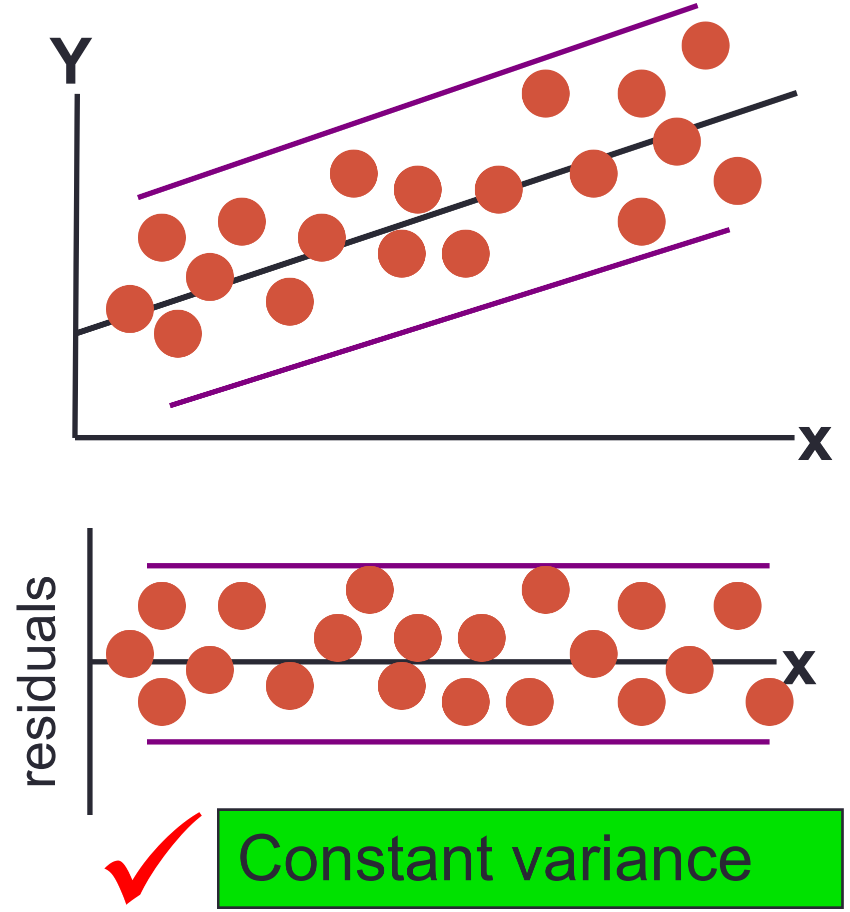
Residual Analysis for Normality

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



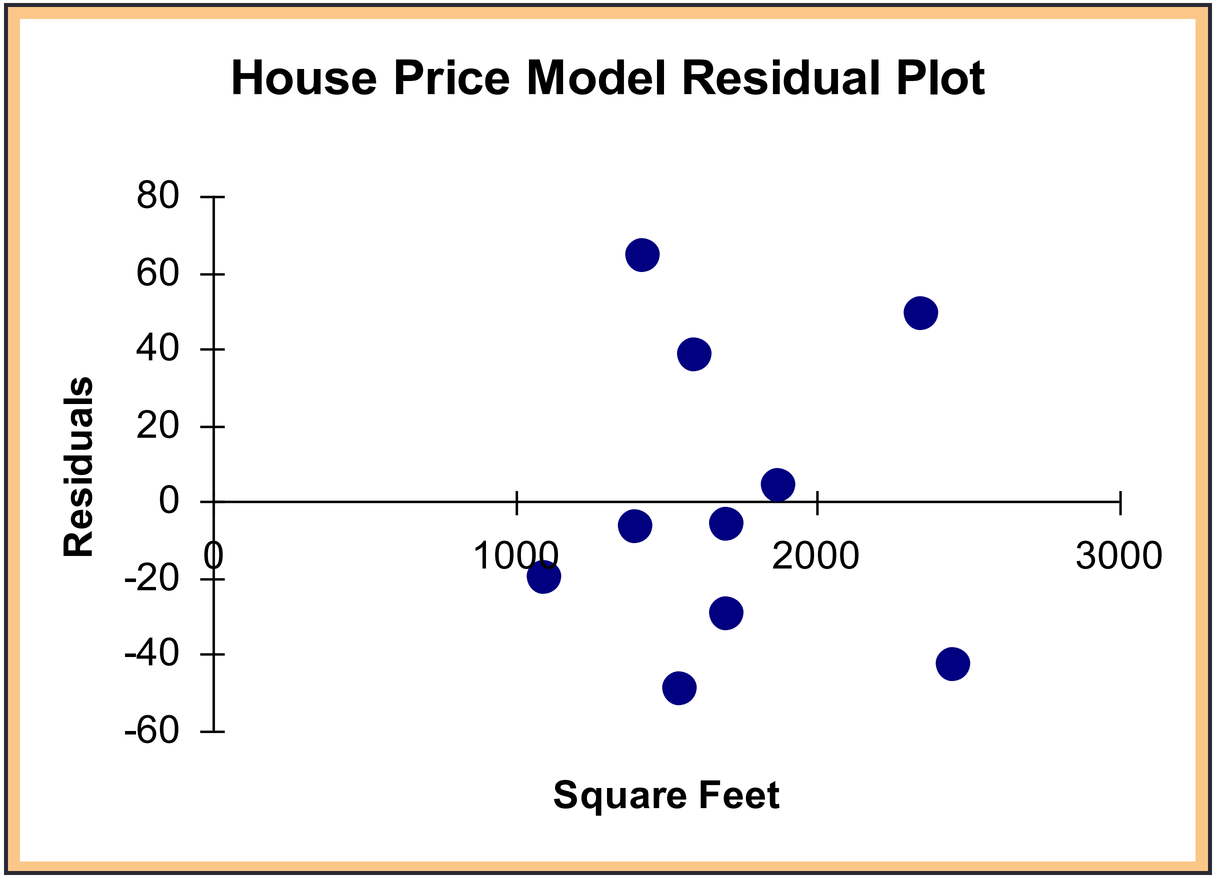
Residual Analysis for Equal Variance





Residual Analysis: Checking For Linearity

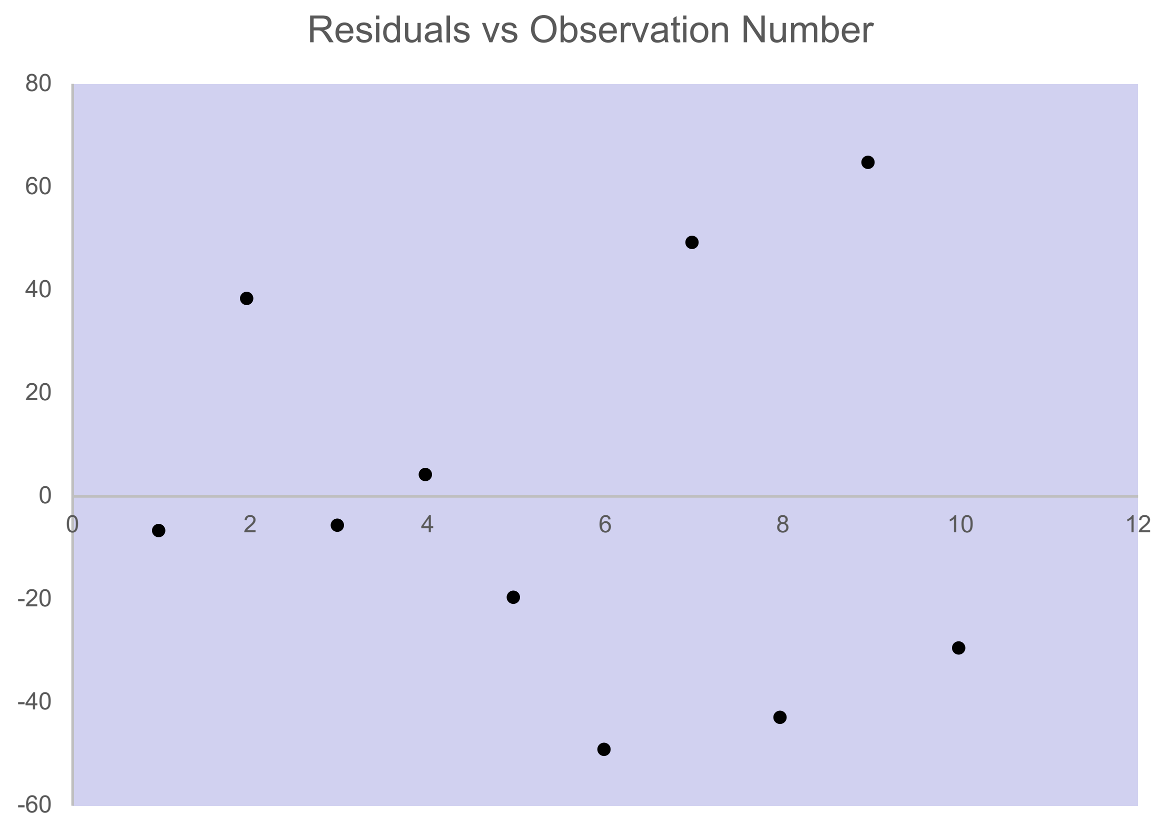
RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.9232	-6.9232
2	273.8767	38.1233
3	284.8535	-5.8535
4	304.0628	3.9371
5	218.9928	-19.9928
6	268.3883	-49.3883
7	356.2025	48.7975
8	367.1793	-43.1793
9	254.6674	64.3326
10	284.8535	-29.8535



Linear Model Assumption Is Appropriate

Residual Analysis: Checking For Independence

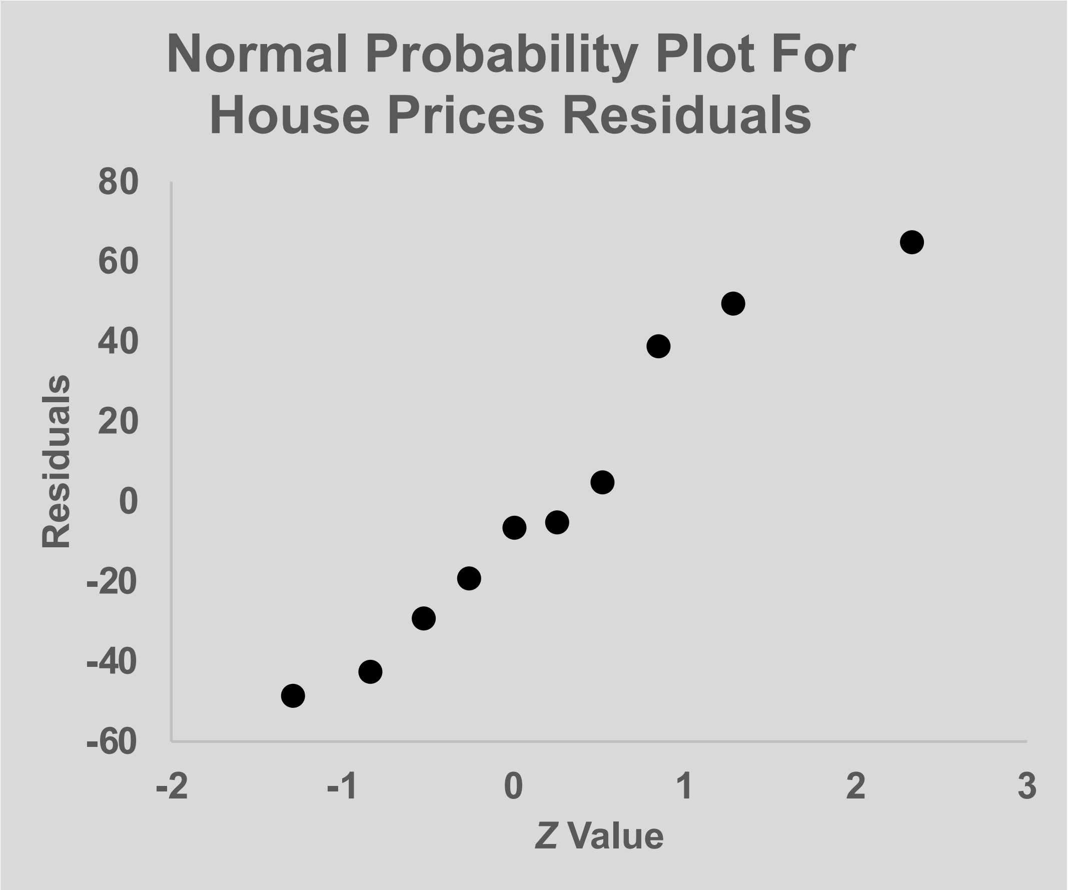
RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.9232	-6.9232
2	273.8767	38.1233
3	284.8535	-5.8535
4	304.0628	3.9371
5	218.9928	-19.9928
6	268.3883	-49.3883
7	356.2025	48.7975
8	367.1793	-43.1793
9	254.6674	64.3326
10	284.8535	-29.8535



Independence Assumption Is Appropriate

Residual Analysis: Checking For Normality

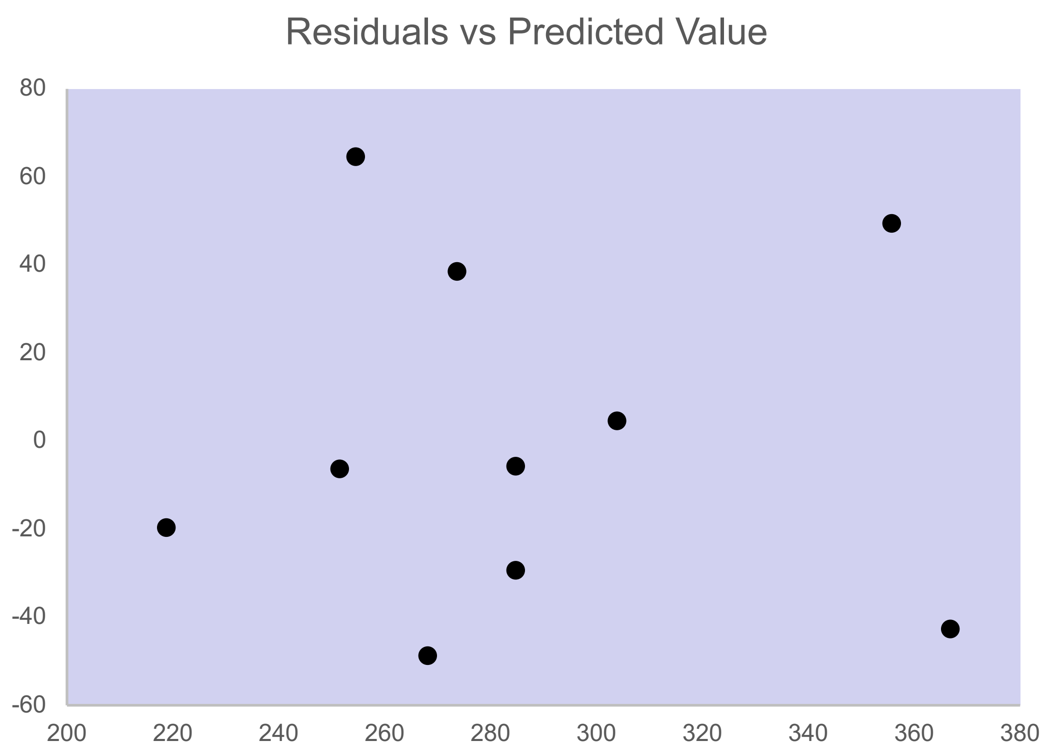
RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.9232	-6.9232
2	273.8767	38.1233
3	284.8535	-5.8535
4	304.0628	3.9371
5	218.9928	-19.9928
6	268.3883	-49.3883
7	356.2025	48.7975
8	367.1793	-43.1793
9	254.6674	64.3326
10	284.8535	-29.8535



Normality Assumption Is Appropriate

Residual Analysis: Checking For Constant Variance

RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.9232	-6.9232
2	273.8767	38.1233
3	284.8535	-5.8535
4	304.0628	3.9371
5	218.9928	-19.9928
6	268.3883	-49.3883
7	356.2025	48.7975
8	367.1793	-43.1793
9	254.6674	64.3326
10	284.8535	-29.8535



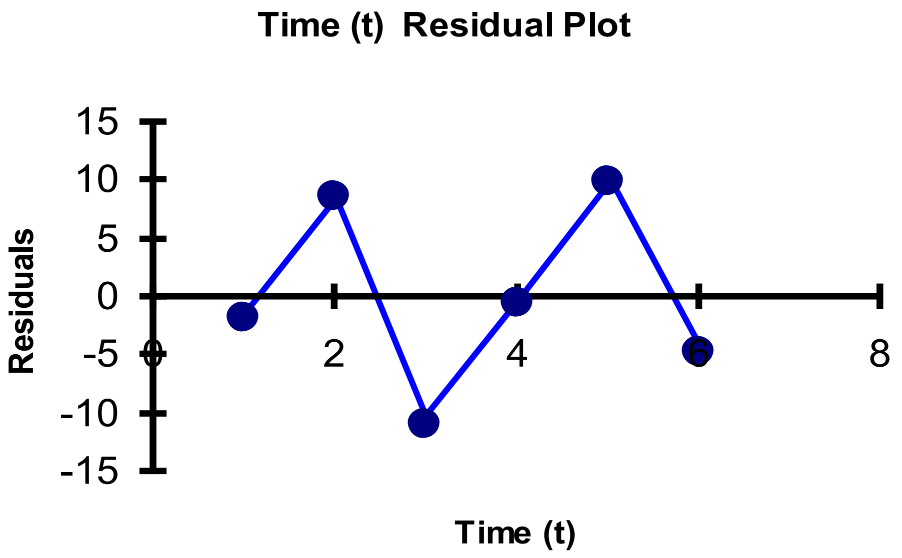
Constant Variance Assumption Is Appropriate

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 cyclical 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: positive autocorrelation does not exist

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

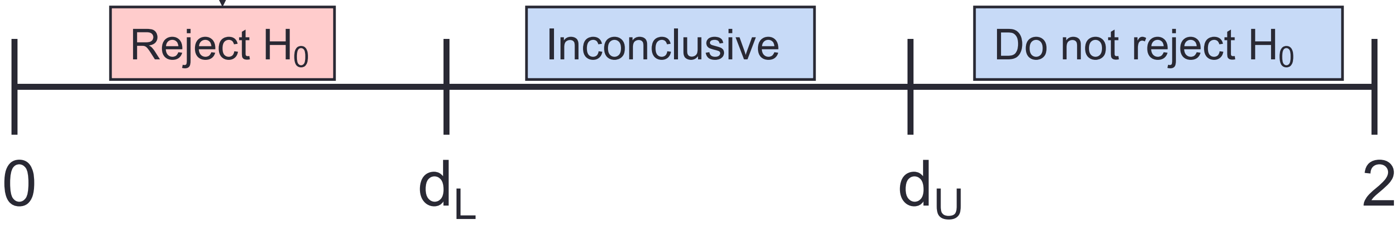
- Calculate the Durbin-Watson test statistic = D.

  (The Durbin-Watson Statistic can be found using Excel.)

- Find the values d_L and d_U from the Durbin-Watson table.

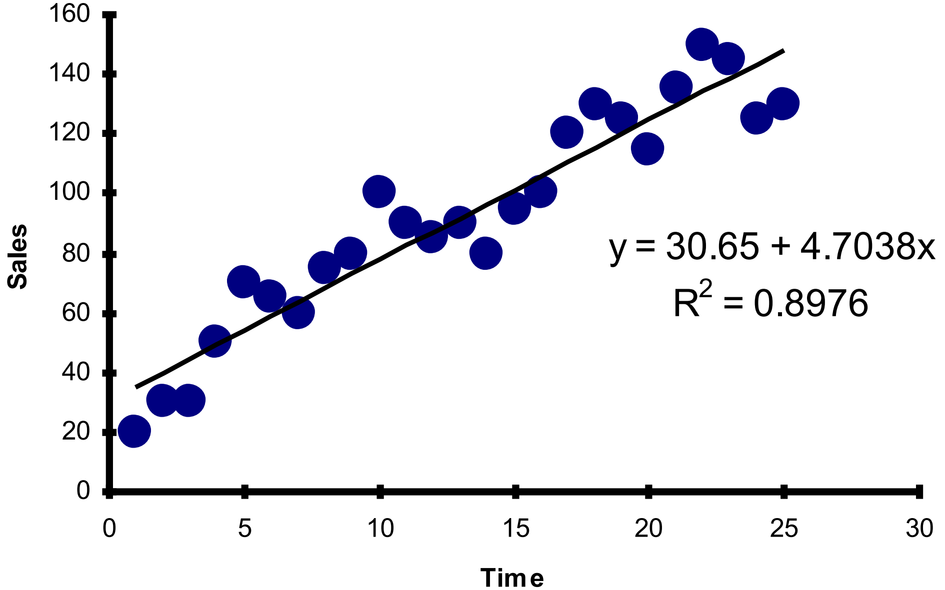
  (for sample size **n** and number of independent variables **k.**)

Decision rule: reject H_0 if $D < d_L$



Testing for Positive Autocorrelation

- Suppose we have the following time series data:



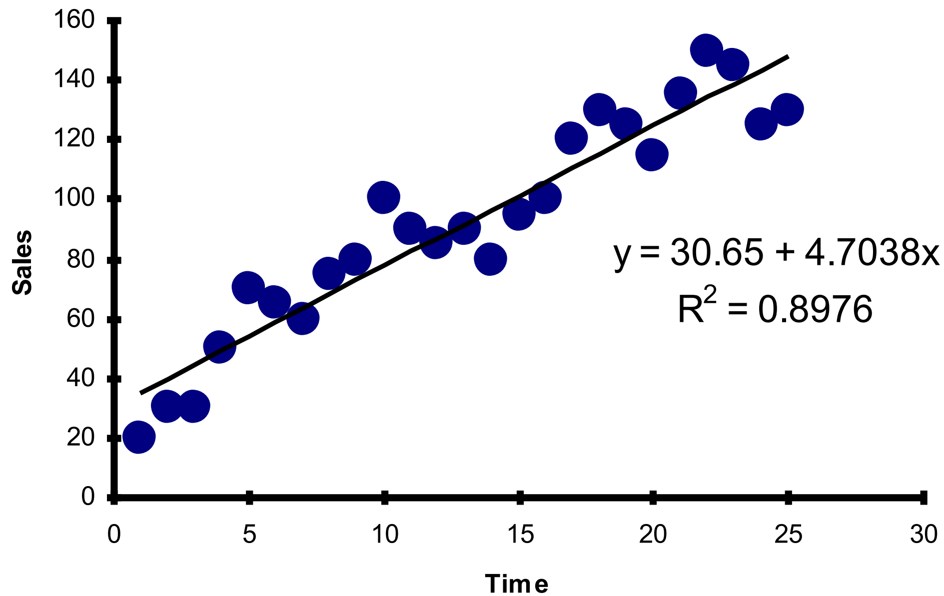
- Is there autocorrelation?

Testing for Positive Autocorrelation

- Example with n = 25:

Excel output:

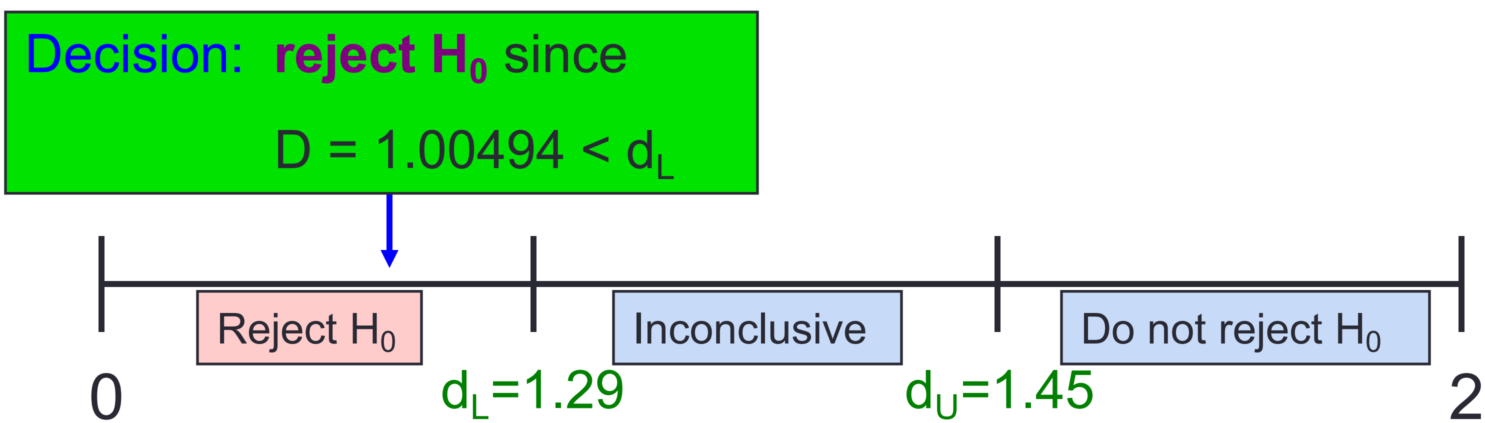
Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	3296.18
Sum of Squared Residuals	3279.98
Durbin-Watson Statistic	**1.00494**



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

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

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

From Excel output:

	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

b_1

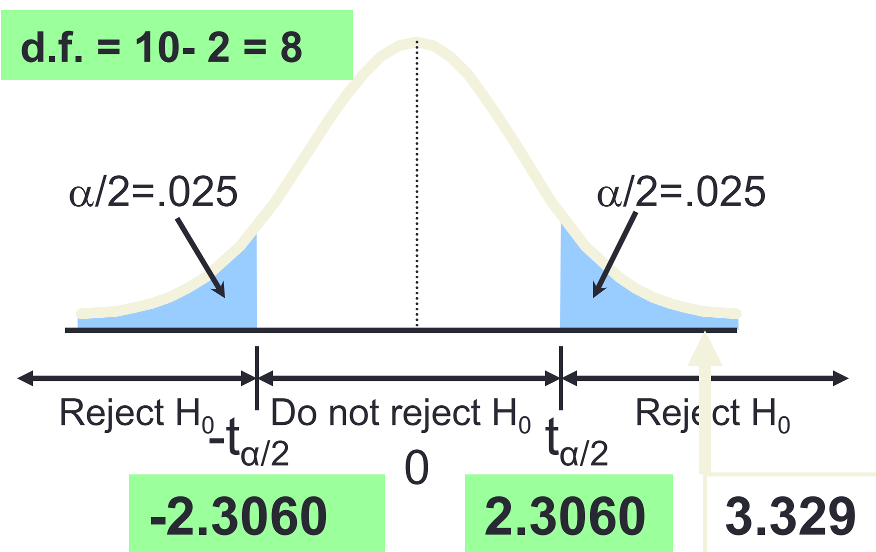
S_{b_1}

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

Inferences About the Slope: t Test Example

Test Statistic: $t_{STAT} = 3.329$

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.

Inferences About the Slope: t Test Example

H_0: $\beta_1 = 0$

From Excel output:

H_1: $\beta_1 \neq 0$

	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

p-value

Decision: Reject H_0, since p-value < α.

There is sufficient evidence that square footage affects house price.

F Test for The Slope

- F Test statistic:

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

where

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

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

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 The Slope: Excel Output

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

$$F_{STAT} = \frac{MSR}{MSE} = \frac{18{,}934.9348}{1{,}708.1957} = 11.0848$$

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			

F Test for The Slope

H_0: $\beta_1 = 0$

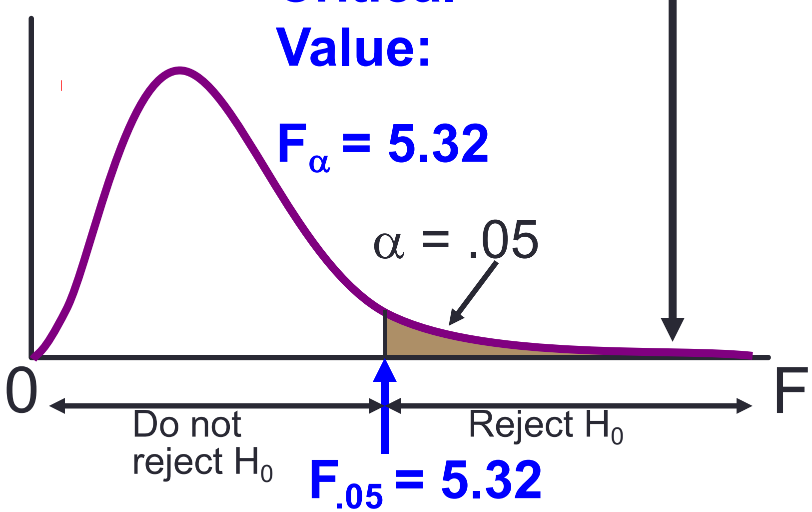
H_1: $\beta_1 \neq 0$

$\alpha = .05$

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

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.

Confidence Interval Estimate for the Slope

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).

Confidence Interval Estimate for the Slope

	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 $1,000s, we are 95% confident that the average impact on sales price is between $33.74 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.

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

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

(with n – 2 degrees of freedom)

where

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

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

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)

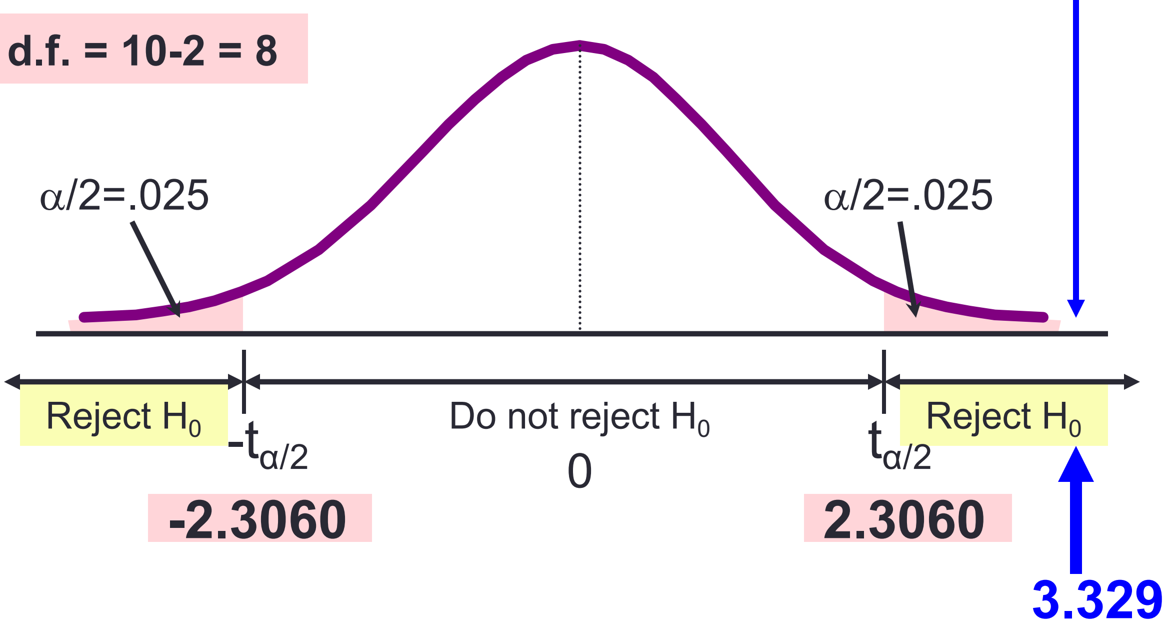
H_1: $\rho \neq 0$ (correlation exists)

$\alpha = .05$, 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

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

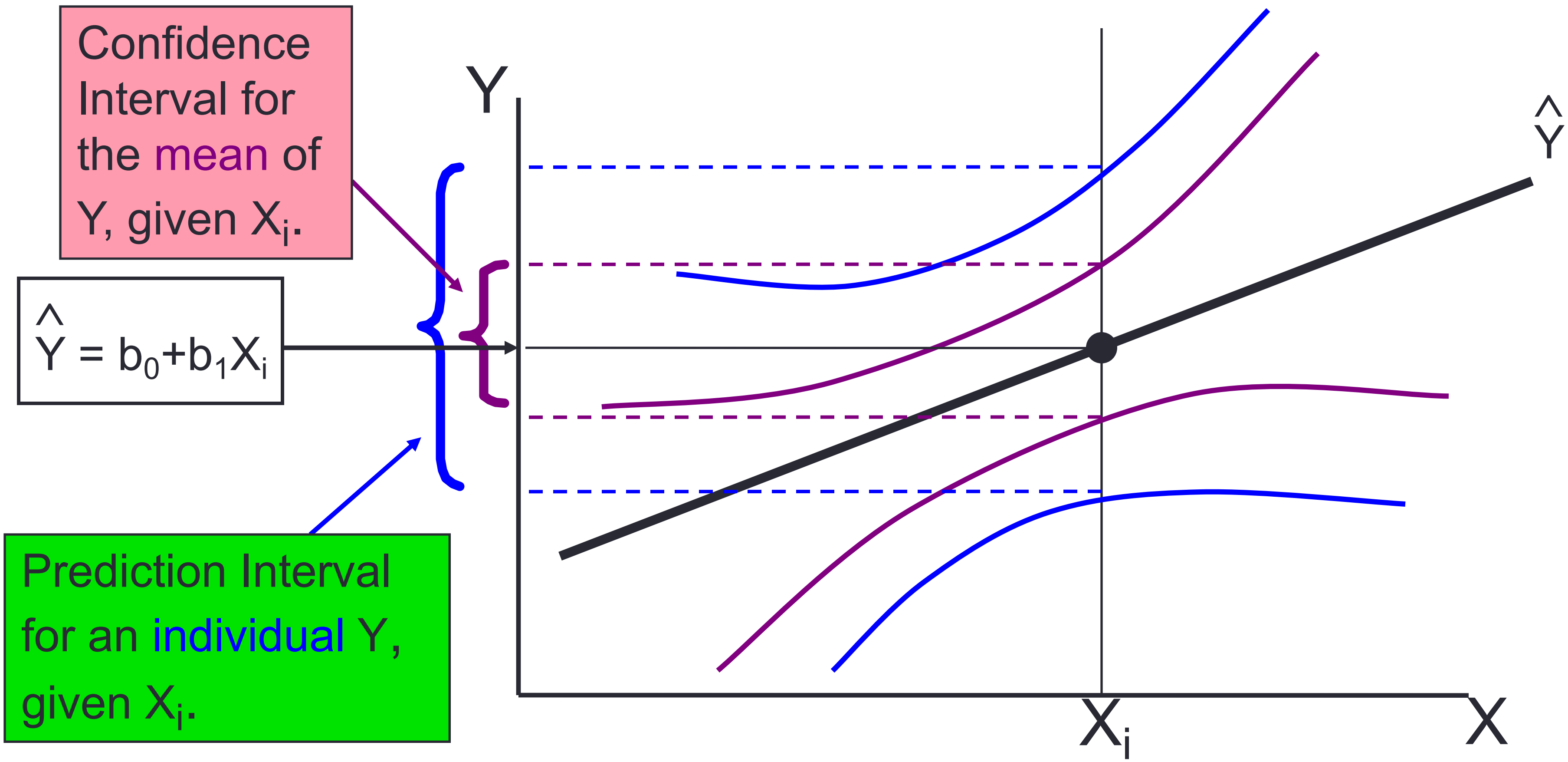


Decision:
Reject H_0.

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 Mean of Y, Given X

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

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

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

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

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

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 $Y_{X=X_i}$:

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i}$$

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)

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints (from Excel) 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)

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 102.28$$

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

Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions of least-squares regression.
- Not knowing how to evaluate the assumptions of least-squares regression.
- 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.
- Concluding that a significant relationship identified always reflects a cause-and-effect relationship.

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.
- Refrain from making predictions or forecasts outside the relevant range.
- Remember that the relationships identified in observational studies may or may not be due to cause-and-effect relationships.

Module Summary

In this module we discussed:

- How to use regression analysis to predict the value of a dependent variable based on a value of an independent variable.
- Understanding the meaning of the regression coefficients b_0 and b_1.
- Evaluating the assumptions of regression analysis and know what to do if the assumptions are violated.
- Making inferences about the slope and correlation coefficient.
- Estimating mean values and predicting individual values.