MODULE THREE: DETERMINING CAUSE AND MAKING RELIABLE FORECASTS

TOPIC 8: SIMPLE LINEAR REGRESSION

Learning Objectives

At the completion of this topic, you should be able to:

- conduct a simple regression and interpret the meaning of the regression coefficients b_0 and b_1
- use regression analysis to predict the value of a dependent variable based on an independent variable
- assess the adequacy of your estimated model
- evaluate the assumptions of regression analysis
- make inferences about the slope and correlation coefficient
- estimate confidence intervals
- comprehend the pitfalls in regression and ethical issues

+Introduction to Regression Analysis

Recall: Correlation Analysis (Topic 3)

Example: Job satisfaction vs productivity

Regression analysis is used to:

- predict the value of a dependent variable (Y) based on the value of at least one independent variable (X)
- explain the impact of changes in an independent variable on the dependent variable e.g. Productivity (Y) Vs Training (X)

Dependent variable (Y): the variable we wish to predict or explain (response variable)

Independent variable (X): the variable used to explain the dependent variable (explanatory variable)

+12.1 Types of Regression Models

Simple Linear Regression Model

Population Y intercept

Population slope coefficient

Independent variable

Random error term

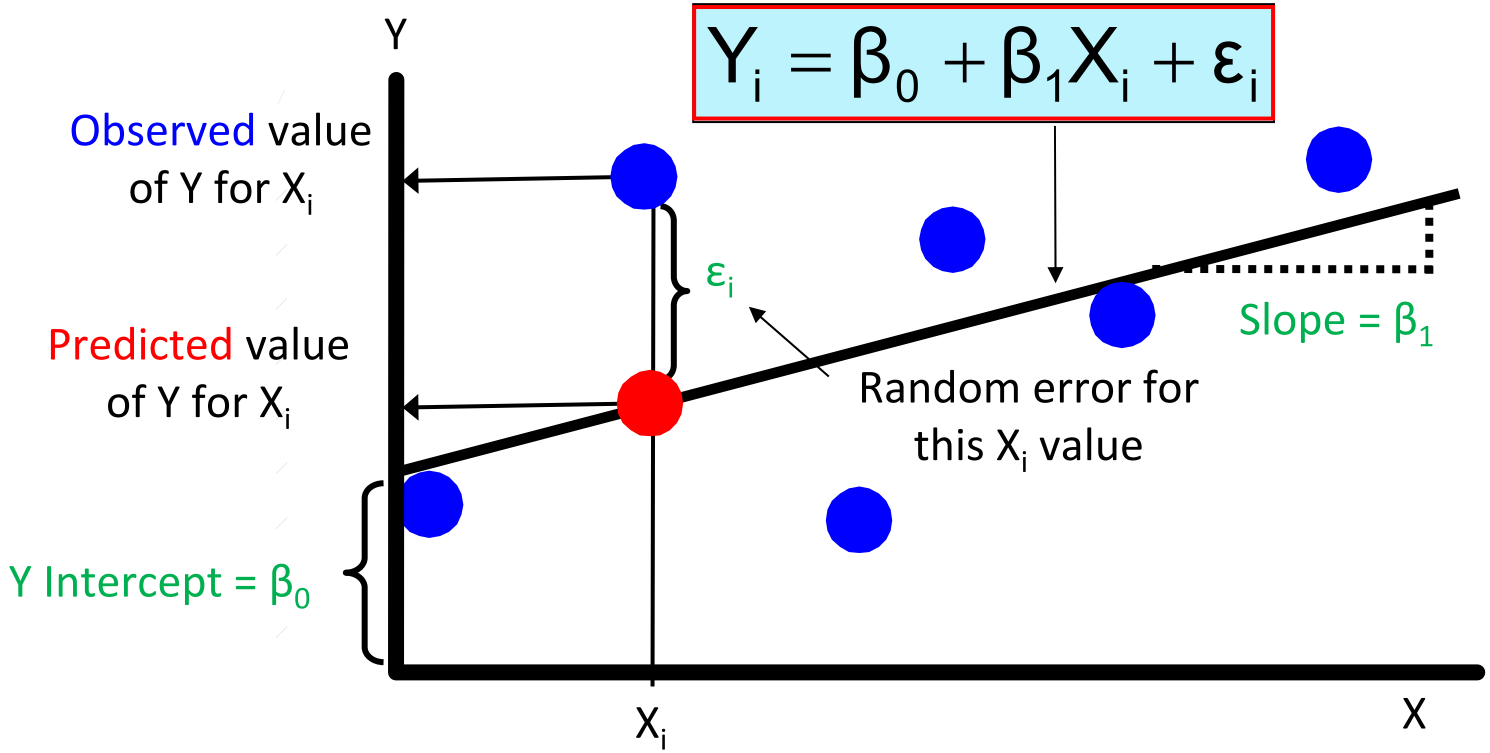
$$Y_i = \underbrace{\boxed{\beta_0} + \boxed{\beta_1} X_i}_{\text{Linear component}} + \underbrace{\boxed{\varepsilon_i}}_{\text{Random error component}}$$

Dependent variable

+12.1 Types of Regression Models

Simple Linear Regression Model (= Regression Equation)

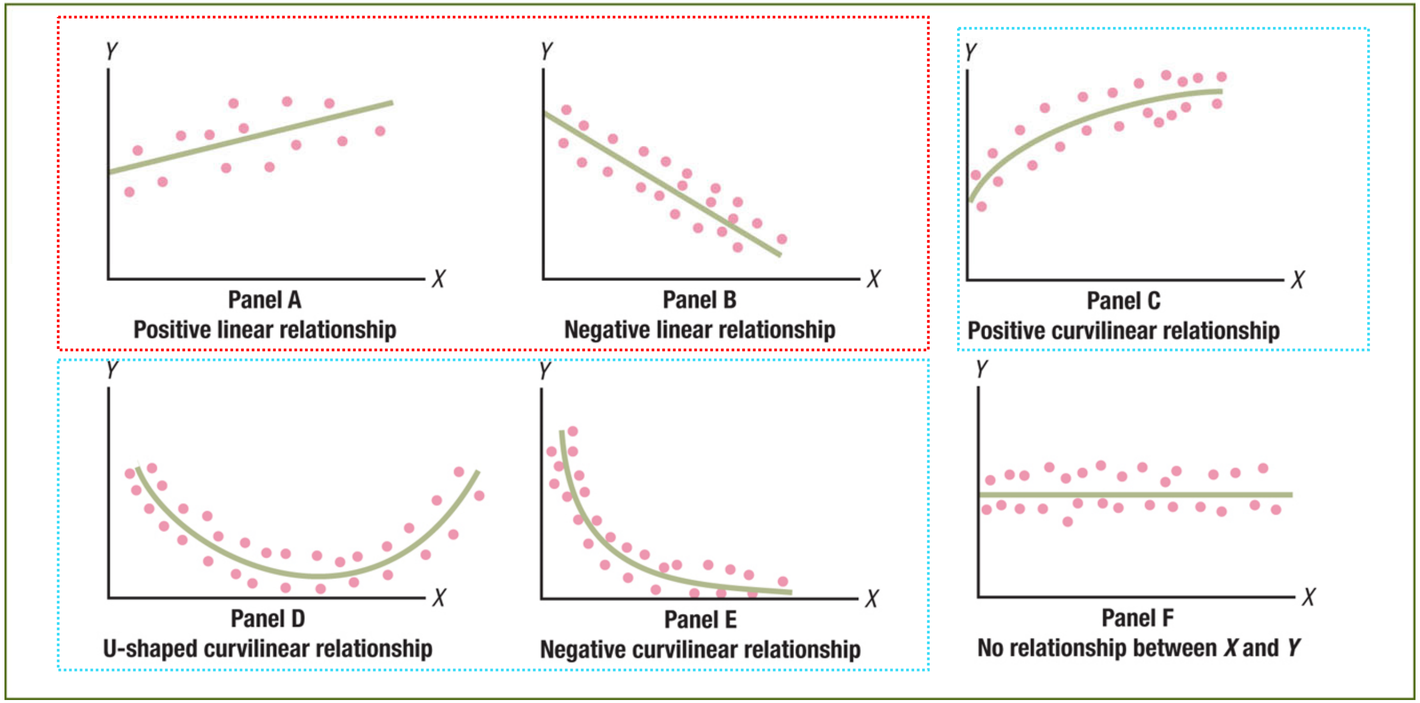
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



+12.1 Types of Regression Models (cont)

No Relationship
Linear relationship (Positive/Negative)
Non-linear Relationship

Figure 12.2
Examples of types of relationships found in scatter diagrams



+Simple Linear Regression

Simple linear regression:

- 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

+12.2 Simple Linear Regression Equation

$$Y_i = \beta_0 + \beta_1 X_i + \boxed{\varepsilon_i}$$

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

$$\boxed{\hat{\ }}Y_i = \boxed{b_0} + \boxed{b_1} X_i$$

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept (β_0)

Estimate of the regression slope (β_1)

Value of X for observation i

+Simple Linear Regression

Example:

A manager of a local computer games store wishes to:

- examine the relationship between weekly sales (Y) and the number of customers making purchases (X) over a 10 week period; and
- use the results of that examination to predict future weekly sales

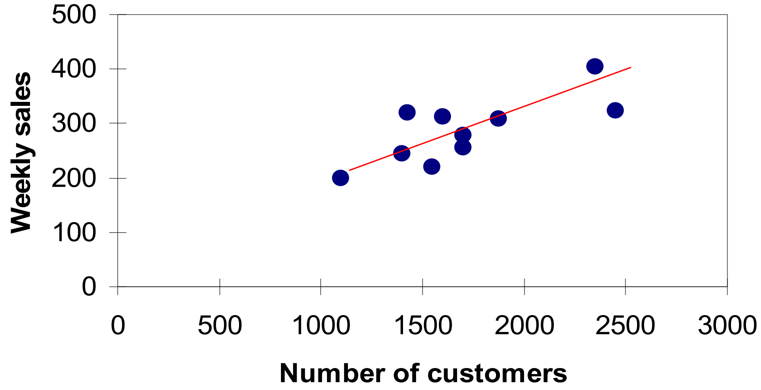
Y - weekly sales

X - number of customers

+Simple Linear Regression (Cont)

Weekly sales in $1,000s (Y)	Number of Customers (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Weekly sales model: scatter plot



+Simple Linear Regression (Cont)

	A	B	C	D	E	F	G
1	*Regression Statistics*						
2	Multiple R	0.762113713			The regression equation is:		
3	R Square	0.580817312					
4	Adjusted R Square	0.528419476	Weekly sales = 98.24833 + 0.10977 (customers)				
5	Standard Error	41.33032365					
6	Observations	10					
7							
8	ANOVA						
9		*df*	*SS*	*MS*	*F*	*Significance F*	
10	Regression	1	18934.93478	18934.93478	11.08475762	0.010394016	
11	Residual	8	13665.56522	1708.195653			
12	Total	9	32600.5				
13							
14		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
15	Intercept	98.24832962	58.03347858	1.692959513	0.128918812	-35.57711186	232.0737711
16	Number of customers	0.109767738	0.032969443	3.329377962	0.010394016	0.033740065	0.18579541

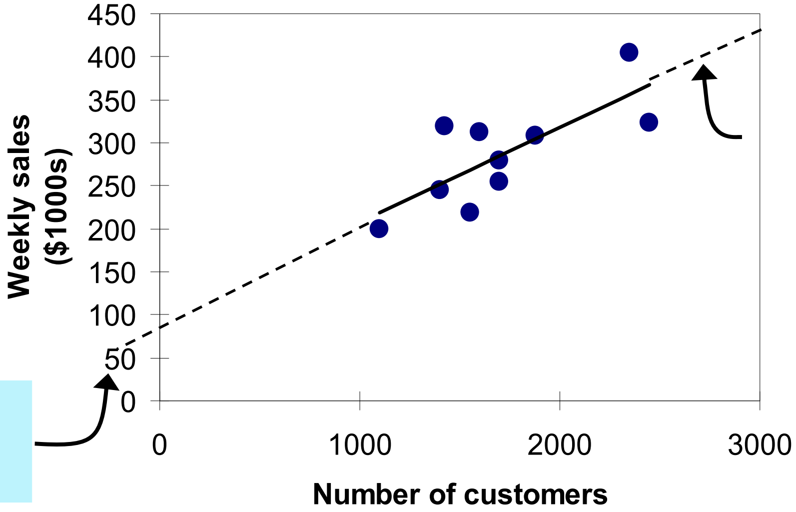
+Simple Linear Regression (Cont)

Weekly sales model: scatter plot and regression line

r = 0.7621

-1 0 1

Slope = 0.10977



Intercept = 98.248

$$\widehat{\text{Weekly sales}} = 98.24833 + 0.10977\,(\text{customers})$$

+Simple Linear Regression (Cont)

$$\widehat{\text{Weekly sales}} = 98.24833 + 0.10977\ (\text{customers})$$

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, for no customers, b_0 = 98.2483 which appears nonsensical. However, the intercept simply indicates that over the sample size selected, the portion of weekly sales not explained by number of customers is $98,248.33. Also note that X=0 is outside the range of observed values

b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X

- Here, b_1 = .10977 tells us that the average value of weekly sales increases by .10977($1,000) = $109.77, on average, for each additional customer

+Simple Linear Regression (Cont)

Predict the weekly sales for the local store for **2,000** customers:

$$\widehat{\text{Weekly sales}} = 98.25 + 0.1098 \boxed{(2000)}$$
$$= 98.25 + 0.1098(2000)$$
$$= 317.85$$

The predicted weekly sales for the local computer games store for 2,000 customers is 317.85 ($1,000s) = $317,850

+The Least-Squares Method

β_0 and β_1 are obtained by finding the values of b_0 and b_1 that **minimise the sum of the squared differences** between actual values (Y) and predicted values ($\hat{Y}$)

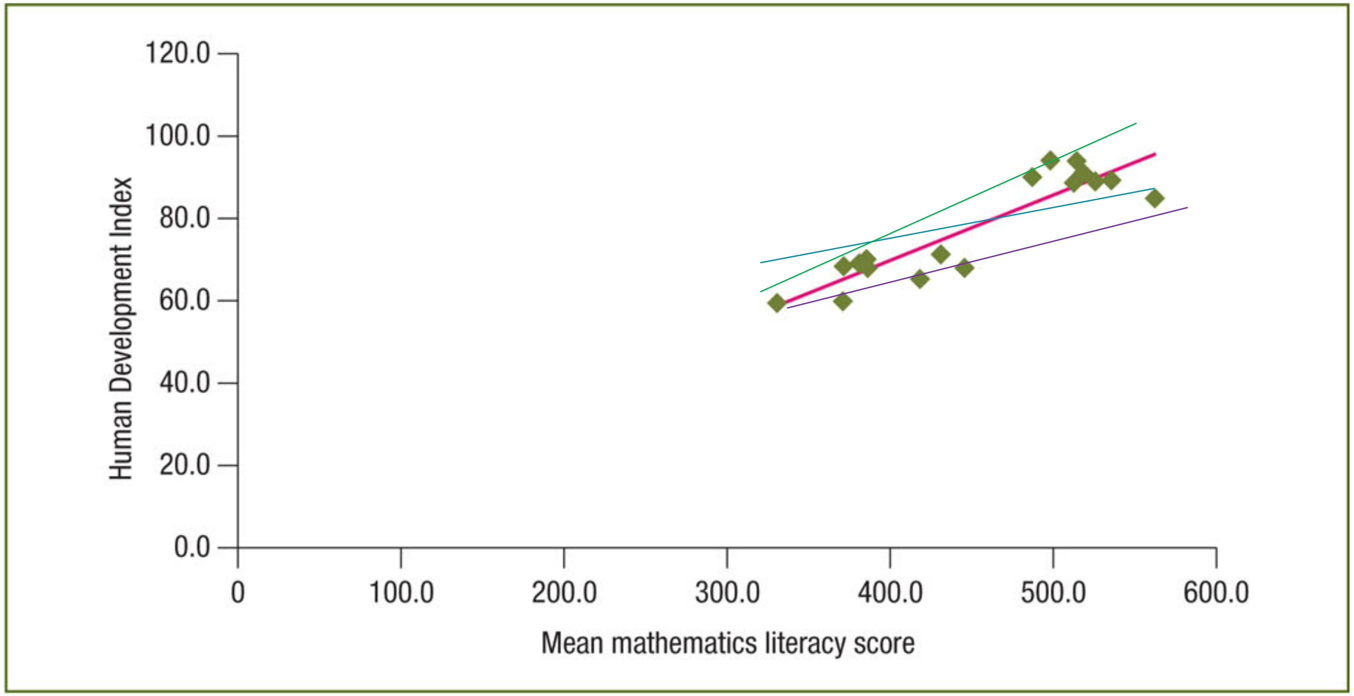
$$\boxed{\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2}$$

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

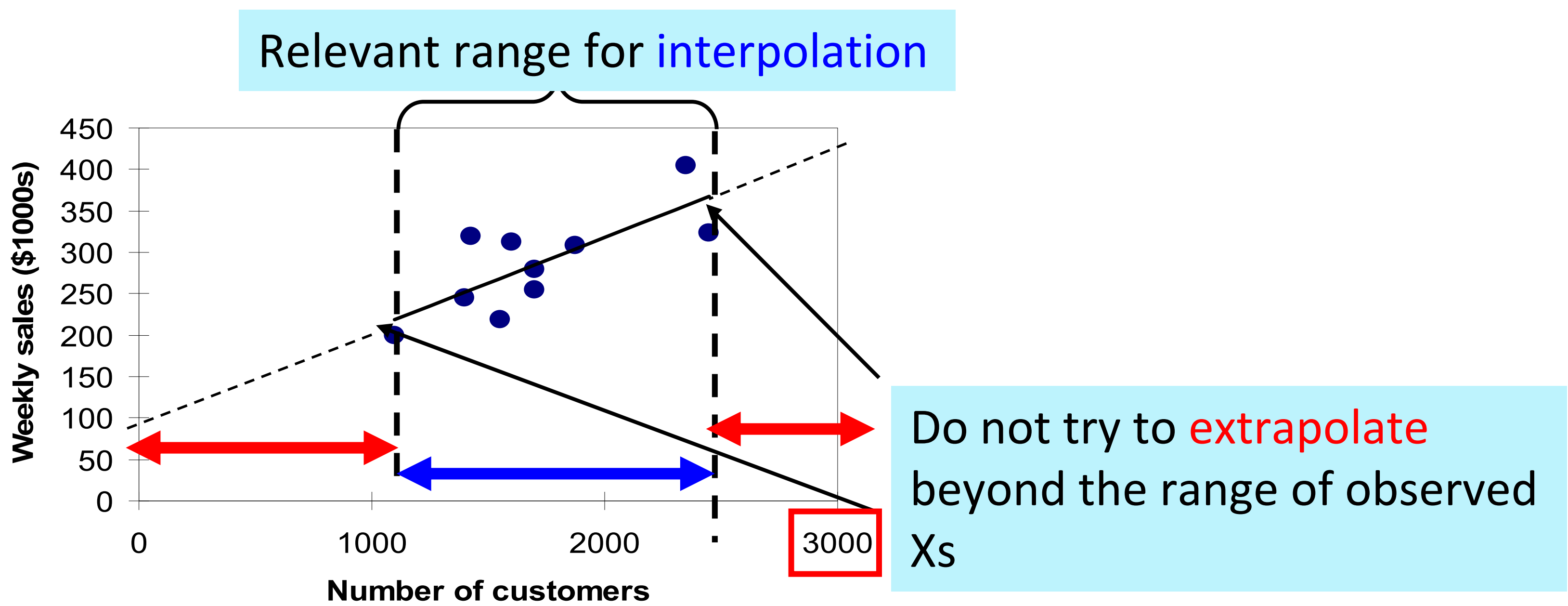
+The Least-Squares Method

Figure 12.5
Microsoft Excel scatter diagram and prediction line for the Human Development Index data

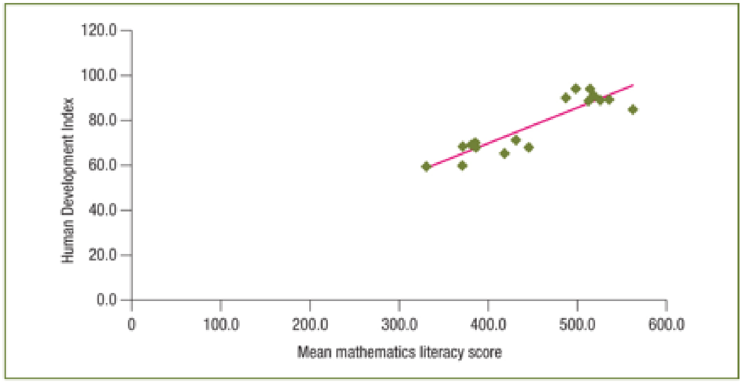


+Predictions in Regression Analysis: Interpolation versus Extrapolation

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



+12.3 Measures of Variati



Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (Y_i - \overline{Y})^2$	$SSR = \sum (\hat{Y}_i - \overline{Y})^2$	$SSE = \sum (Y_i - \hat{Y}_i)^2$
Measures the variation of the Y_i values around their mean Y	Explained variation attributable to the relationship between X and Y	Variation attributable to factors other than the relationship between X and Y

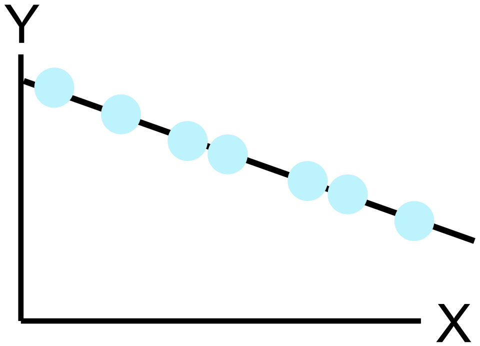
+The Coefficient of Determination, r^2

The Coefficient of Determination (r^2) is equal to the regression sum of squares (i.e. the explained variation) divided by the total sum of squares (i.e. the total variation)

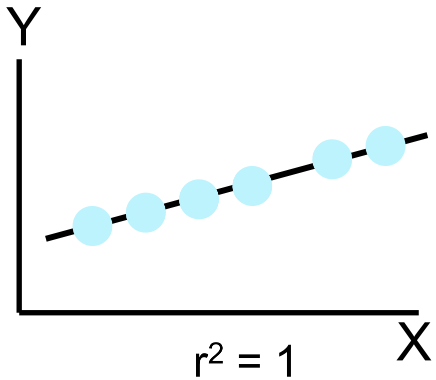
$$r^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{SSR}{SST}$$

It measures the proportion of the variation in Y that is explained by the Independent variable X in the regression model

+The Coefficient of Determination, r^2 (Cont)

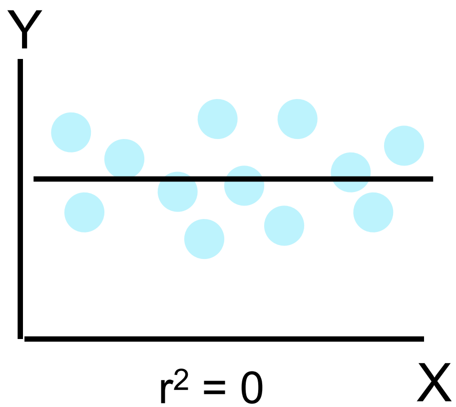


$r^2 = 1$



$r^2 = 1$

- Perfect linear relationship between X and Y
- 100% of the variation in Y is explained by variation in X

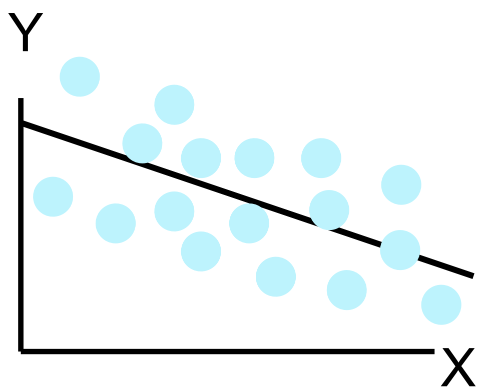


$r^2 = 0$

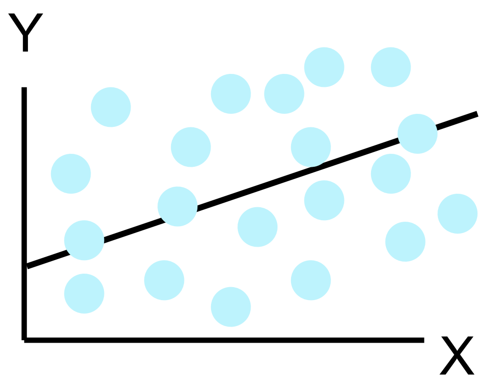
- No linear relationship between X and Y
- The value of Y does not depend on X (none of the variation in Y is explained by variation in X)

+The Coefficient of Determination, r^2 (Cont)

$0 < r^2 < 1$



Moderate/Weaker linear relationships between X and Y:



Some, but not all, of the variation in Y is explained by variation in X

+The Coefficient of Determination, r^2 (Cont)

	A	B	C	D	E	F	G
1	*Regression Statistics*						
2	Multiple R	0.762113713					
3	R Square	0.580817312					
4	Adjusted R Square	0.528419476					
5	Standard Error	41.33032365					
6	Observations	10					
7							
8	ANOVA						
9		*df*	*SS*	*MS*	*F*	*Significance F*	
10	Regression	1	18934.93478	18934.93478	11.08475762	0.010394016	
11	Residual	8	13665.56522	1708.195653			
12	Total	9	32600.5				
13							
14		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
15	Intercept	98.24832962	58.03347858	1.692959513	0.128918812	-35.57711186	232.0737711
16	Number of customers	0.109767738	0.032969443	3.329377962	0.010394016	0.033740065	0.18579541

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

58.08% of the variation in weekly sales is explained by variation in number of customers

About 42% is explained by other factors …

+Standard Error of the Estimate

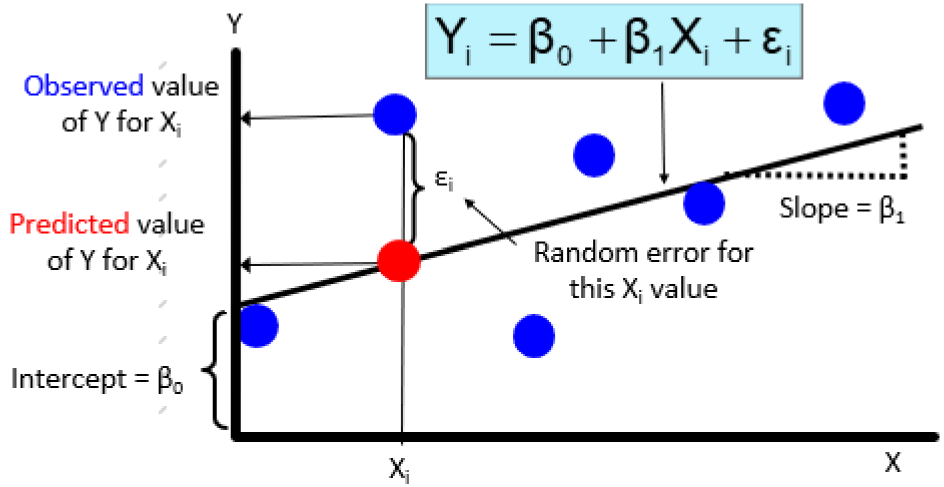
The standard deviation of the variation of observations around the regression line is estimated by:

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

Where:

SSE = error sum of squares

n = sample size



+Standard Error of the Estimate (Cont)

Excel Output:

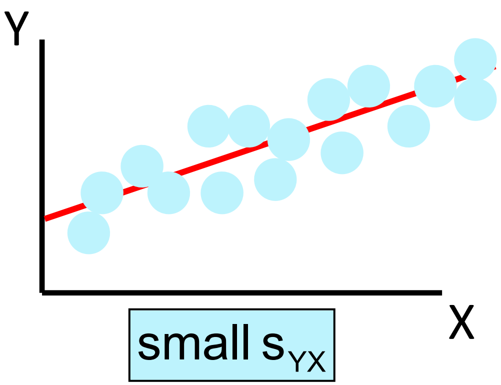
	A	B	C	D	E	F	G
1	*Regression Statistics*						
2	Multiple R	0.762113713					
3	R Square	0.580817312					
4	Adjusted R Square	0.528419476					
5	Standard Error	41.33032365					
6	Observations	10					
7							
8	ANOVA						
9		*df*	*SS*	*MS*	*F*	*Significance F*	
10	Regression	1	18934.93478	18934.93478	11.08475762	0.010394016	
11	Residual	8	13665.56522	1708.195653			
12	Total	9	32600.5				
13							
14		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
15	Intercept	98.24832962	58.03347858	1.692959513	0.128918812	-35.57711186	232.0737711
16	Number of customers	0.109767738	0.032969443	3.329377962	0.010394016	0.033740065	0.18579541

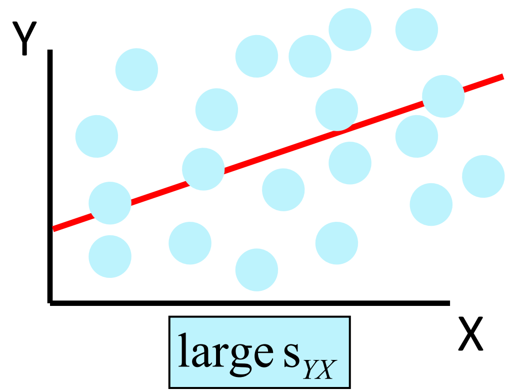
$$S_{YX} = 41.33032$$

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

+Standard Error of the Estimate - 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 weekly sales in the $200K - $300K range

+12.4 Assumptions

Use the acronym LINE:

Linearity

- The underlying relationship between X and Y is linear

Independence of errors

- Error values are statistically independent

Normality of error

- Error values (ε) are normally distributed for any given value of X

Equal variance (homoscedasticity)

- The probability distribution of the errors has constant variance

+12.5 Residual Analysis

The residual for observation i, e_i, is the difference between its observed and predicted value

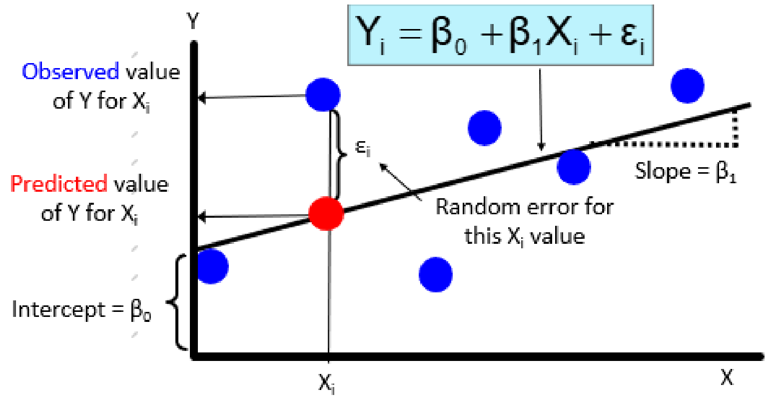
$$e_i = Y_i - \hat{Y}_i$$

Check the assumptions of regression by examining the residuals:

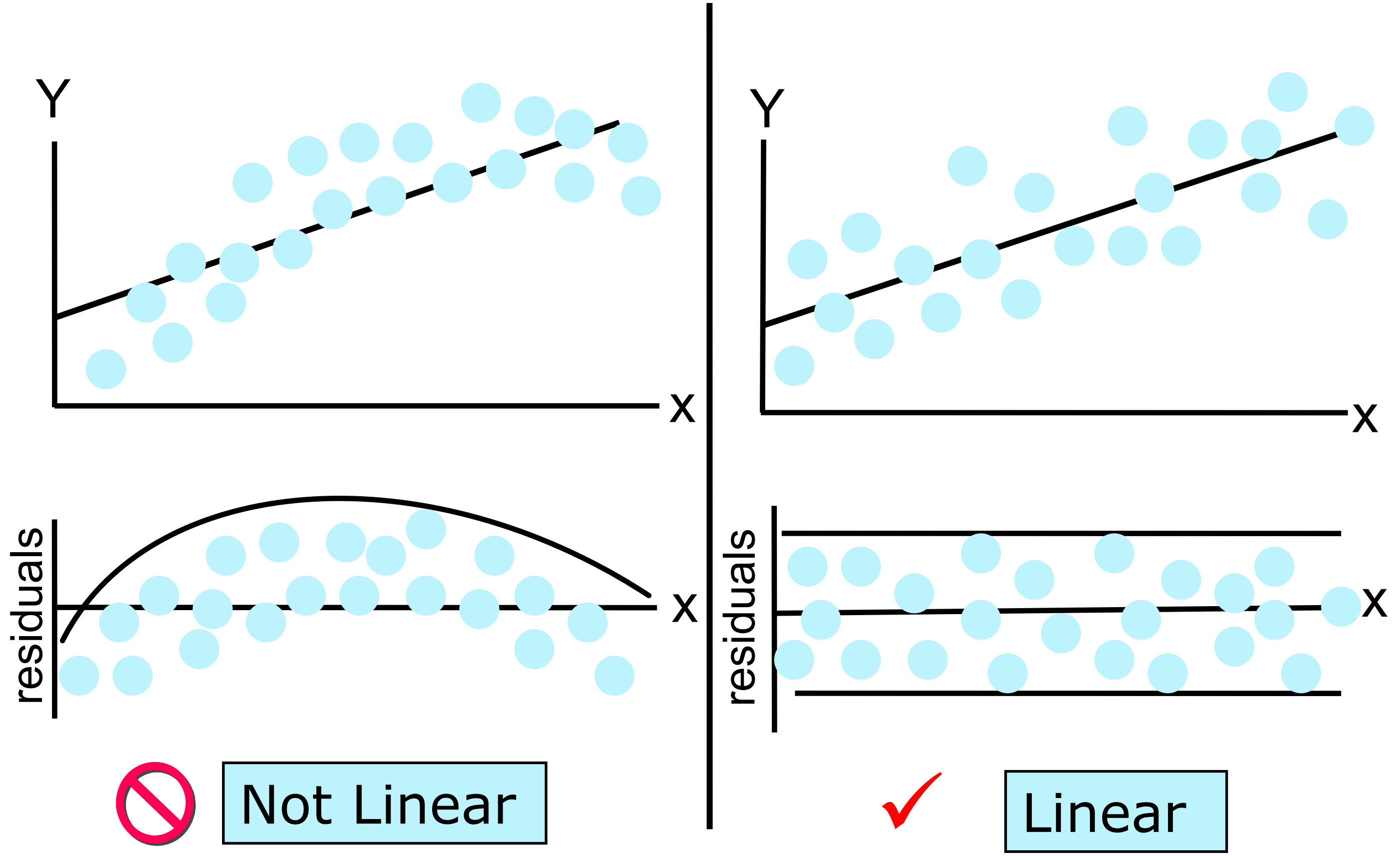
- Examine for linearity assumption
- Evaluate independence assumption
- Evaluate normal distribution assumption
- Examine for constant variance for all levels of X (homoscedasticity)

Graphical Analysis of Residuals

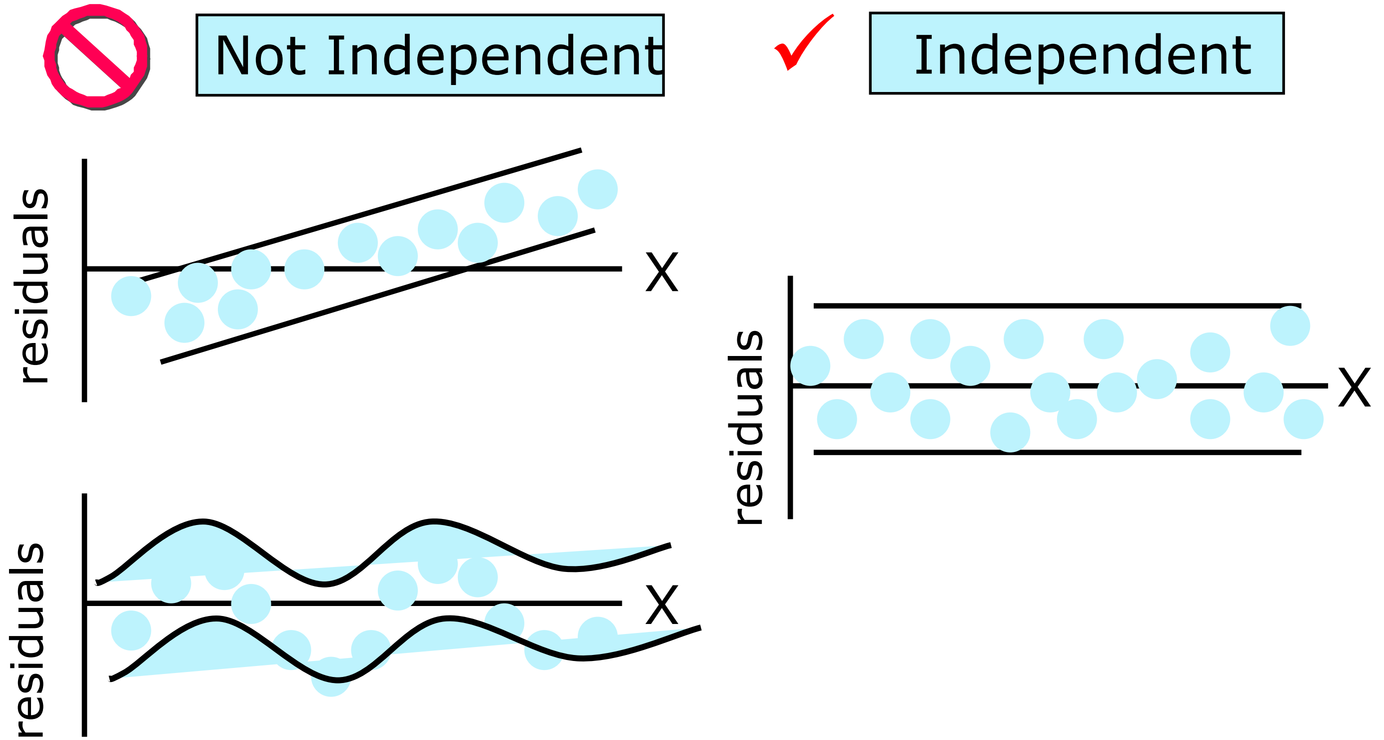
Can plot residuals vs. X



+12.5 Residual Analysis for Linearity

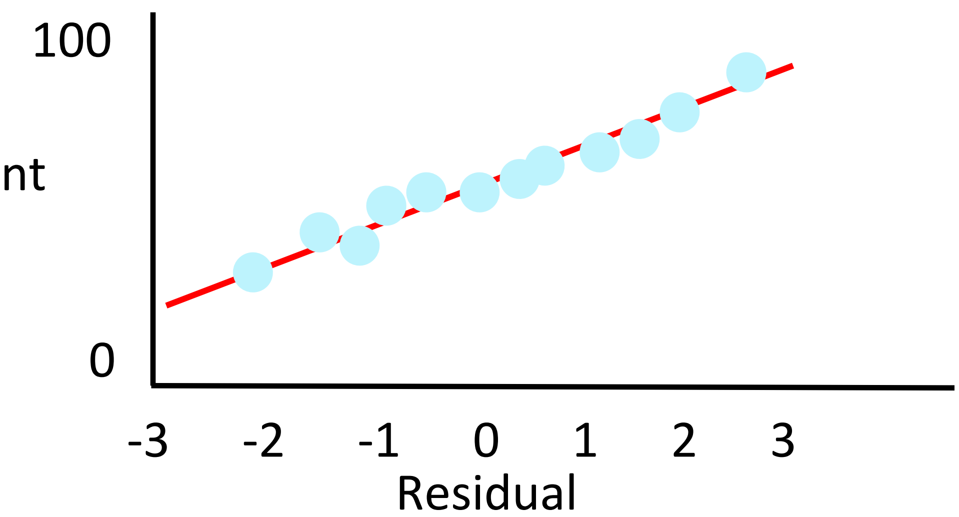


12.5 Residual Analysis for Independence

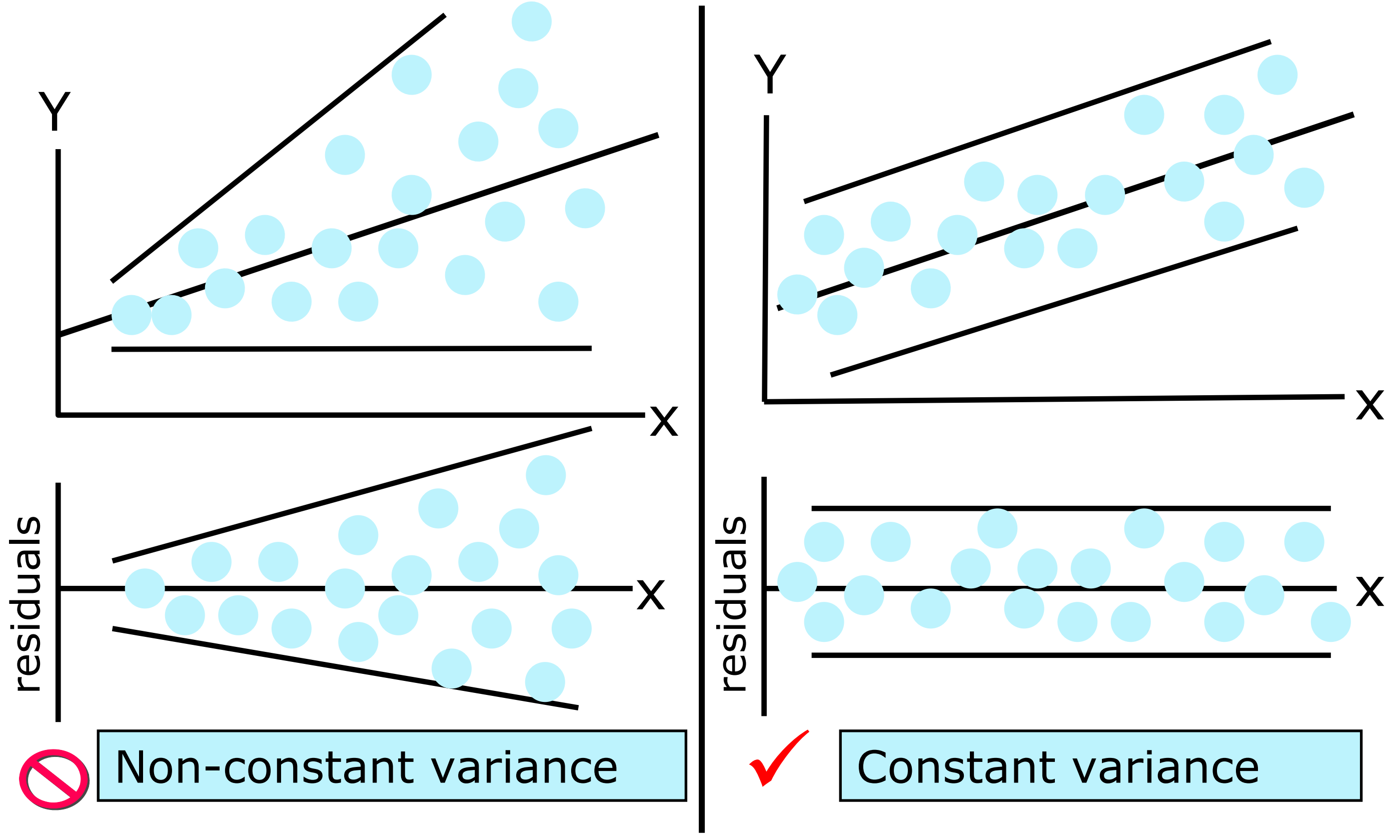


+12.5 Residual Analysis for Normality

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

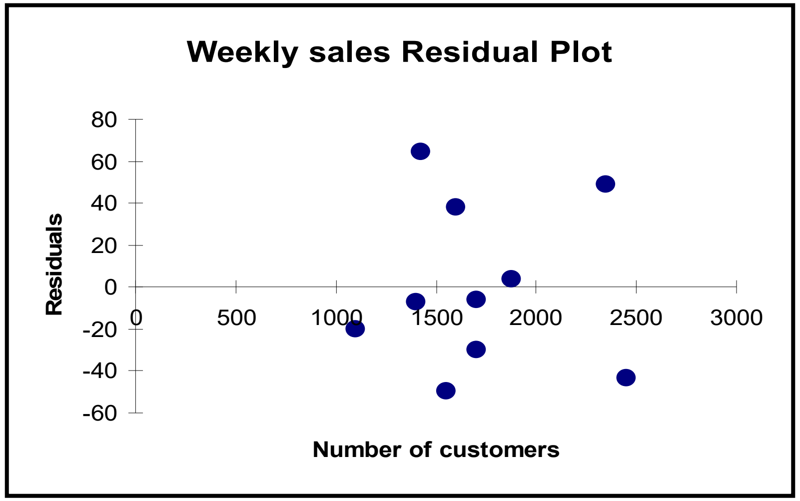


+12.5 Residual Analysis for Equal Variance (Homoscedasticity)



+12.5 Residual Analysis – Excel Residual Output

RESIDUAL OUTPUT		
	Predicted Weekly Sales	*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

LINE ✓

+12.7 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 least squares slope

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

+12.7 Inferences About the Slope – Excel Output

	A	B	C	D	E	F	G
1	*Regression Statistics*						
2	Multiple R	0.762113713					
3	R Square	0.580817312					
4	Adjusted R Square	0.528419476					
5	Standard Error	41.33032365					
6	Observations	10					
7							
8	ANOVA						
9		*df*	*SS*	*MS*	*F*	*Significance F*	
10	Regression	1	18934.93478	18934.93478	11.08475762	0.010394016	
11	Residual	8	13665.56522	1708.195653			
12	Total	9	32600.5				
13							
14		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
15	Intercept	98.24832962	58.03347858	1.692959513	0.128918812	-35.57711186	232.0737711
16	Number of customers	0.109767738	0.032969443	3.329377962	0.010394016	0.033740065	0.18579541

$$S_{b_1} = 0.03297$$

+ *t* Test for the Slope (β_1)

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 with d.f. = n-2

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

Where: b_1 = regression slope coefficient

β_1 = hypothesised slope (population)

S_b = standard error of the slope

+ *t* Test for the Slope (β_1)

$$\widehat{\text{Weekly sales}} = 98.25 + 0.1098\,(\text{customers})$$

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

The slope of this model is 0.1098
Does number of customers affect weekly sales?

b_1 $\quad$ S_{b_1}

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Number of customers	**0.10977**	**0.03297**	**3.32938**	**0.01039**

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

P-value = 0.01039
$\alpha = 0.05$
P-value < α

P-value Approach

If P-value < α, reject H_0
If P-value > α, fail to reject H_0

Decision: Reject H_0

Conclusion: There is sufficient evidence that number of customers affects weekly sales

+*t* Test for the Slope (β_1)

$H_0: \beta_1 = 0$

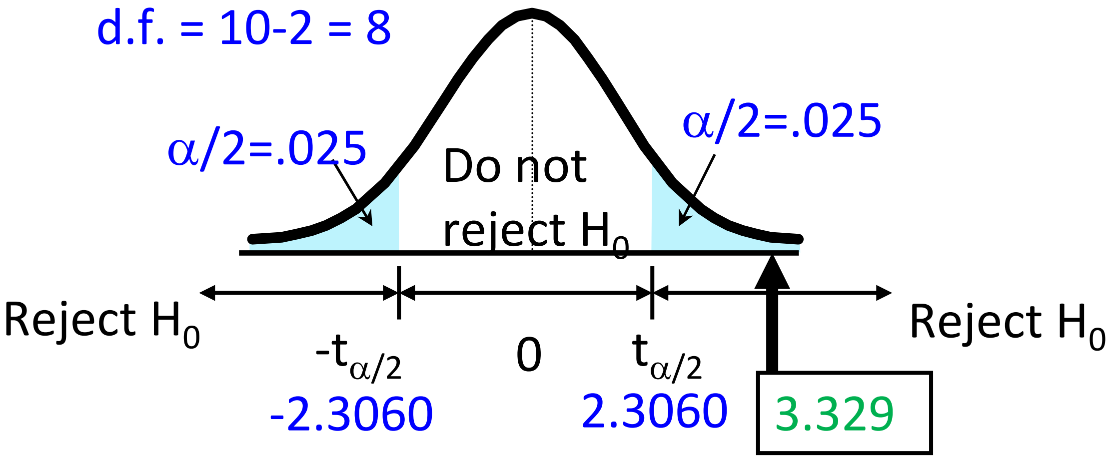
$H_1: \beta_1 \neq 0$

Critical value Approach

If t test statistic < $-t_{\alpha/2}$ or t test statistic > $t_{\alpha/2}$, reject H_0
Otherwise, fail to reject H_0

t Test Statistic: $t = 3.329$

t critical values = +/- 2.3060 (from t tables)



Decision: Reject H_0

Conclusion: There is sufficient evidence that number of customers affects weekly sales

+ *F* Test for Significance

F Test statistic

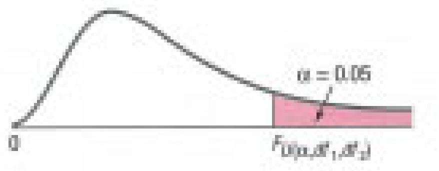
$$F = \frac{MSR}{MSE}$$

where:

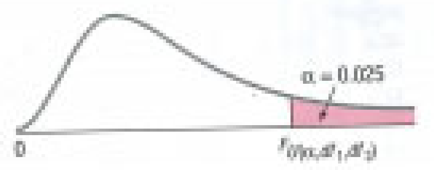
$$MSR = \frac{SSR}{k}$$

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

F follows an F distribution with k numerator and (n – k - 1) denominator degrees of freedom (Table E.5)



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



+*F* Test for Significance – Excel Output

	A	B	C	D	E	F	G
2	Multiple R	0.762113713					
3	R Square	0.580817312					
4	Adjusted R Square	0.528419476					
5	Standard Error	41.33032365					
6	Observations	10					
7			With 1 and 8 degrees of freedom			P-value for the F Test	
8	ANOVA						
9		*df*	*SS*	*MS*	*F*	*Significance F*	
10	Regression	1	18934.93478	18934.93478	11.08475762	0.010394016	
11	Residual	8	13665.56522	1708.195653			
12	Total	9					
13							
14		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
15	Intercept	98.24832962	58.03347858	1.692959513	0.128918812	-35.57711186	232.0737711
16	Number of customers	0.109767738	0.032969443	3.329377962	0.010394016	0.033740065	0.18579541

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

+ *F* Test for Significance - Example

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

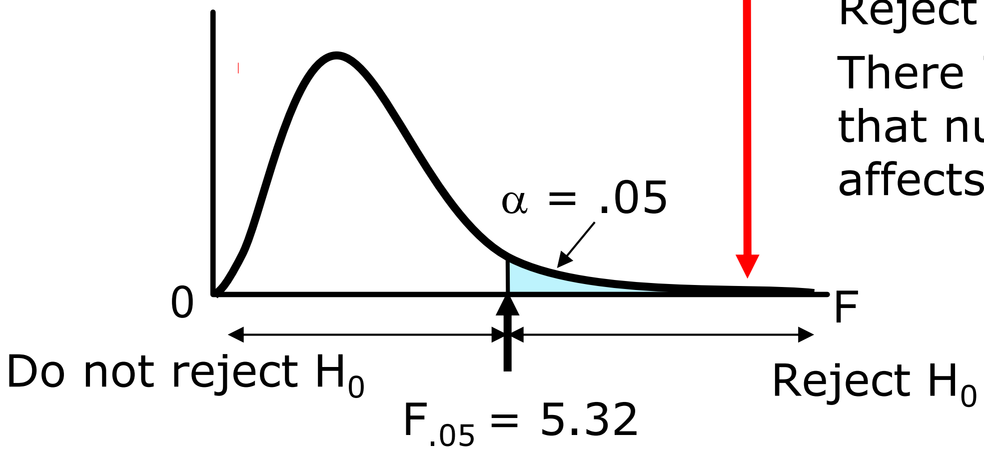
$\alpha = .05$

$df_1 = 1 \qquad df_2 = 8$

Critical Value: $F_\alpha = 5.32$

F Test Statistic:

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



Conclusion:

Reject H_0 at $\alpha = 0.05$

There is sufficient evidence that number of customers affects weekly sales

+Confidence Interval Estimation for the Slope (β_1)

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

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

d.f. = n - 2 Excel Printout for Weekly sales:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Customers	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

At 95% level of confidence, the confidence interval for the slope is (0.03374, 0.18580); i.e. we are 95% confident that the average impact on weekly sales is between $33.74 and $185.80 per customer

This 95% confidence interval does not include 0.
Conclusion: There is a significant relationship between weekly sales and number of customers at the .05 level of significance

+ *t* Test for the Correlation Coefficient

(-1< r < 1) r is an estimate of the true correlation coefficient ρ

Hypotheses

H_0: $\rho = 0$	no association (correlation) between X and Y
H_1: $\rho \neq 0$	statistically significant association (correlation) exists

t Test statistic

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

where

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

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

(with n – 2 degrees of freedom)

+ *t* Test for the Correlation Coefficient (r) – Example

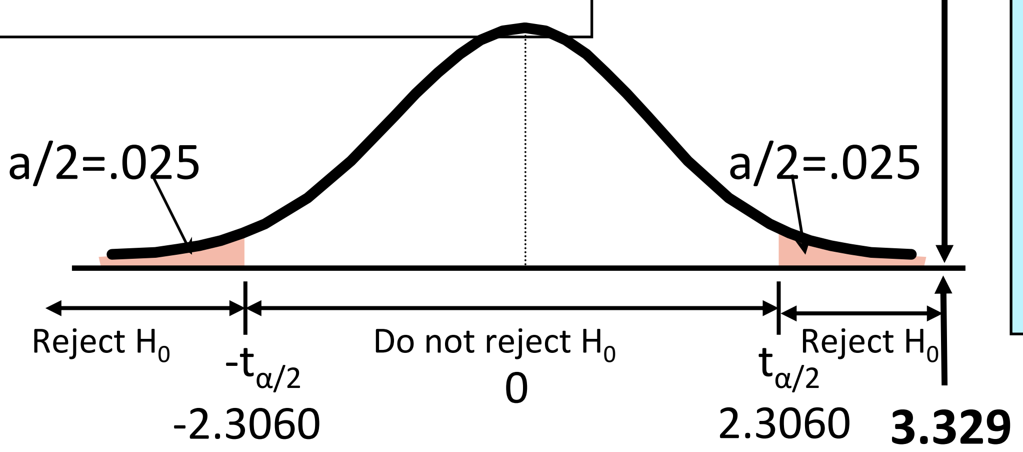
Is there evidence of a significant linear relationship between weekly sales and number of customers at the 5% level of significance?

H_0: ρ = 0 (No correlation)

H_1: ρ ≠ 0 (correlation exists)

α =.05 , df = 10 - 2 = 8

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

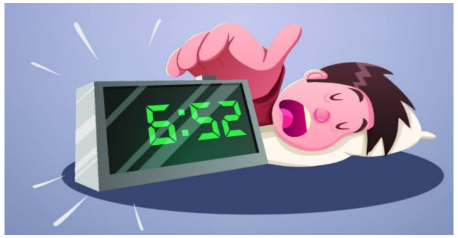


Decision: Reject H_o

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

+ 12.9 Pitfalls in Regression and Ethical Issues

- 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 (e.g. Height Vs Age)
- Concluding that a significant relationship in observational study is due to a cause and effect relationship



WORD OF WARNING!

Correlation Isn't Causation!

As ice cream sales increase, the rate of drowning deaths increases sharply.

Therefore, ice cream consumption causes drowning?!

This conclusion is wrong!

"a strong association is not a proof of causation"

EXERCISE: SALES VS ADVERTISING

A company has collected data over the last 10 years relating to its annual expenditure on advertising as well as its total sales (all figures scaled for inflation).

Sales ($m)	30.2	37.3	29.9	35.2	35	33.5	36	31.1	34.1	36.9
Advertising ($m)	0.5	1.2	0.6	1.1	1.8	1.4	1	0.7	0.7	1.3

Develop a regression model (Using Excel) and answer the following questions:

- How well does the model predict sales?
- Interpret b0 and b1.
- At the 0.05 level of significance, is there a significant linear relationship between the sales and the expenditure on advertising?
- What would you estimate sales to be when $1m is spent on advertising?

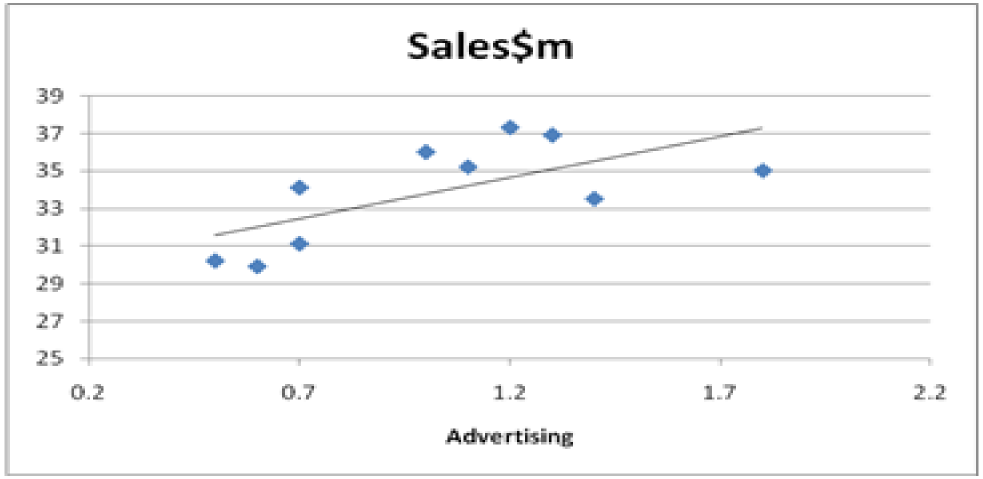
Sales^ ($Million) = 29.414 +4.375*Advertising
= 29.414 +4.375*1
= $33.789 m

Sales vs Advertising

Correlation coefficients

	Advertising$m	*Sales$m*
Advertising$m	1	
Sales$m	0.666	1

Scatter diagram



Regression output

Regression Statistics	
Multiple R	0.666
R Square	0.444
Adj R Square	0.374
Standard Error	2.136
Observations	10

H_o: There is no linear relationship between sales and Advertising

H_1: There is a linear relationship between sales and Advertising

ANOVA

	df	*SS*	*MS*	*F*	*Sig F*
Regression	1	29.110	29.110	6.383	0.035
Residual	8	36.486	4.561		
Total	9	65.596			

	Coefficients	*St Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	29.414	1.907	15.423	0.000	25.016	33.812
Advertising$m	4.375	1.732	2.526	0.035	0.382	8.368

Sales^ ($Million) = 29.414 +4.375*Advertising