Linear Regression

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

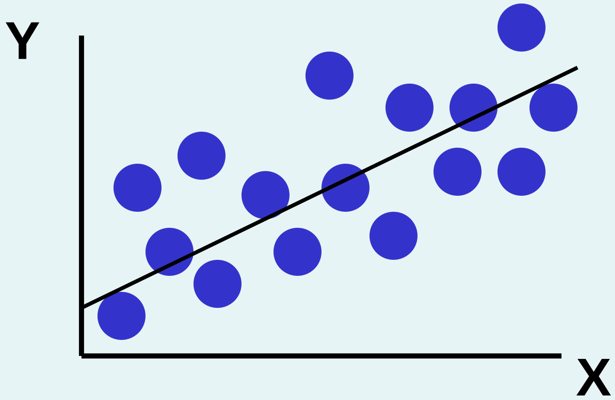
Independent variable: the variable used to predict or explain the dependent variable

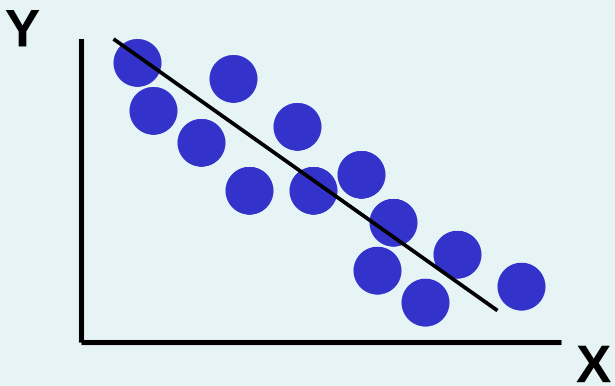
Simple Linear Regression Model

- 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

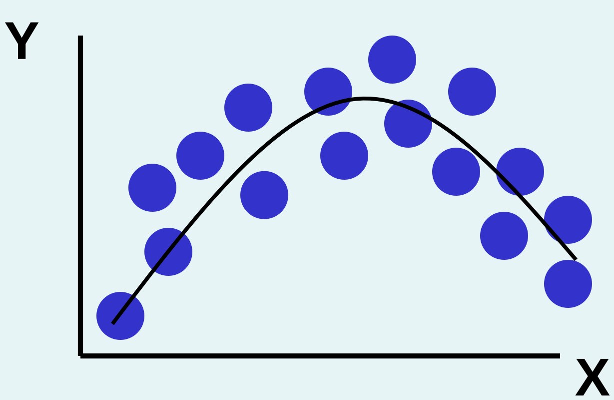
Types of Relationships

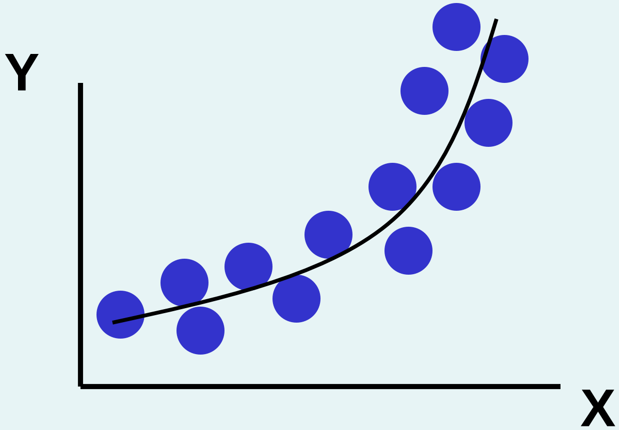
Linear relationships





Nonlinear relationships

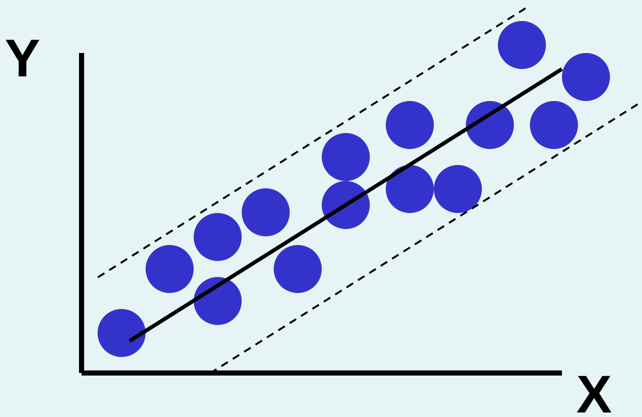


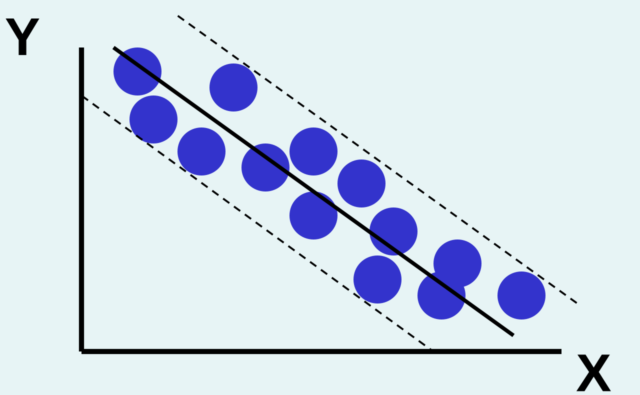


Types of Relationships

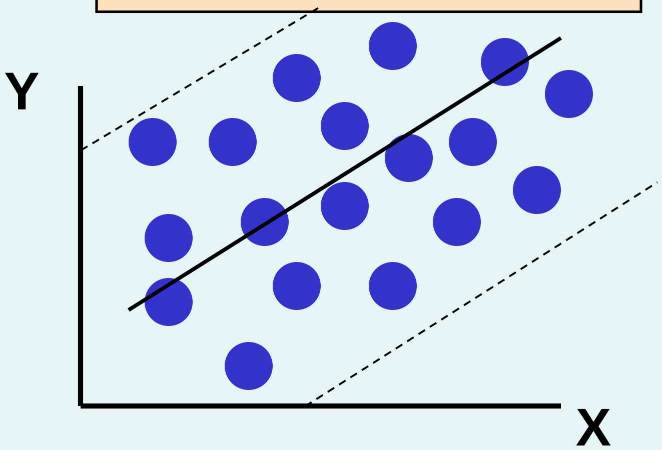
(continued)

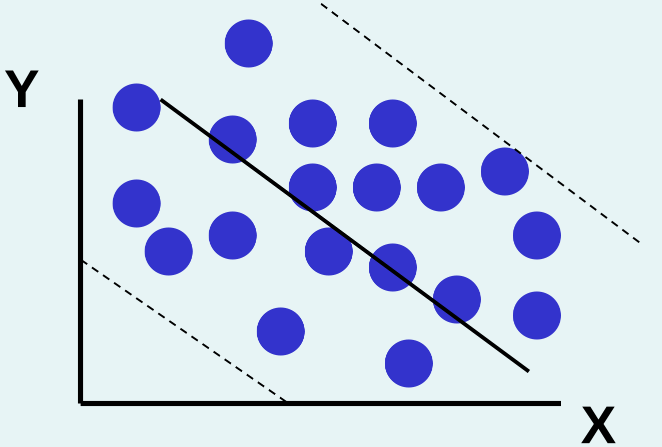
Strong relationships





Weak relationships

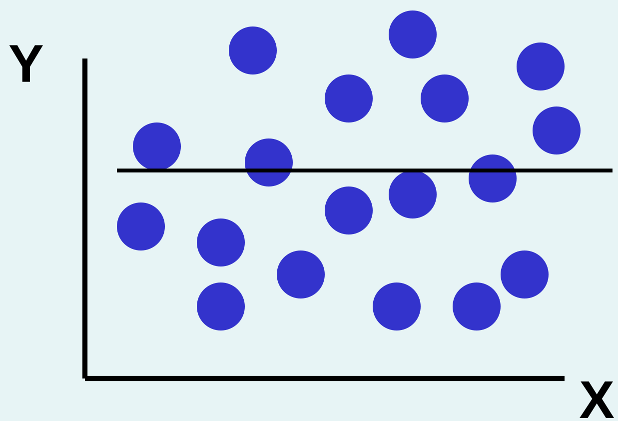


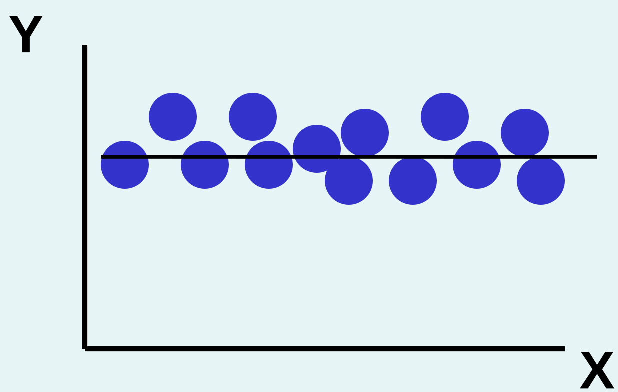


Types of Relationships

(continued)

No relationship





Simple Linear Regression Model

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

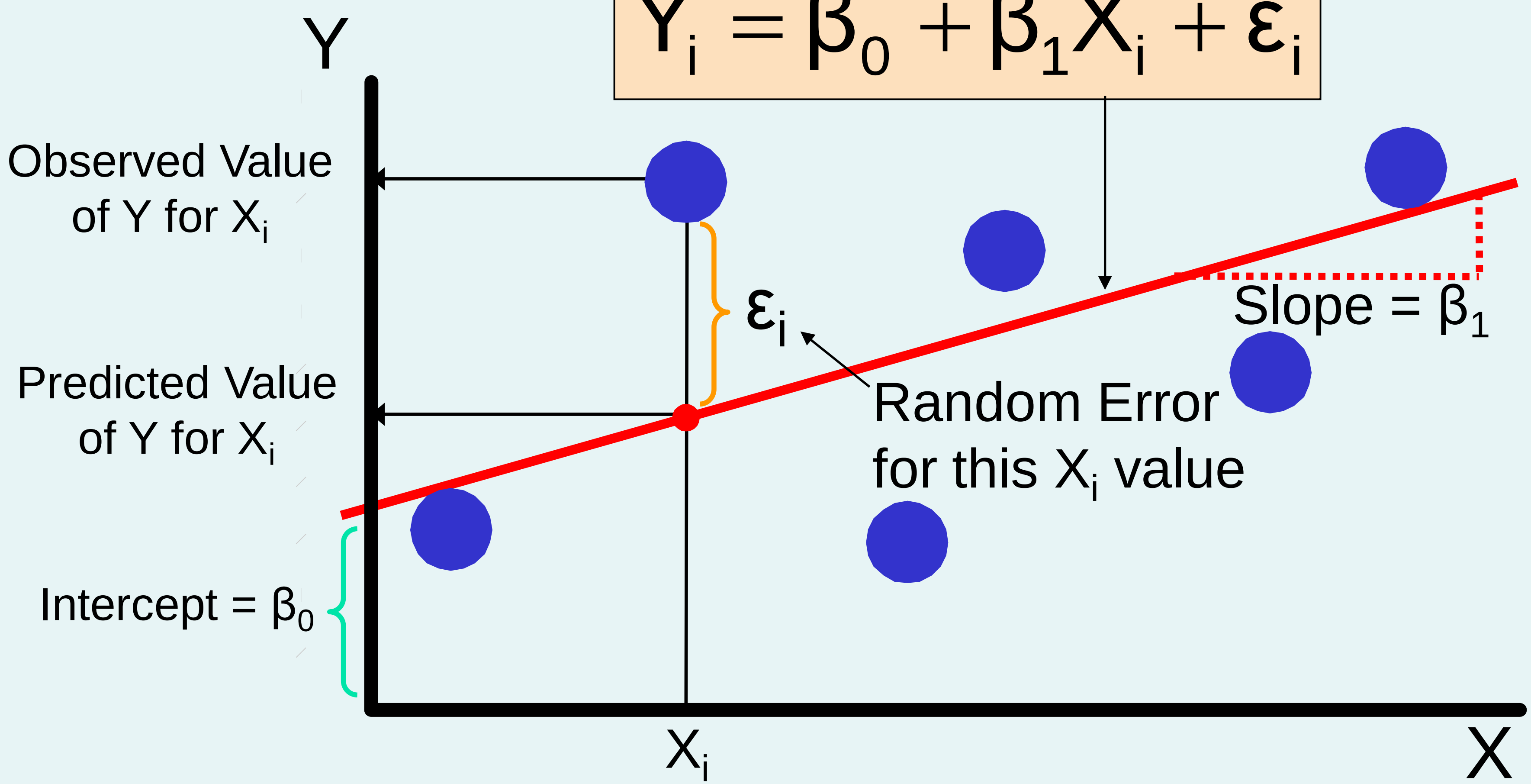
Random Error term

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

Simple Linear Regression Model

(continued)

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



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

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

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 increase in X

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 $\hat{Y}$:

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

The Least Squares Estimates

Slope $$b_1 = \frac{SSXY}{SSX}$$

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

where

$$SSXY = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y}$$

$$SSX = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^{\,2}$$

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)

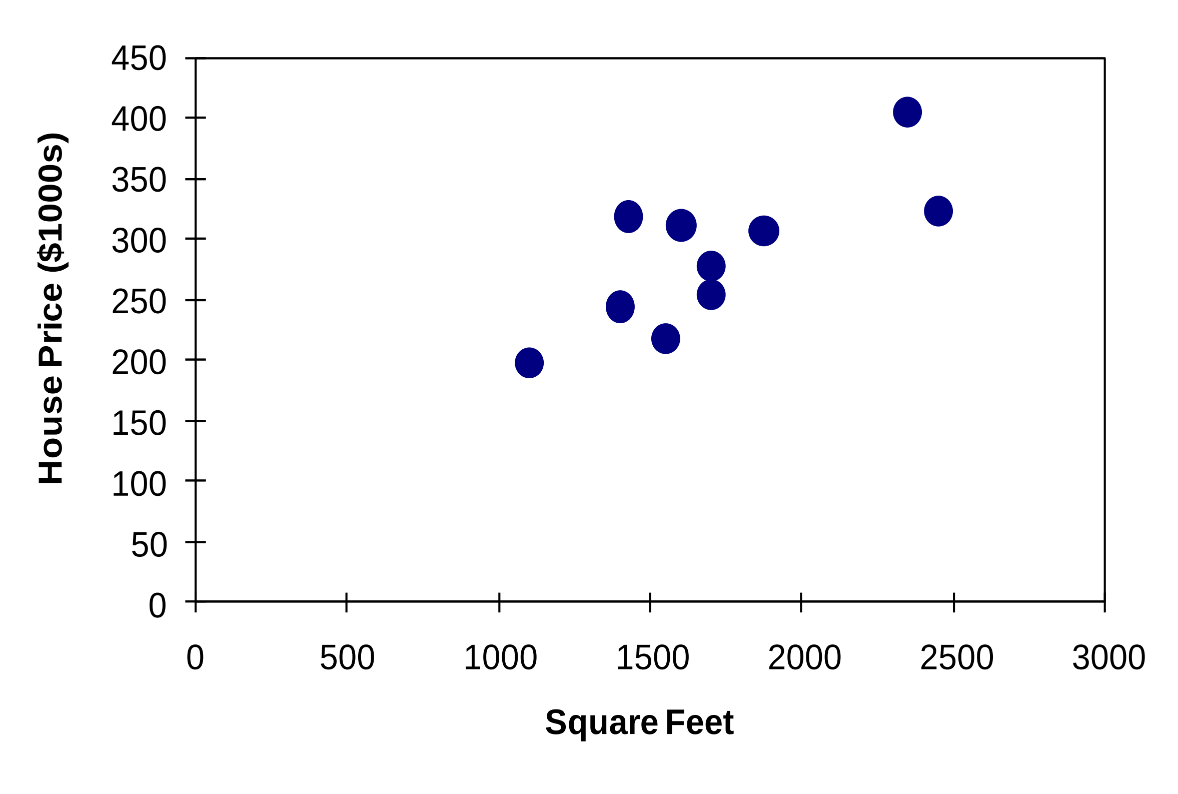
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

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: 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:

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

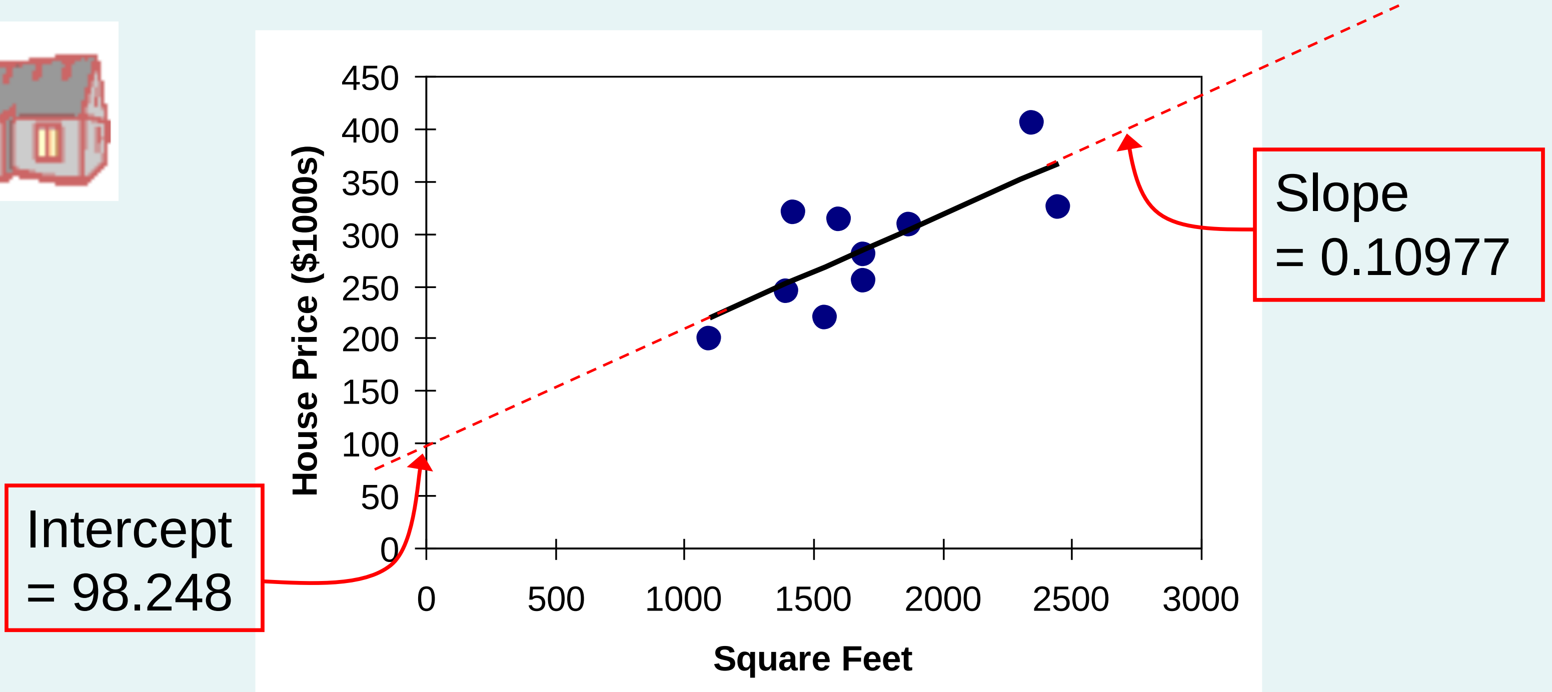
ANOVA

	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 average 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 average 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 .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:

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

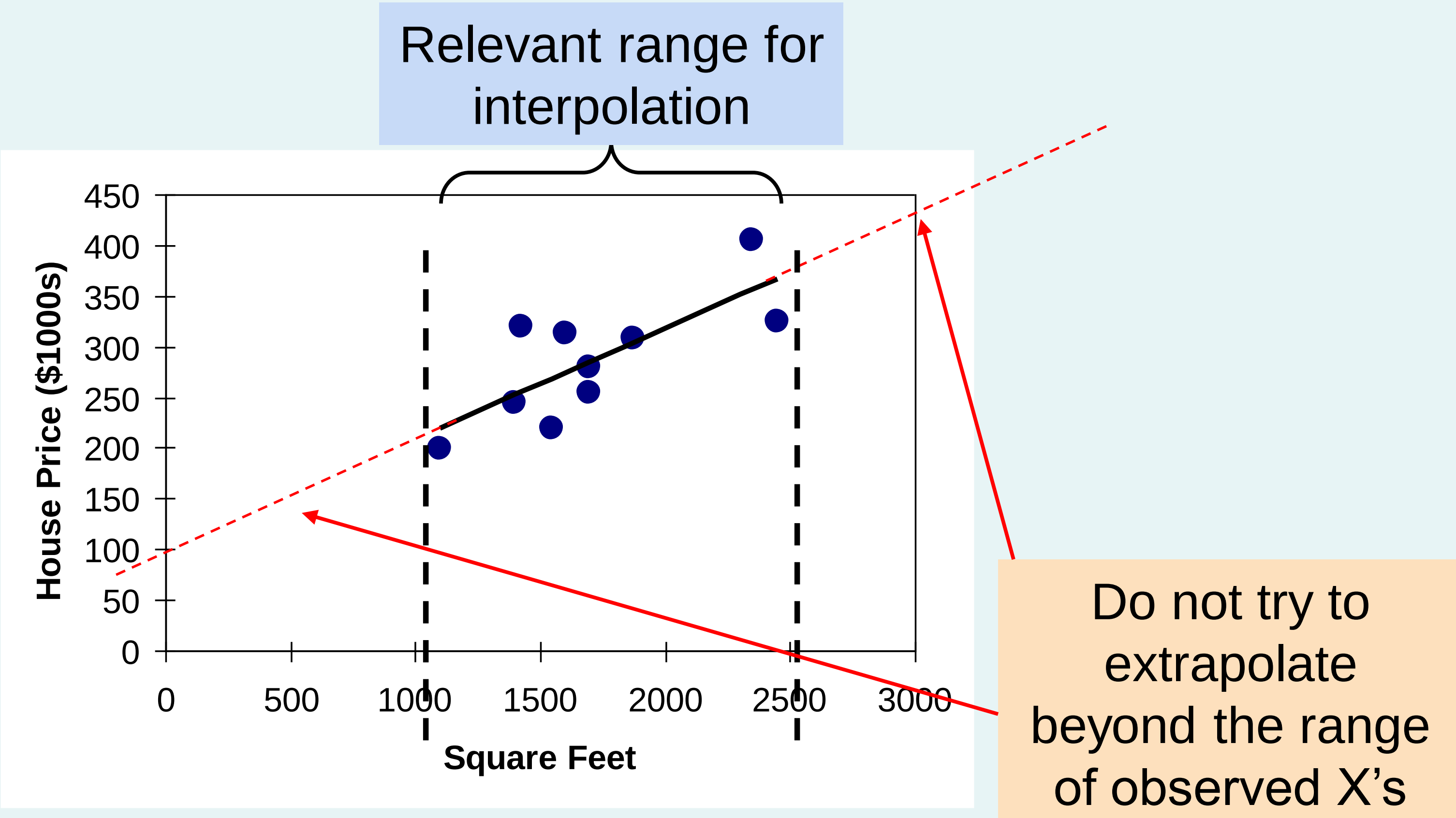
$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

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

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

$$SST = \sum (Y_i - \bar{Y})^2$$

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

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

where:

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

Y_i = Observed value of the dependent variable

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

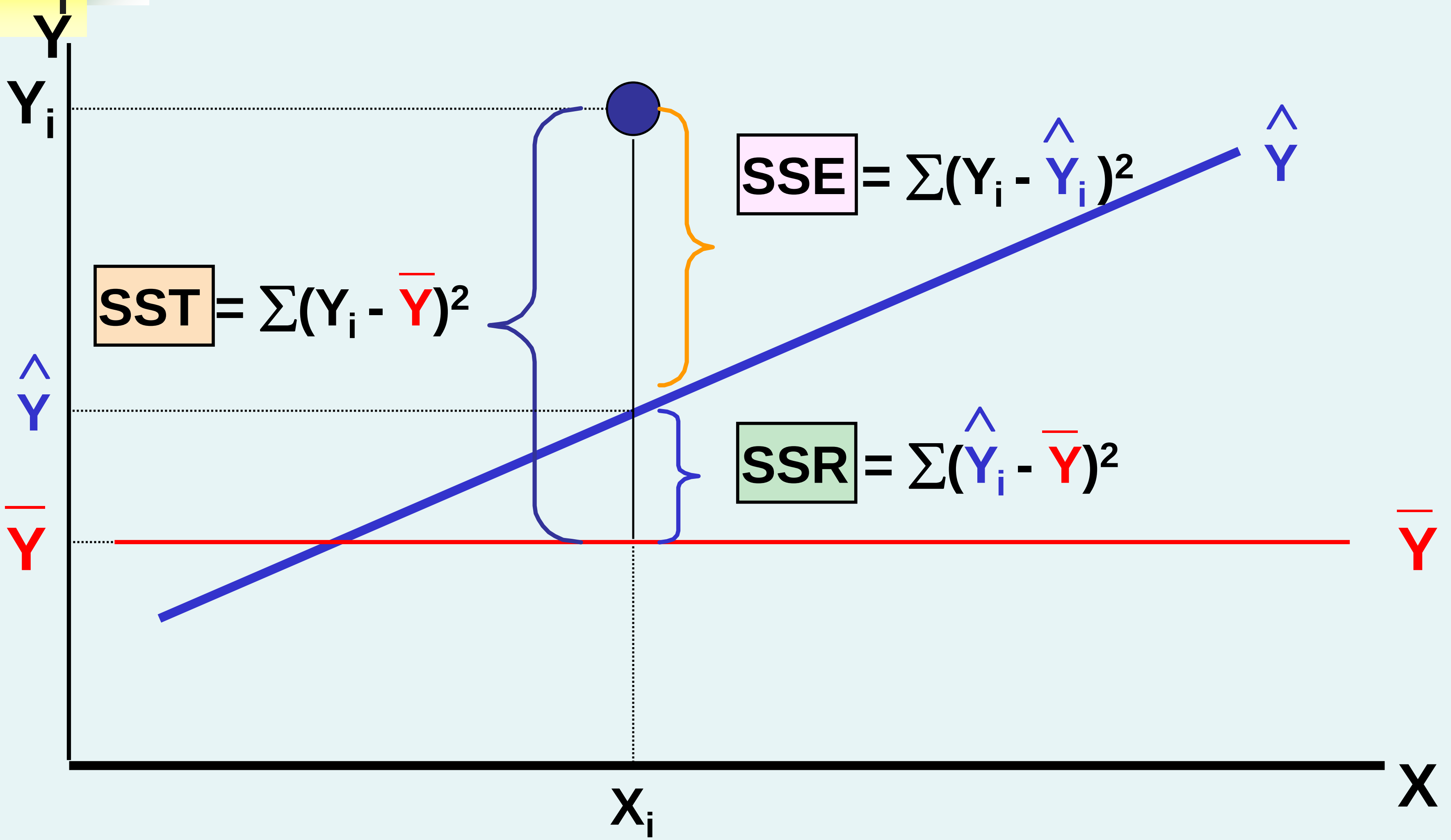
Measures of Variation

(continued)

- 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

(continued)



Coefficient of Determination, r^2

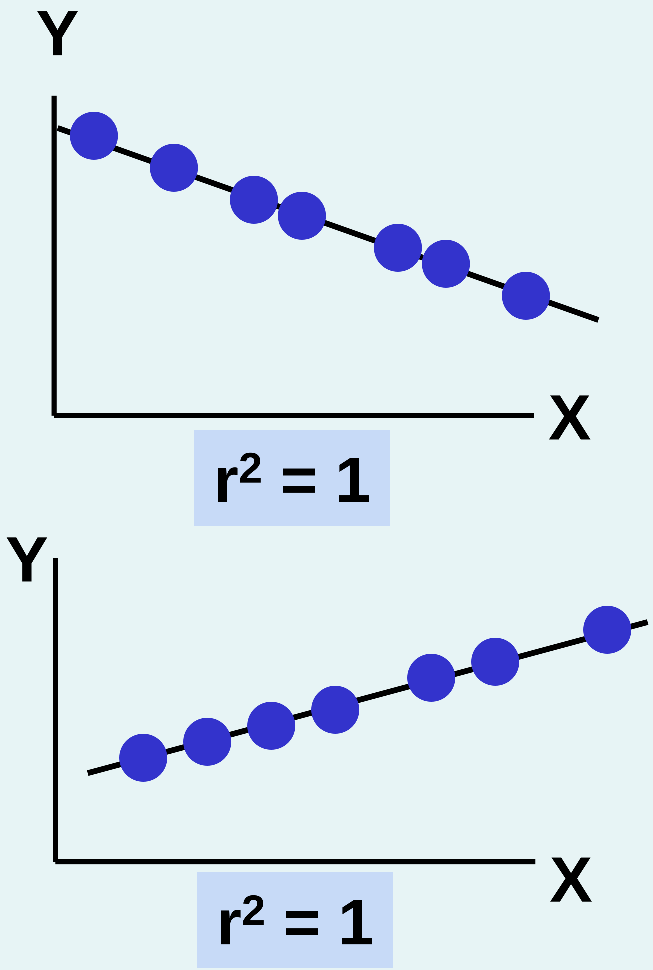
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called r-squared and is denoted as r^2

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

$$0 \leq r^2 \leq 1$$

- r^2 is also the sample correlation coefficient

Examples of Approximate r^2 Values

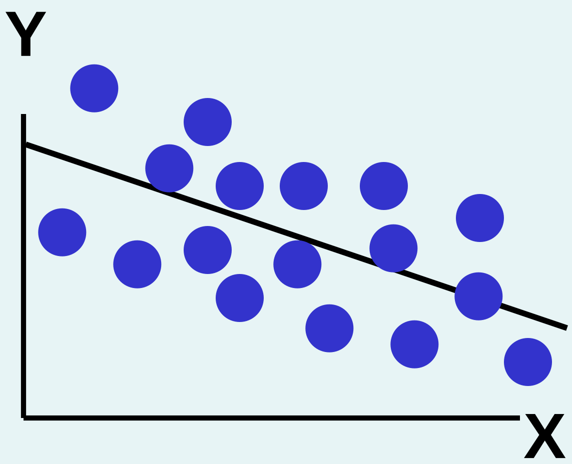


$r^2 = 1$

Perfect linear relationship between X and Y:

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

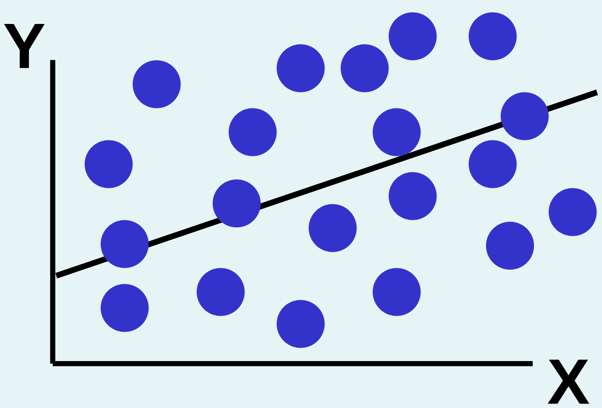
Examples of Approximate r^2 Values



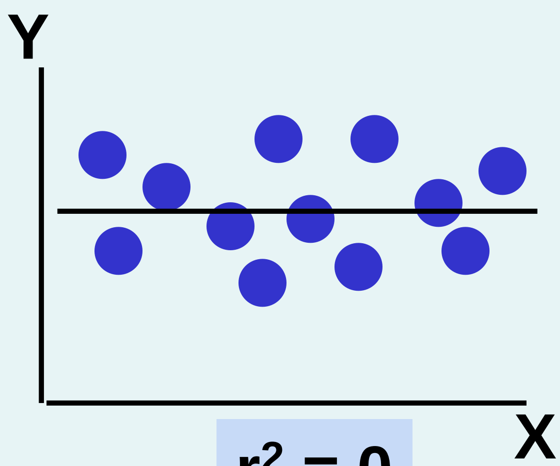
$0 < r^2 < 1$

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



$r^2 = 0$

$r^2 = 0$

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

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

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

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

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

$$S_{YX} = 41.33032$$

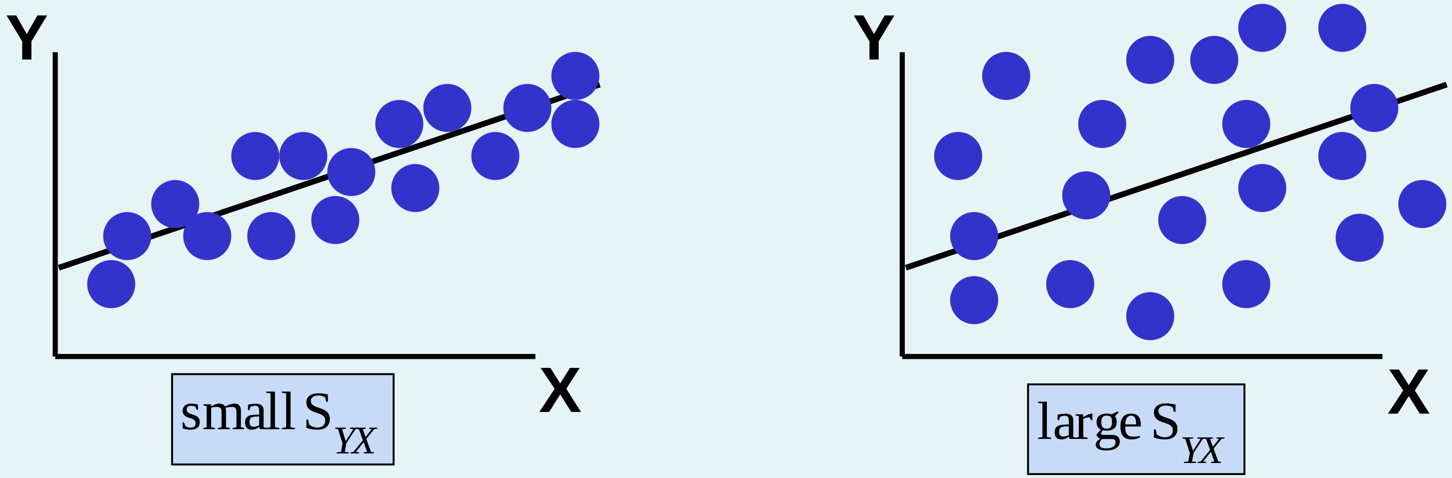
ANOVA

	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

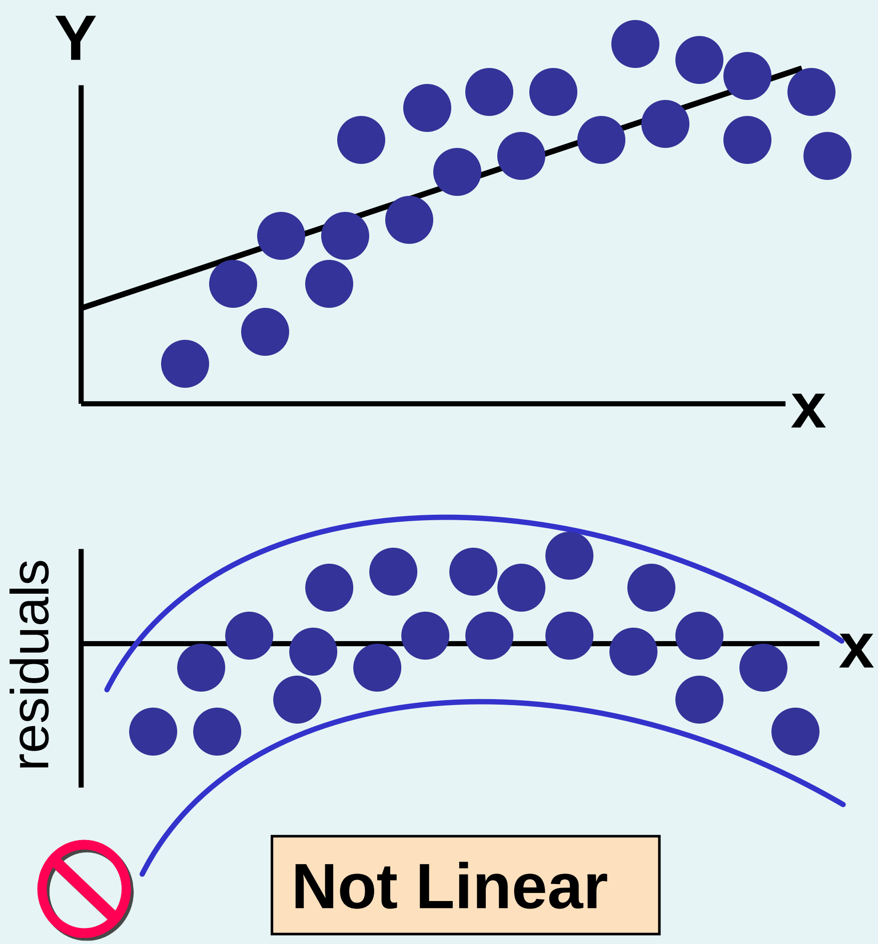
- Linearity
  - The relationship between X and Y is linear
- Independence of Errors
  - Error values are statistically independent
- Normality of Error
  - Error values are normally distributed for any given value of X
- Equal Variance (also called homoscedasticity)
  - The probability distribution of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

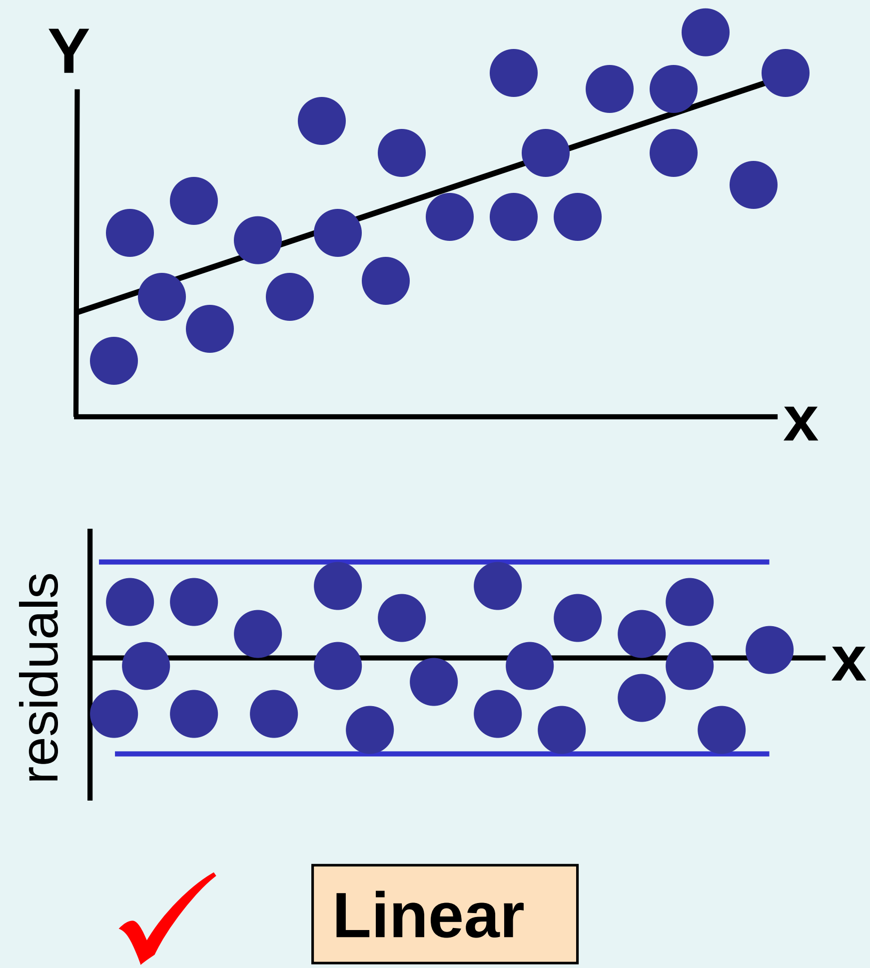
Residual Analysis

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

- 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 for all levels of X (homoscedasticity)
- Graphical Analysis of Residuals
  - Can plot residuals vs. X

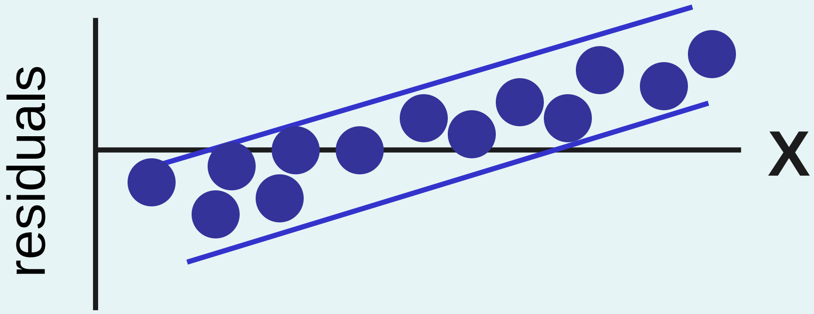
Residual Analysis for Linearity

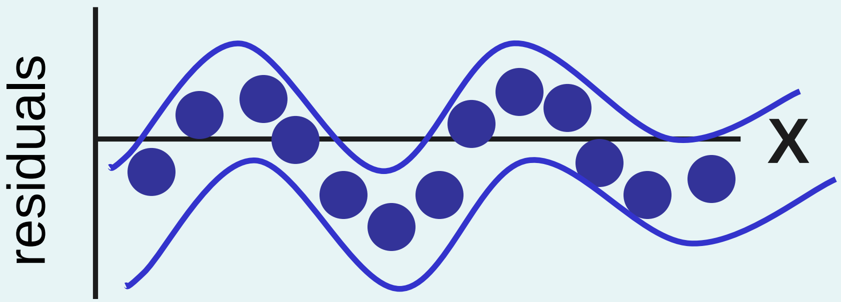




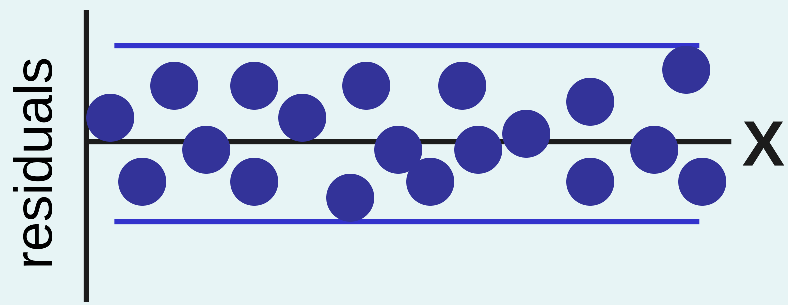
Residual Analysis for Independence

Not Independent





✓ Independent

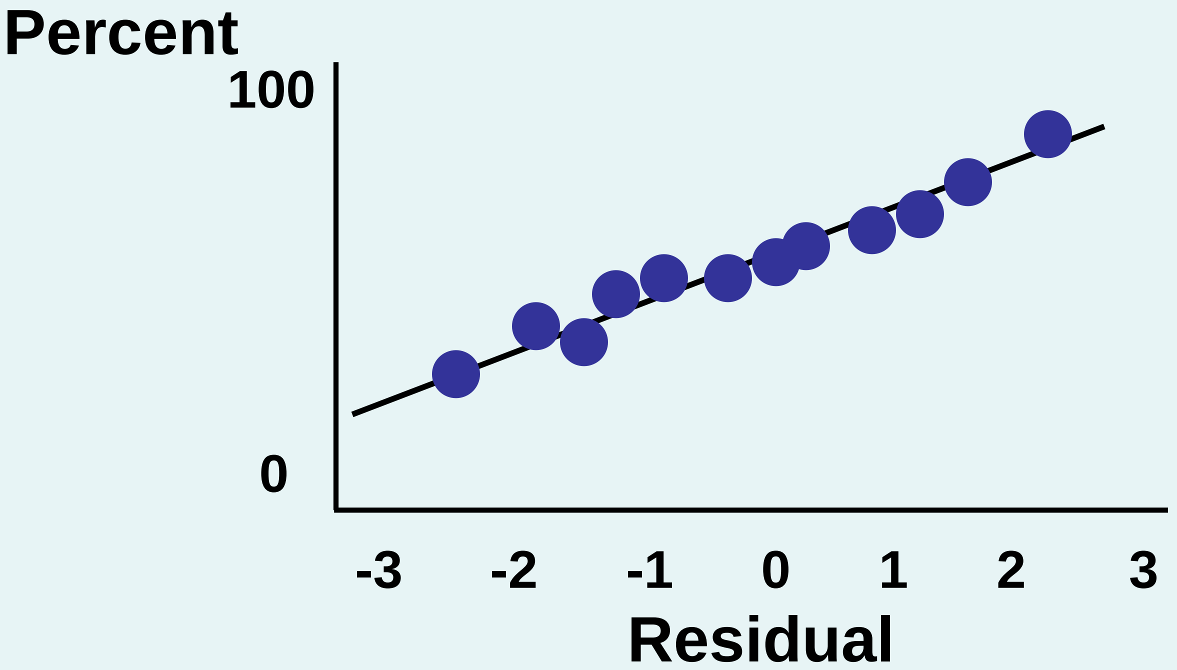


Checking for Normality

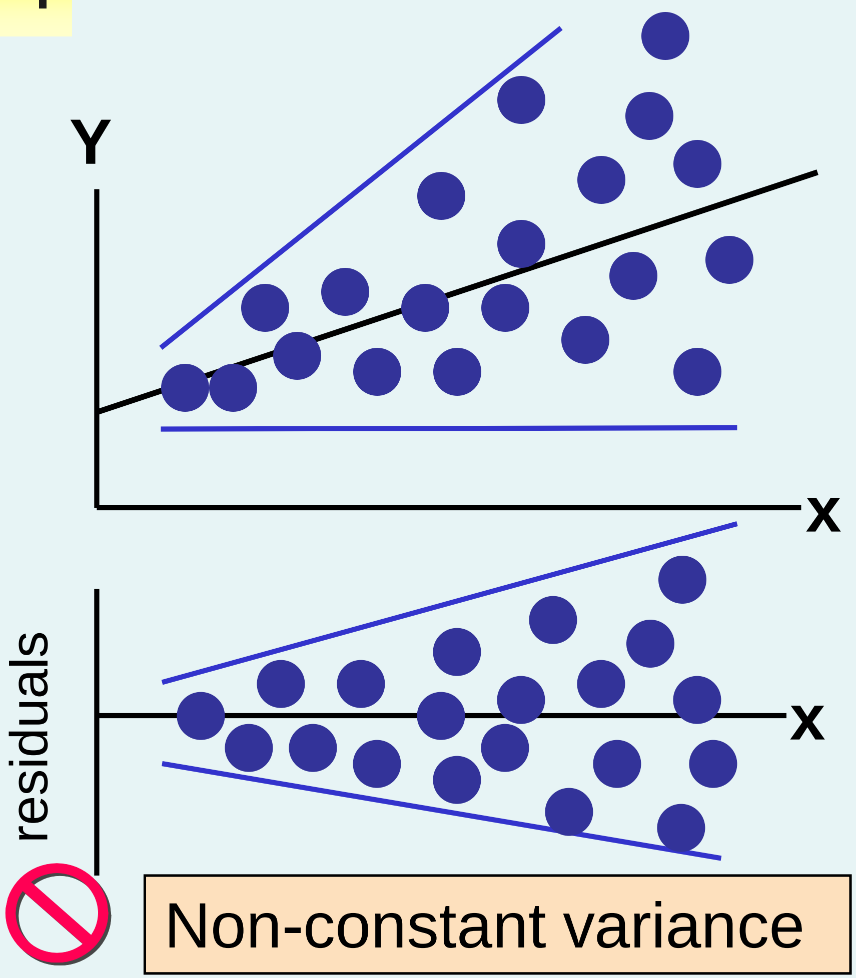
- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Boxplot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals

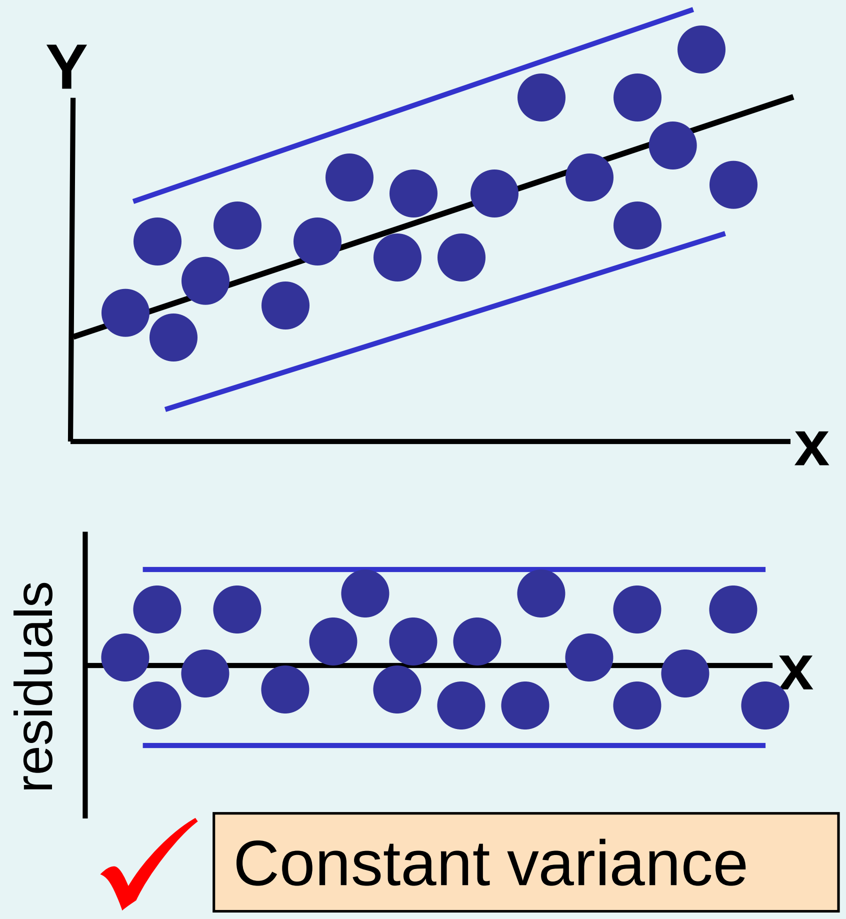
Residual Analysis for Normality

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



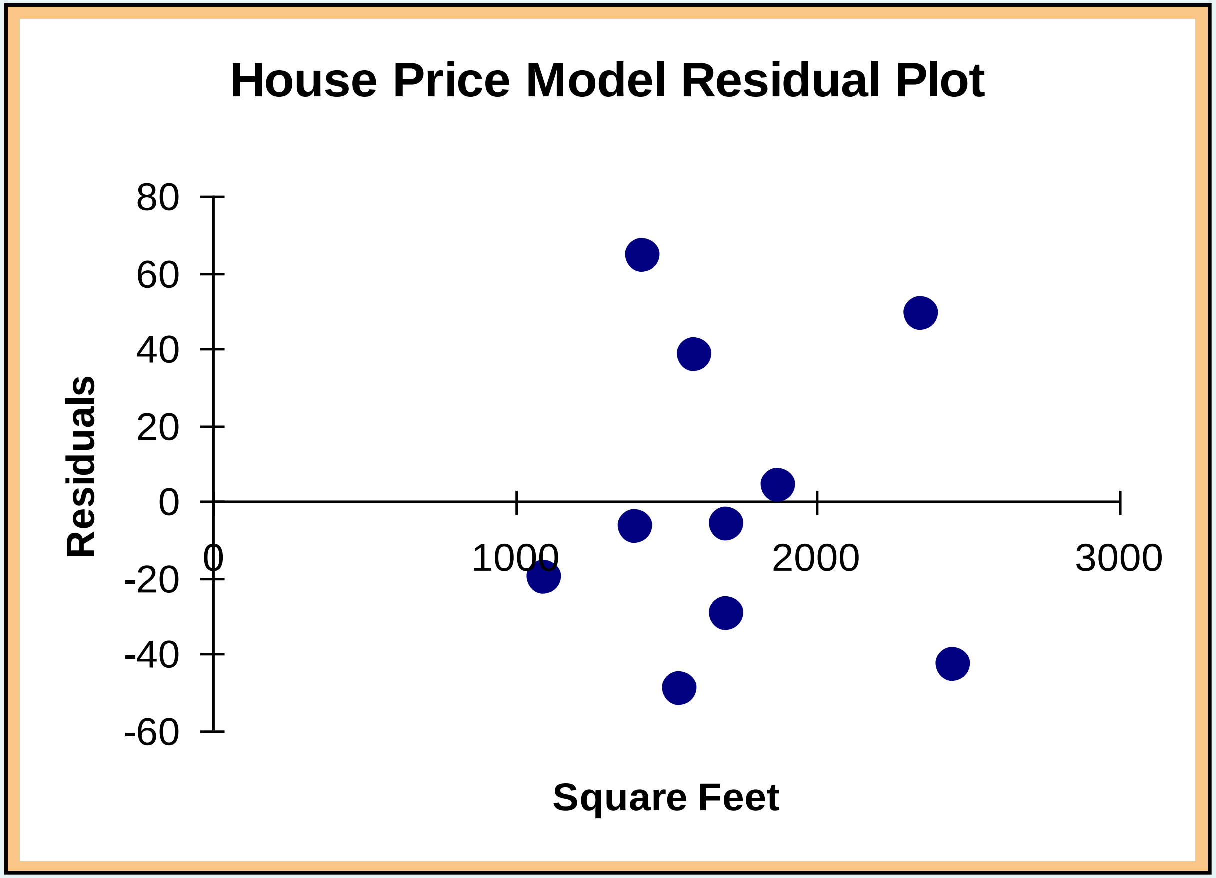
Residual Analysis for Equal Variance





Simple Linear Regression Example: 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

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

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

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

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

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**

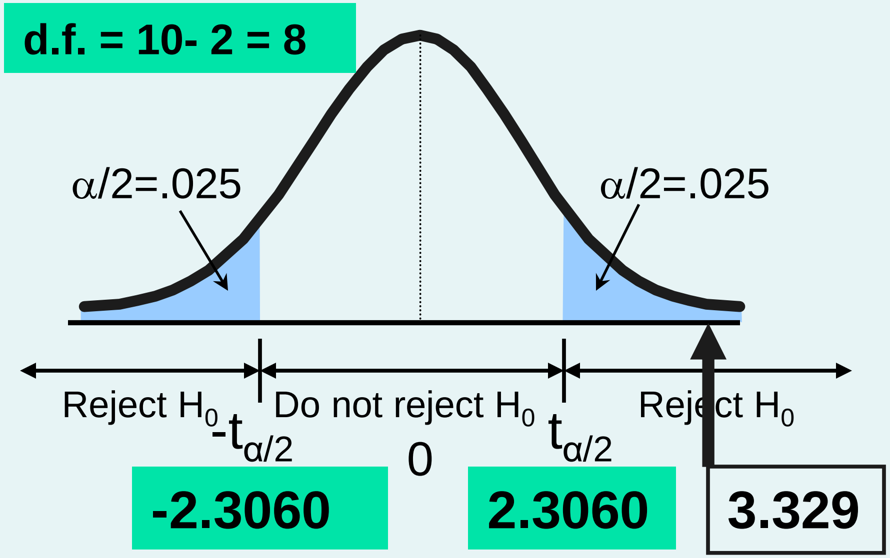
b_1 S_{b_1}

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

Inferences About the Slope: t Test Example

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

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



Decision: Reject H_0

There is sufficient evidence that square footage affects house price

Inferences About the Slope: t Test Example

H_0: $\beta_1 = 0$

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**

p-value

Decision: Reject H_0, since p-value $< \alpha$

There is sufficient evidence that square footage affects house price.

F Test for Significance

- 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 Significance Excel Output

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

$$F_{STAT} = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$

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			

F Test for Significance

(continued)

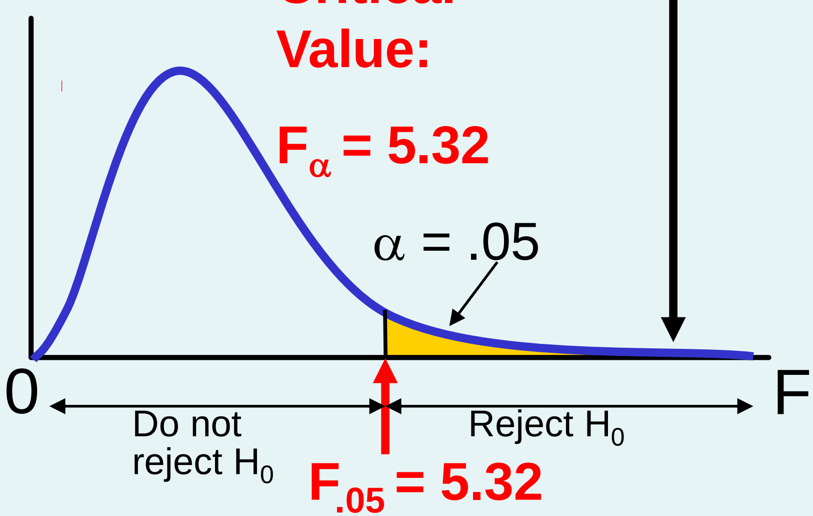
H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

$\alpha = .05$

$df_1 = 1 \qquad df_2 = 8$

Critical Value:



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

(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**

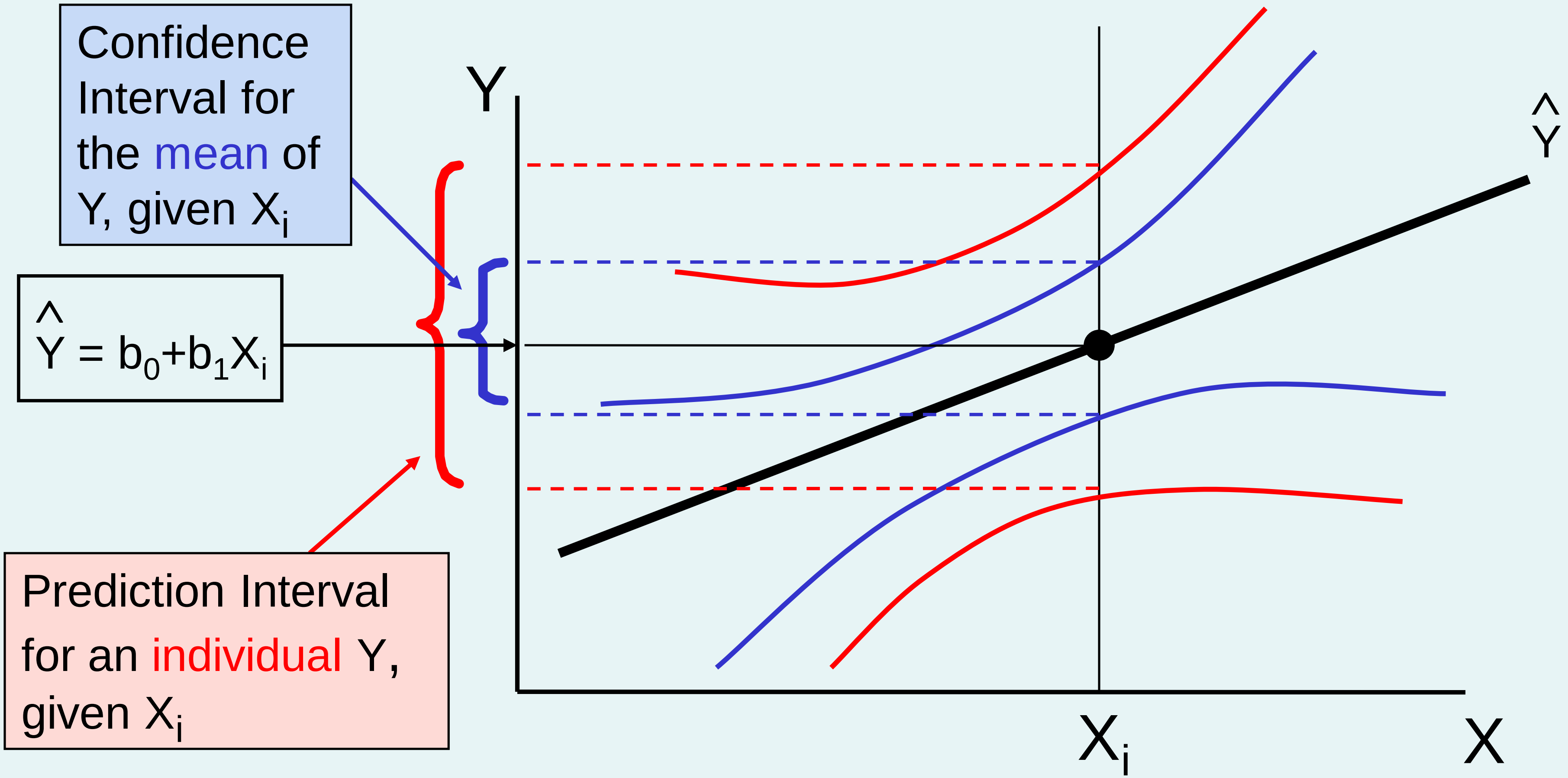
Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.74 and $185.80 per square foot of house size

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

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

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

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

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

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

Prediction Interval for an Individual Y, Given X

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

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

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

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

Multivariate Regression

The Multiple Regression Model

Idea: Examine the linear relationship between
1 dependent (Y) & *2 or more independent variables (X_i)*

Multiple Regression Model with k Independent Variables:

Y-intercept

Population slopes

Random Error

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Multiple Regression Equation

The coefficients of the multiple regression model are estimated using sample data

Multiple regression equation with k independent variables:

Estimated (or predicted) value of Y

Estimated intercept

Estimated slope coefficients

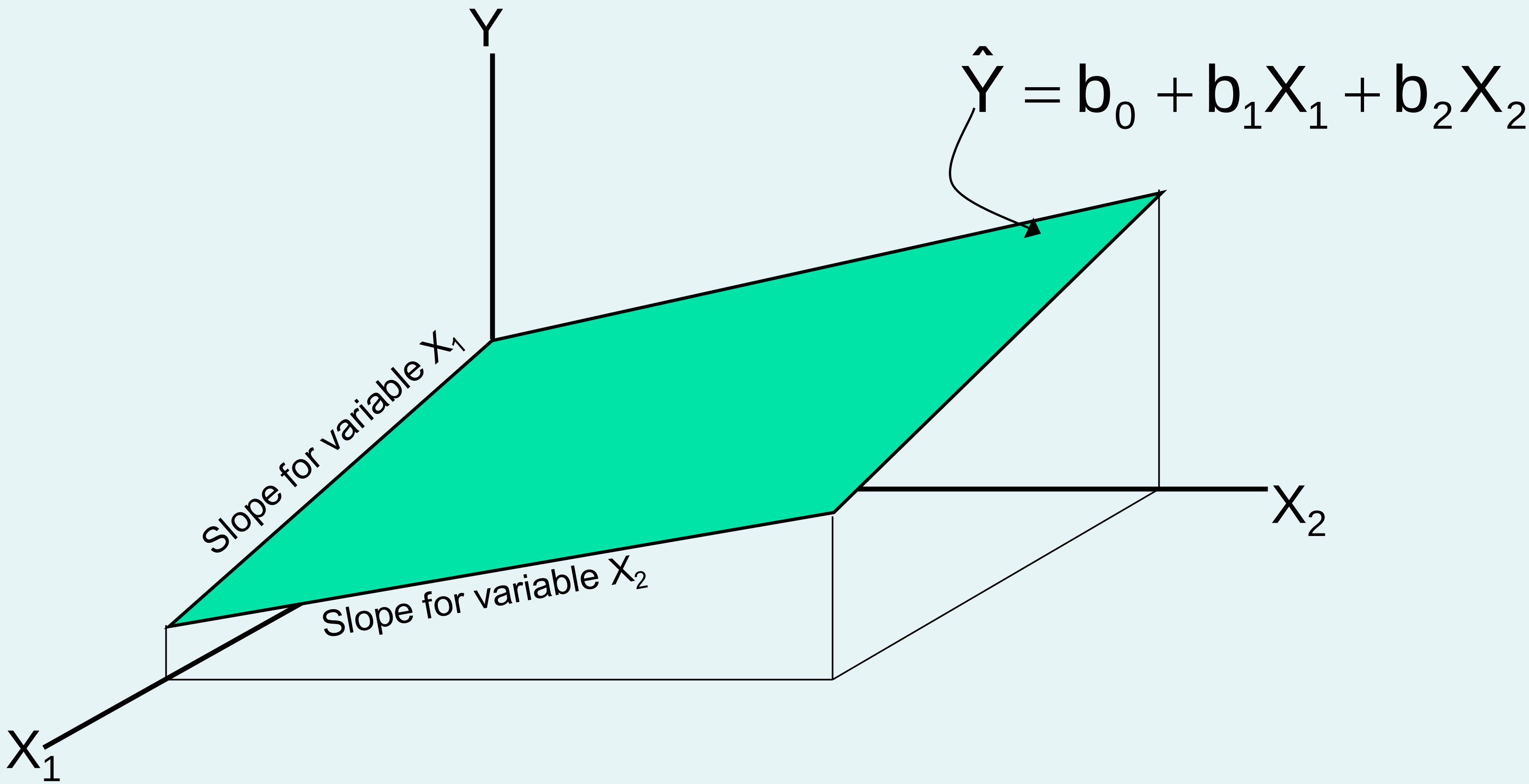
$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$

In this chapter we will use Excel to obtain the regression slope coefficients and other regression summary measures.

Multiple Regression Equation

(continued)

Two variable model



Example: 2 Independent Variables

- A distributor of frozen dessert pies wants to evaluate factors thought to influence demand
  - Dependent variable: Pie sales (units per week)
  - Independent variables: Price (in $)
    Advertising ($100's)
- Data are collected for 15 weeks

Pie Sales Example

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Multiple regression equation:

$$\widehat{\text{Sales}} = b_0 + b_1 (\text{Price}) + b_2 (\text{Advertising})$$

Excel Multiple Regression Output

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

The Multiple Regression Equation

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

where
Sales is in number of pies per week
Price is in \$
Advertising is in \$100's.

b_1 = -24.975: sales will decrease, on average, by 24.975 pies per week for each \$1 increase in selling price, net of the effects of changes due to advertising

b_2 = 74.131: sales will increase, on average, by 74.131 pies per week for each \$100 increase in advertising, net of the effects of changes due to price

Using The Equation to Make Predictions

Predict sales for a week in which the selling price is \$5.50 and advertising is \$350:

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$
$$= 306.526 - 24.975(5.50) + 74.131(3.5)$$
$$= 428.62$$

Predicted sales is 428.62 pies

Note that Advertising is in \$100s, so \$350 means that $X_2 = 3.5$

Coefficient of Multiple Determination

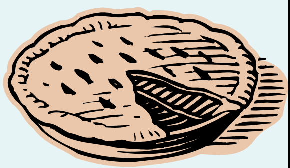
- Reports the proportion of total variation in Y explained by all X variables taken together

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

Multiple Coefficient of Determination In Excel

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$



52.1% of the variation in pie sales is explained by the variation in price and advertising

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Adjusted r^2

- r^2 never decreases when a new X variable is added to the model
  - This can be a disadvantage when comparing models
- What is the net effect of adding a new variable?
  - We lose a degree of freedom when a new X variable is added
  - Did the new X variable add enough explanatory power to offset the loss of one degree of freedom?

Adjusted r^2

(continued)

- Shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used

$$r_{adj}^2 = 1 - \left[(1 - r^2) \left(\frac{n-1}{n-k-1} \right) \right] = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

(where n = sample size, k = number of independent variables)

- Penalizes excessive use of unimportant independent variables
- Smaller than r^2
- Useful in comparing among models

Adjusted r^2 in Excel

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$r^2_{adj} = .44172$$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Is the Model Significant?

- F Test for Overall Significance of the Model
- Shows if there is a linear relationship between all of the X variables considered together and Y
- Use F-test statistic
- Hypotheses:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (no linear relationship)

H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

F Test for Overall Significance

- Test statistic:

$$\boxed{F_{STAT} = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}}$$

where F_{STAT} has numerator d.f. = k and

denominator d.f. = (n – k - 1)

F Test for Overall Significance In Excel

(continued)

Regression Statistics	
Multiple R	**0.72213**
R Square	**0.52148**
Adjusted R Square	**0.44172**
Standard Error	**47.46341**
Observations	**15**

$$F_{STAT} = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F Test

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	**2**	**29460.027**	**14730.013**	**6.53861**	**0.01201**
Residual	**12**	**27033.306**	**2252.776**		
Total	**14**	**56493.333**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**306.52619**	**114.25389**	**2.68285**	**0.01993**	**57.58835**	**555.46404**
Price	**-24.97509**	**10.83213**	**-2.30565**	**0.03979**	**-48.57626**	**-1.37392**
Advertising	**74.13096**	**25.96732**	**2.85478**	**0.01449**	**17.55303**	**130.70888**

F Test for Overall Significance

(continued)

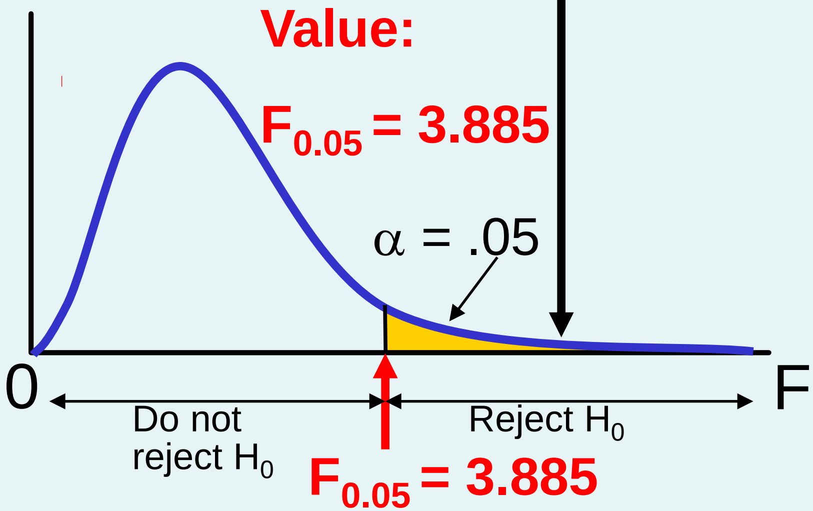
H_0: $\beta_1 = \beta_2 = 0$

H_1: β_1 and β_2 not both zero

$\alpha = .05$

$df_1 = 2 \qquad df_2 = 12$

Critical Value:



Test Statistic:

$$F_{STAT} = \frac{MSR}{MSE} = 6.5386$$

Decision:

Since F_{STAT} test statistic is in the rejection region (p-value < .05), reject H_0

Conclusion:

There is evidence that at least one independent variable affects Y

Are Individual Variables Significant?

- Use t tests of individual variable slopes
- Shows if there is a linear relationship between the variable X_j and Y holding constant the effects of other X variables
- Hypotheses:
  - H_0: $\beta_j = 0$ (no linear relationship)
  - H_1: $\beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Are Individual Variables Significant?

(continued)

H_0: β_j = 0 (no linear relationship)

H_1: $\beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Test Statistic:

$$\boxed{t_{STAT} = \frac{b_j - 0}{S_{b_j}}}$$

(df = n – k – 1)

Are Individual Variables Significant? Excel Output *(continued)*

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

t Stat for Price is t_{STAT} = -2.306, with p-value .0398

t Stat for Advertising is t_{STAT} = 2.855, with p-value .0145

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Inferences about the Slope: t Test Example

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

d.f. = 15-2-1 = 12

α **= .05**

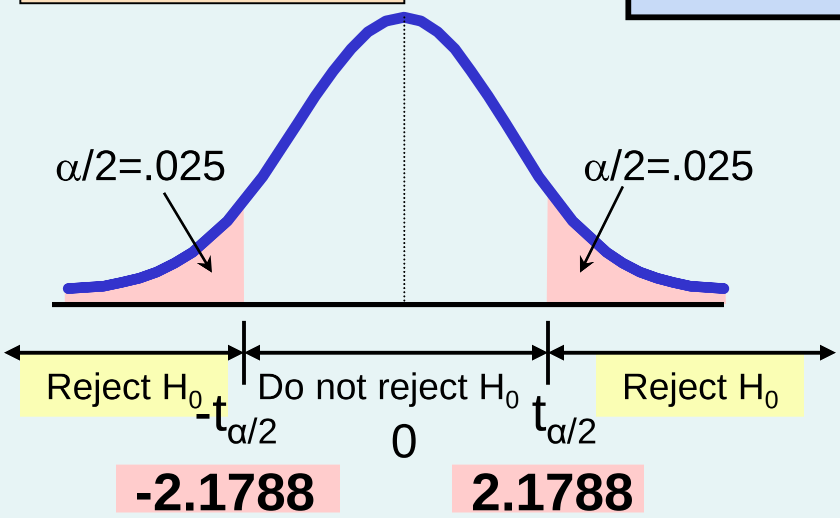
$t_{\alpha/2}$ **= 2.1788**

From the Excel output:

For Price t_{STAT} = -2.306, with p-value .0398

For Advertising t_{STAT} = 2.855, with p-value .0145

The test statistic for each variable falls in the rejection region (p-values < .05)



Decision:

Reject H_0 for each variable

Conclusion:

There is evidence that both Price and Advertising affect pie sales at α = .05

Confidence Interval Estimate for the Slope

Confidence interval for the population slope β_j

$$b_j \pm t_{\alpha/2} S_{b_j}$$

where t has (n – k – 1) d.f.

	Coefficients	Standard Error
Intercept	306.52619	114.25389
Price	-24.97509	10.83213
Advertising	74.13096	25.96732

Here, t has

(15 – 2 – 1) = 12 d.f.

Example: Form a 95% confidence interval for the effect of changes in price (X_1) on pie sales:

$$-24.975 \pm (2.1788)(10.832)$$

So the interval is (-48.576 , -1.374)

(This interval does not contain zero, so price has a significant effect on sales)

Confidence Interval Estimate for the Slope

(continued)

Confidence interval for the population slope β_j

	Coefficients	*Standard Error*	…	*Lower 95%*	*Upper 95%*
Intercept	**306.52619**	**114.25389**	**…**	**57.58835**	**555.46404**
Price	**-24.97509**	**10.83213**	**…**	**-48.57626**	**-1.37392**
Advertising	**74.13096**	**25.96732**	**…**	**17.55303**	**130.70888**

Example: Excel output also reports these interval endpoints:

Weekly sales are estimated to be reduced by between 1.37 to 48.58 pies for each increase of $1 in the selling price, holding the effect of advertising constant

Testing Portions of the Multiple Regression Model

- Contribution of a Single Independent Variable X_j

$$SSR(X_j \mid \text{all variables except } X_j)$$
$$= SSR\text{ (all variables)} - SSR(\text{all variables except } X_j)$$

  - This is extra Sum of Squares attributed to X_j
  - Measures the contribution of X_j in explaining the total variation in Y (SST)

Testing Portions of the Multiple Regression Model

(continued)

Contribution of a Single Independent Variable X_j, assuming all other variables are already included (consider here a 2-variable model):

$SSR(X_1 \mid X_2)$
$= SSR \text{ (all variables)} - SSR(X_2)$

From ANOVA section of regression for

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

From ANOVA section of regression for

$$\hat{Y} = b_0 + b_2 X_2$$

Measures the contribution of X_1 in explaining SST

The Partial F-Test Statistic

- Consider the hypothesis test:

H_0: variable X_j does not significantly improve the model after all other variables are included

H_1: variable X_j significantly improves the model after all other variables are included

- Test using the F-test statistic:

(with 1 and n-k-1 d.f.)

$$F_{STAT} = \frac{\text{SSR }(X_j \mid \text{all variables except j})}{\text{MSE}}$$

Testing Portions of Model: Example

Example: Frozen dessert pies

Test at the α = .05 level to determine whether the price variable significantly improves the model given that advertising is included

Testing Portions of Model: Example

(continued)

H_0: X_1 (price) does not improve the model with X_2 (advertising) included

H_1: X_1 does improve model

$\alpha = .05$, df = 1 and 12

$F_{0.05} = 4.75$

(For X_1 and X_2)

ANOVA			
	df	*SS*	*MS*
Regression	2	**29460.02687**	14730.01343
Residual	12	27033.30647	**2252.775539**
Total	14	56493.33333	

(For X_2 only)

ANOVA		
	df	*SS*
Regression	1	**17484.22249**
Residual	13	39009.11085
Total	14	56493.33333

Testing Portions of Model: Example

(continued)

(For X_1 and X_2)

ANOVA			
	df	*SS*	*MS*
Regression	2	**29460.02687**	14730.01343
Residual	12	27033.30647	**2252.775539**
Total	14	56493.33333	

(For X_2 only)

ANOVA		
	df	*SS*
Regression	1	**17484.22249**
Residual	13	39009.11085
Total	14	56493.33333

$$F_{STAT} = \frac{\text{SSR}\ (X_1 \mid X_2)}{\text{MSE(all)}} = \frac{29{,}460.03 - 17{,}484.22}{2252.78} = 5.316$$

Conclusion: Since $F_{STAT} = 5.316 > F_{0.05} = 4.75$ Reject H_0;
Adding X_1 does improve model

Testing Portions of Model: Example

(continued)

H_0: X_2 (advertising) does not improve the model with X_1 (price) included

H_1: X_2 does improve model

$\alpha = .05$, df = 1 and 12

$F_{0.05} = 4.75$

(For X_1 and X_2)

ANOVA			
	df	*SS*	*MS*
Regression	2	**29460.02687**	14730.01343
Residual	12	27033.30647	**2252.775539**
Total	14	56493.33333	

(For X_1 only)

ANOVA		
	df	*SS*
Regression	1	**11100.43803**
Residual	13	45392.8953
Total	14	56493.33333

Testing Portions of Model: Example

(continued)

(For X_1 and X_2)

ANOVA			
	df	*SS*	*MS*
Regression	2	**29460.02687**	14730.01343
Residual	12	27033.30647	**2252.775539**
Total	14	56493.33333	

(For X_1 only)

ANOVA		
	df	*SS*
Regression	1	**11100.43803**
Residual	13	45392.8953
Total	14	56493.33333

$$F_{STAT} = \frac{\text{SSR}(X_2 \mid X_1)}{\text{MSE(all)}} = \frac{29{,}460.03 - 11{,}100.44}{2252.78} = 8.150$$

Conclusion: Since $F_{STAT} = 8.150 > F_{0.05} = 4.75$ Reject H_0;
Adding X_2 does improve model

Simultaneous Contribution of Independent Variables

- Use partial F test for the simultaneous contribution of multiple variables to the model
  - Let m variables be an additional set of variables added simultaneously
  - To test the hypothesis that the set of m variables improves the model:

$$F_{STAT} = \frac{[\text{SSR(all)} - \text{SSR (all except new set of m variables)}]/\,\text{m}}{\text{MSE(all)}}$$

(where F_{STAT} has m and n-k-1 d.f.)

Model Building

- Goal is to develop a model with the best set of independent variables
- Stepwise regression procedure
  - Provide evaluation of alternative models as variables are added and deleted, with partial F tests.
- Best-subset approach
  - Try all combinations and select the best using the highest adjusted r^2 and lowest standard error

Stepwise Regression

- Idea: develop the least squares regression equation in steps, adding one independent variable at a time and evaluating whether existing variables should remain or be removed
- Evaluate significance of newly added variables by partial F tests

Best Subsets Regression

- Idea: estimate all possible regression equations using all possible combinations of independent variables

- Choose the best fit by looking for the highest adjusted r^2 and lowest standard error