Inferential Statistics and Probability
a Holistic Approach

Chapter 13
Correlation and Linear Regression

1

Mathematical Model

- You have a small business producing custom t-shirts.
- Without marketing, your business has revenue (sales) of $1000 per week.
- Every dollar you spend marketing will increase revenue by 2 dollars.
- Let variable X represent amount spent on marketing and let variable Y represent revenue per week.
- Write a mathematical model that relates X to Y

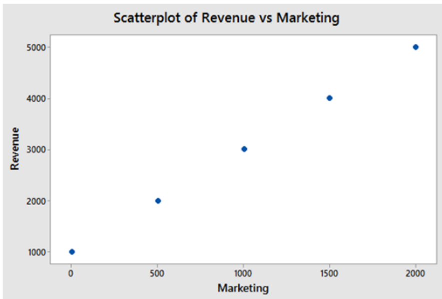
2

Mathematical Model - Table

X=marketing	Y=revenue
$0	$1000
$500	$2000
$1000	$3000
$1500	$4000
$2000	$5000

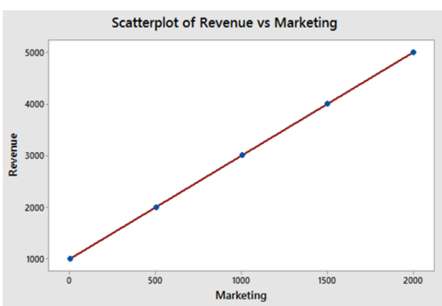
3

Mathematical Model - Scatterplot



4

Mathematical Model - Linear



5

Mathematical Linear Model

Linear Model	Example
$Y = \beta_0 + \beta_1 X$	$Y = 1000 + 2X$
Y: *Dependent Variable*	Y: *Revenue*
X: *Independent Variable*	X: *Marketing*
β_0: *Y-intercept*	β_0: \$1000
β_1: *Slope*	β_1: \$2 *per* \$1 *marketing*

6

Statistical Model

- You have a small business producing custom t-shirts.
- Without marketing, your business has revenue (sales) of $1000 per week.
- Every dollar you spend marketing will increase revenue by an expected value of 2 dollars.
- Let variable X represent amount spent on marketing and let variable Y represent revenue per week.
- Let ε represent the difference between Expected Revenue and Actual Revenue (Residual Error)
- Write a statistical model that relates X to Y

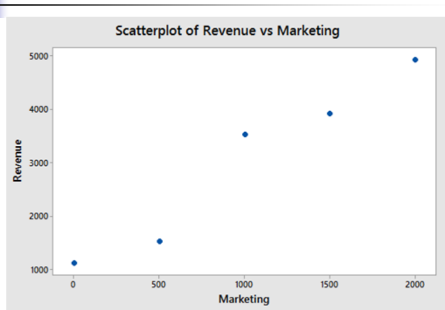
7

Statistical Model - Table

X=Marketing	Expected Revenue	Y=Actual Revenue	ε=Residual Error
$0	$1000	$1100	+$100
$500	$2000	$1500	-$500
$1000	$3000	$3500	+$500
$1500	$4000	$3900	-$100
$2000	$5000	$4900	-$100

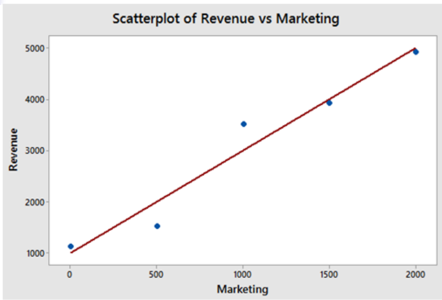
8

Statistical Model - Scatterplot



9

Statistical Model - Linear



10

Statistical Linear Model

Regression Model

$Y = \beta_0 + \beta_1 X + \varepsilon$

Y: Dependent Variable

X: Independent Variable

β_0: Y-intercept

β_1: Slope

ε: $Normal(0, \sigma)$

Example

$Y = 1000 + 2X + \varepsilon$

Y: Revenue

X: Marketing

β_0: \$1000

β_1: \$2 per \$1 marketing

11

Regression Analysis

- Purpose: to determine the regression equation; it is used to predict the value of the dependent response variable (Y) based on the independent explanatory variable (X).
- Procedure:
  - select a sample from the population
  - list the paired data for each observation
  - draw a scatter diagram to give a visual portrayal of the relationship
  - determine the regression equation.

12

Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
Y : *Dependent Variable*
X : *Independen t Variable*
β_0 : *Y – intercept*
β_1 : *Slope*
ε : *Normal* $(0, \sigma)$

13

Estimation of Population Parameters

- From sample data, find statistics that will estimate the 3 population parameters
- Slope parameter
  - b_1 will be an estimator for β_1
- Y-intercept parameter
  - b_o 1 will be an estimator for β_o
- Standard deviation
  - s_e will be an estimator for σ

14

Regression Analysis

- the regression equation: $\hat{Y} = b_0 + b_1 X$, where:
- $\hat{Y}$ is the average predicted value of *Y* for any *X*.
- b_0 is the Y-intercept, or the estimated *Y* value when *X=0*
- b_1 is the slope of the line, or the average change in $\hat{Y}$ for each change of one unit in *X*
- the least squares principle is used to obtain b_1 *and* b_0

$$SSX = \Sigma X^2 - \tfrac{1}{n}(\Sigma X)^2$$
$$SSY = \Sigma Y^2 - \tfrac{1}{n}(\Sigma Y)^2$$
$$SSXY = \Sigma XY - \tfrac{1}{n}(\Sigma X \cdot \Sigma y)$$

$$b_1 = \frac{SSXY}{SSX}$$
$$b_0 = \overline{Y} - b_1 \overline{X}$$

15

Assumptions Underlying Linear Regression

- For each value of X, there is a group of Y values, and these Y values are *normally distributed*.
- The *means* of these normal distributions of Y values all lie on the straight line of regression.
- The *standard deviations* of these normal distributions are equal.
- The Y values are statistically independent. This means that in the selection of a sample, the Y values chosen for a particular X value do not depend on the Y values for any other X values.

16

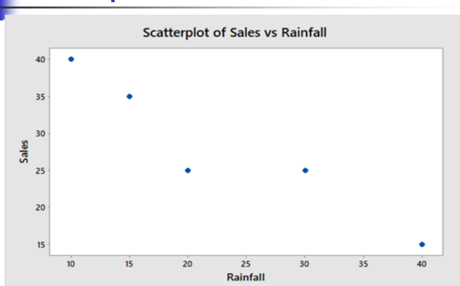
Example

- X = Average Annual Rainfall (Inches)
- Y = Average Sale of Sunglasses/1000
  - Make a Scatterplot
  - Find the least square line

X	10	15	20	30	40
Y	40	35	25	25	15

17

Example *continued*



18

Example *continued*

	X	Y	X^2	Y^2	XY
	10	40	100	1600	400
	15	35	225	1225	525
	20	25	400	625	500
	30	25	900	625	750
	40	15	1600	225	600
Σ	115	140	3225	4300	2775

19

Example *continued*

- Find the least square line
  - SSX = 580
  - SSY = 380
  - SSXY = -445
  - b_1 = -.767
  - b_0 = 45.647
  - $\hat{Y}$ = 45.647 - .767X

20

The Standard Error of Estimate

- The standard error of estimate measures the scatter, or dispersion, of the observed values around the line of regression
- The formulas that are used to compute the standard error:

$$SSR = b_1 \cdot SSXY$$

$$SSE = \sum (Y - \hat{Y})^2 = SSY - SSR$$

$$MSE = SSE / (n-2)$$

$$s_e = \sqrt{MSE}$$

21

Example *continued*

- Find SSE and the standard error:
  - SSR = 341.422
  - SSE = 38.578
  - MSE = 12.859
  - s_e = 3.586

x	y	ŷ	y - ŷ	$(y - ŷ)^2$
10	40	37.97	2.03	4.104
15	35	34.14	0.86	0.743
20	25	30.30	-5.30	28.108
30	25	22.63	2.37	5.620
40	15	14.96	0.04	0.002
			Total	38.578

22

Correlation Analysis

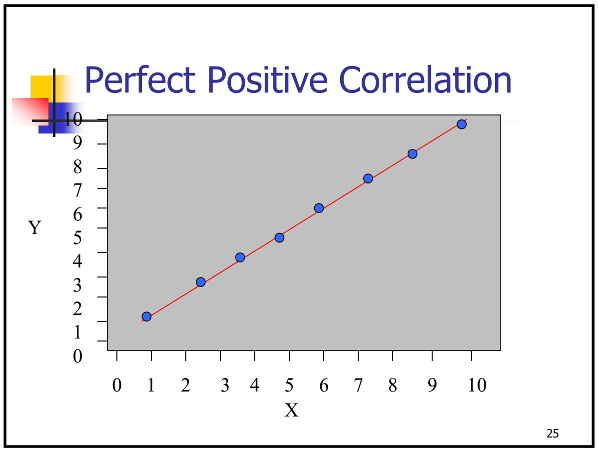
- Correlation Analysis: A group of statistical techniques used to measure the strength of the relationship (correlation) between two variables.
- Scatter Diagram: A chart that portrays the relationship between the two variables of interest.
- Dependent Variable: The variable that is being predicted or estimated. "Effect"
- Independent Variable: The variable that provides the basis for estimation. It is the predictor variable. "Cause?" (Maybe!)

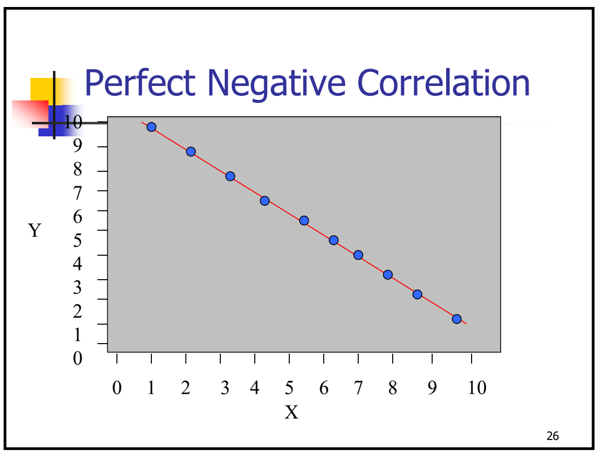
23

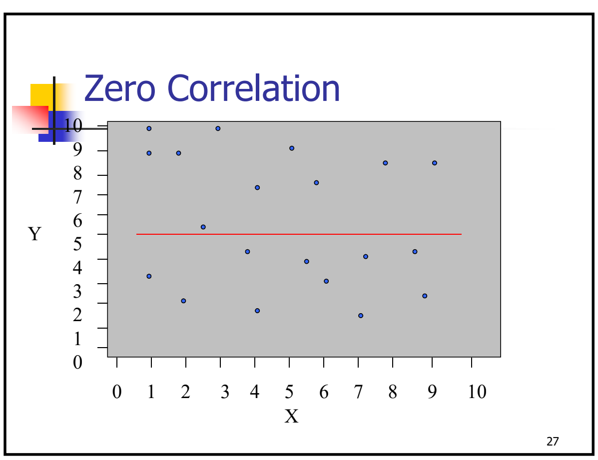
The Coefficient of Correlation, r

- The Coefficient of Correlation (r) is a measure of the **strength** of the relationship between two variables.
  - It requires interval or ratio-scaled data (variables).
  - It can range from -1.00 to 1.00.
  - Values of -1.00 or 1.00 indicate perfect and strong correlation.
  - Values close to 0.0 indicate weak correlation.
  - Negative values indicate an inverse relationship and positive values indicate a direct relationship.

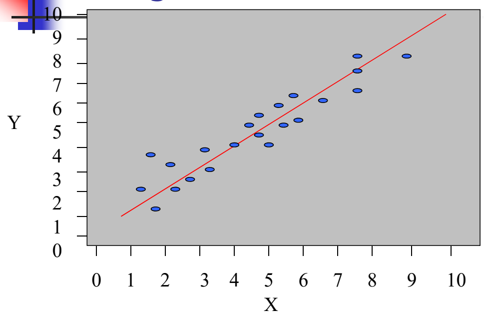
24


Perfect Positive Correlation
Y
10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10
X
25


Perfect Negative Correlation
Y
10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10
X
26

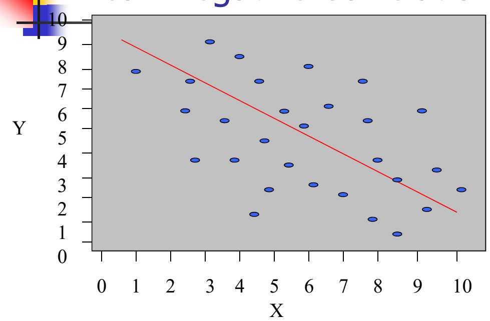

Zero Correlation
Y
10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10
X
27

Strong Positive Correlation



28

Weak Negative Correlation



29

Causation

- Correlation does not necessarily imply causation.
- There are 4 possibilities if X and Y are correlated:
  1. X causes Y
  2. Y causes X
  3. X and Y are caused by something else.
  4. Confounding - The effect of X and Y are hopelessly mixed up with other variables.

30

Causation - Examples

- City with more police per capita have more crime per capita.
- As Ice cream sales go up, shark attacks go up.
- People with a cold who take a cough medicine feel better after some rest.

31

r^2: Coefficient of Determination

- r^2 is the proportion of the total variation in the dependent variable Y that is explained or accounted for by the variation in the independent variable X.
- The coefficient of determination is the square of the coefficient of correlation, and ranges from 0 to 1.

32

Formulas for r and r^2

$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}} \qquad r^2 = \frac{SSR}{SSY}$$

$$SSX = \Sigma X^2 - \tfrac{1}{n}(\Sigma X)^2$$

$$SSY = \Sigma Y^2 - \tfrac{1}{n}(\Sigma Y)^2$$

$$SSXY = \Sigma XY - \tfrac{1}{n}(\Sigma X \cdot \Sigma Y)$$

$$SSR = SSY - \left(SSXY^2 / SSX\right)$$

33

Example

- X = Average Annual Rainfall (Inches)
- Y = Average Sale of Sunglasses/1000

X	10	15	20	30	40
Y	40	35	25	25	15

34

Example *continued*

- Make a Scatter Diagram
- Find r and r^2

35

Example *continued*

X	Y	X^2	Y^2	XY
10	40	100	1600	400
15	35	225	1225	525
20	25	400	625	500
30	25	900	625	750
40	15	1600	225	600
115	140	3225	4300	2775

- SSX = $3225 - 115^2/5$ = 580
- SSY = $4300 - 140^2/5$ = 380
- SSXY= 2775 - (115)(140)/5 = -445

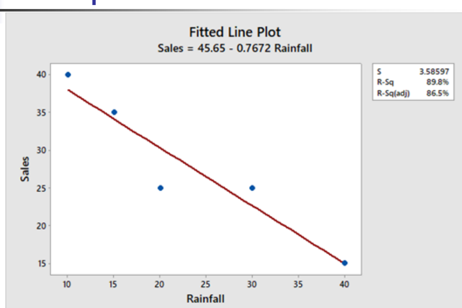
36

Example *continued*

- r = -445/sqrt(580 x 330) = -.9479
  - Strong negative correlation
- r^2 = .8985
  - About 89.85% of the variability of sales is explained by rainfall.

37

Example *continued*



38

Characteristics of F-Distribution

- There is a "family" of *F* Distributions.
- Each member of the family is determined by two parameters: the numerator degrees of freedom and the denominator degrees of freedom.
- *F* cannot be negative, and it is a continuous distribution.
- The *F* distribution is positively skewed.
- Its values range from 0 *to* ∞. As $F \rightarrow \infty$ the curve approaches the X-axis.

39

Hypothesis Testing in Simple Linear Regression

- The following Tests are equivalent:
  - H_0: X and Y are uncorrelated
  - H_a: X and Y are correlated
  - H_0: $\beta_1 = 0$
  - H_a: $\beta_1 \neq 0$
- Both can be tested using ANOVA

40

ANOVA Table for Simple Linear Regression

Source	SS	df	MS	F
Regression	SSR	1	SSR/dfR	MSR/MSE
Error/Residual	SSE	n-2	SSE/dfE	
TOTAL	SSY	n-1		

41

Example *continued*

- Test the Hypothesis $H_o: \beta_1 = 0$, α=5%

Source	SS	df	MS	F	p-value
Regression	341.422	1	341.422	26.551	0.0142
Error	38.578	3	12.859		
TOTAL	380.000	4			

- Reject Ho p-value < α

42

Confidence Interval

- The confidence interval for the mean value of Y for a given value of X is given by:

$$\hat{Y} \pm t \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{SSX}}$$

- Degrees of freedom for t =n-2

43

Prediction Interval

- The prediction interval for an individual value of Y for a given value of X is given by:

$$\hat{Y} \pm t \cdot s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{SSX}}$$

- Degrees of freedom for t =n-2

44

Example *continued*

- Find a 95% Confidence Interval for Sales of Sunglasses when rainfall = 25 inches.
- Find a 95% Prediction Interval for Sales of Sunglasses when rainfall = 25 inches.

45

Example – Minitab output

- Sales = 45.65 - 0.767 Rainfall
- Variable Setting
- Rainfall   25
- Fit   SE Fit   95% CI   95% PI
- 26.4655 1.63111 (21.2746, 31.6564) (13.9282, 39.0028)

46

Example *continued*

- 95% Confidence Interval

$$22.63 \pm 6.60$$

- 95% Confidence Interval

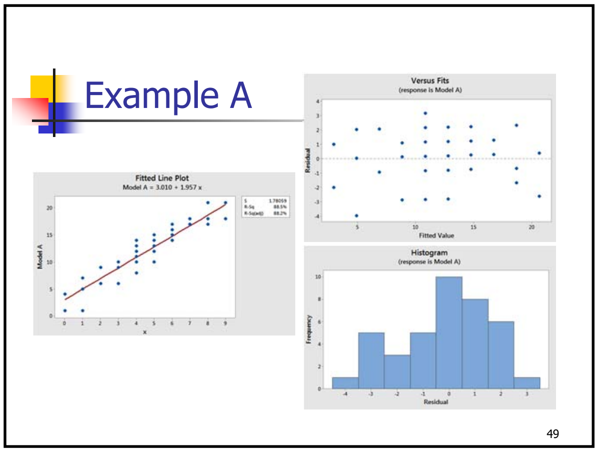
$$22.63 \pm 13.18$$

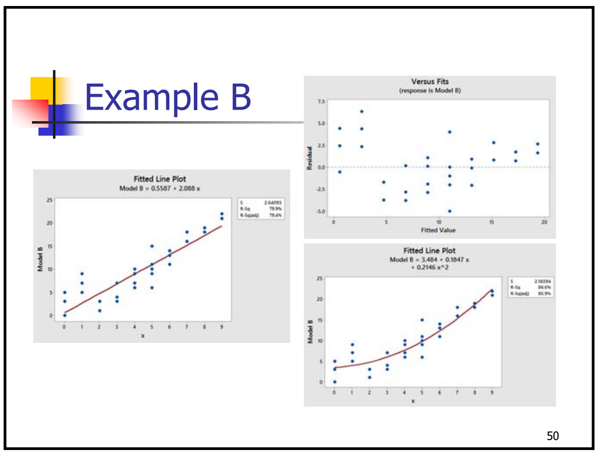
47

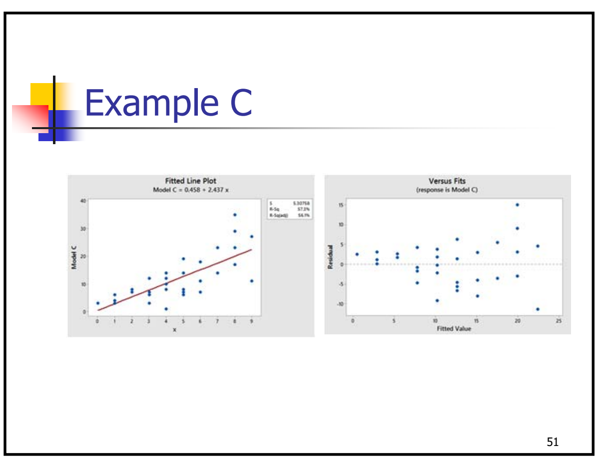
Residual Analysis

- Residuals should
  - have a normal distribution with constant σ
  - be mutually independent
  - not follow a pattern
  - be checked for outliers
    - with respect the line
    - with respect to X

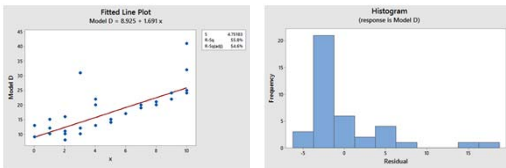
48


Example A
Versus Fits
(response is Model A)
Residual
Fitted Value
Fitted Line Plot
Model A = 3.010 + 1.957 x
S 1.78059
R-Sq 88.5%
R-Sq(adj) 88.2%
Model A
x
Histogram
(response is Model A)
Frequency
Residual
49


Example B
Versus Fits
(response is Model B)
Residual
Fitted Value
Fitted Line Plot
Model B = 0.5587 + 2.088 x
S 2.64093
R-Sq 79.9%
R-Sq(adj) 79.4%
Model B
x
Fitted Line Plot
Model B = 3.484 + 0.1847 x
+ 0.2146 x^2
S 2.18394
R-Sq 86.6%
R-Sq(adj) 85.9%
Model B
x
50

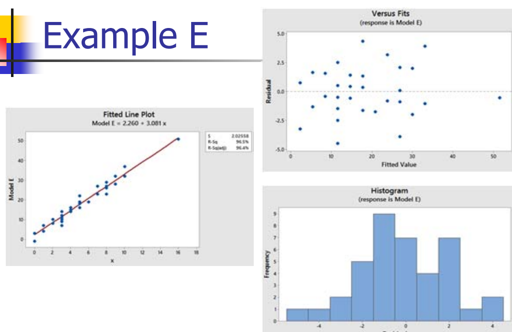

Example C
Fitted Line Plot
Model C = 0.458 + 2.437 x
S 5.30758
R-Sq 57.3%
R-Sq(adj) 56.1%
Model C
x
Versus Fits
(response is Model C)
Residual
Fitted Value
51

Example D



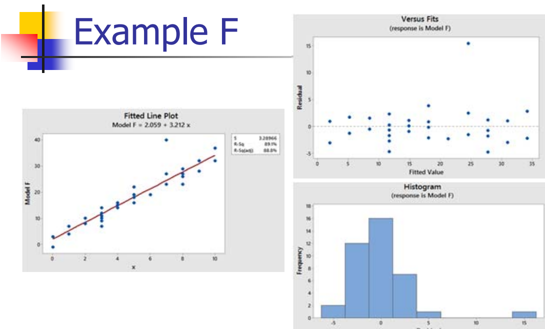
52

Example E



53

Example F



54

Using Minitab to Run Regression

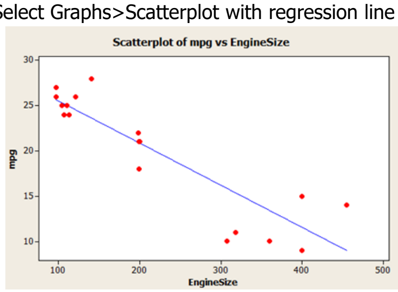
- Data shown is engine size in cubic inches (X) and MPG (Y) for 20 cars.

x	y		x	y
400	15		104	25
455	14		121	26
113	24		199	21
198	22		360	10
199	18		307	10
200	21		318	11
97	27		400	9
97	26		97	27
110	25		140	28
107	24		400	15

55

Using Minitab to Run Regression

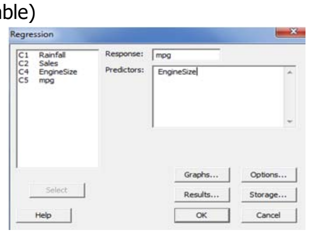
Select Graphs>Scatterplot with regression line



56

Using Minitab to Run Regression

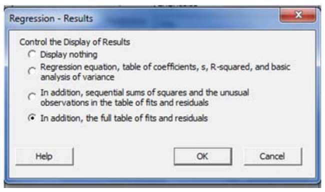
Select Statistics>Regression>Regression, then choose the Response (Y-variable) and model (X-variable)



57

Using Minitab to Run Regression

Click the results box, and choose the fits and residuals to get all predictions.



58

Using Minitab to Run Regression

The results at the beginning are the regression equation, the intercept and slope, the standard error of the residuals, and the r^2

```
The regression equation is
mpg = 30.2 - 0.0466 EngineSize
```

Predictor	Coef	SE Coef	T	P
Constant	30.203	1.361	22.20	0.000
EngineSize	-0.046598	0.005378	-8.66	0.000

```
S = 2.95688   R-Sq = 80.7%   R-Sq(adj) = 79.6%
```

59

Using Minitab to Run Regression

Next is the ANOVA table, which tests the significance of the regression model.

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	656.42	656.42	75.08	0.000
Residual Error	18	157.38	8.74		
Total	19	813.80			

60

Using Minitab to Run Regression

Finally, the residuals show the potential outliers.

Obs	EngineSize	mpg	Fit	SE Fit	Residual	St Resid
1	400	15.000	11.564	1.167	3.436	1.26
2	455	14.000	9.001	1.421	4.999	1.93
3	113	24.000	24.937	0.880	-0.937	-0.33
4	198	22.000	20.976	0.673	1.024	0.36
5	199	18.000	20.930	0.672	-2.930	-1.02
6	200	21.000	20.883	0.671	0.117	0.04
7	97	27.000	25.683	0.939	1.317	0.47
8	97	26.000	25.683	0.939	0.317	0.11
9	110	25.000	25.077	0.891	-0.077	-0.03
10	107	24.000	25.217	0.902	-1.217	-0.43
11	104	25.000	25.357	0.913	-0.357	-0.13
12	121	26.000	24.565	0.853	1.435	0.51
13	199	21.000	20.930	0.672	0.070	0.02
14	360	10.000	13.427	0.998	-3.427	-1.23
15	307	10.000	15.897	0.807	-5.897	-2.07R
16	318	11.000	15.385	0.842	-4.385	-1.55
17	400	9.000	11.564	1.167	-2.564	-0.94
18	97	27.000	25.683	0.939	1.317	0.47
19	140	28.000	23.679	0.792	4.321	1.52
20	400	15.000	11.564	1.167	3.436	1.26

61

Using Minitab to Run Regression

- Find a 95% confidence interval for the **expected** MPG of a car with an engine size of 250 ci.
- Find a 95% prediction interval for the **actual** MPG of a car with an engine size of 250 ci.

```
mpg = 30.20 - 0.04660 EngineSize


Variable    Setting
EngineSize      250


    Fit    SE Fit        95% CI              95% PI
18.5533  0.679201  (17.1264, 19.9803)  (12.1793, 24.9273)
```

62

Residual Analysis

- Residuals for Simple Linear Regression
  - The residuals should represent a linear model.
  - The standard error (standard deviation of the residuals) should not change when the value of X changes.
  - The residuals should follow a normal distribution.
  - Look for any potential extreme values of X.
  - Look for any extreme residual errors

63