

14.3 Coefficient of Determination

Best fit line

GOALS:

1. Associate a regression equation with a coefficient of determination, r^2 .
2. Understand that r^2 provides a measure of how well the line fits the data by comparing amounts of variation.
3. Compute r^2 as the ratio of the variation in the response variable, \hat{y} (from the regression equation), to the total variation in the observed y (from original data).
- *4. Interpret r^2 as the proportion of total variation that is explained by the regression line.
5. Use a calculator to compute r^2
6. As a proportion, $r^2 \leq 1$

Study Ch. 14.3, #83, 87, 89, b-d for 99, 101

[#79-83 (by hand), 89, 91 (by calc)]

Class Notes: Prof. G. Battaly, Westchester Community College, NY

[Statistics Home Page](#)

[Class Notes](#)

[Homework](#)

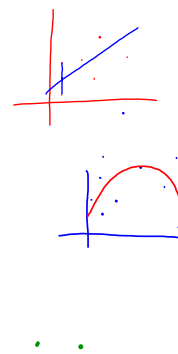
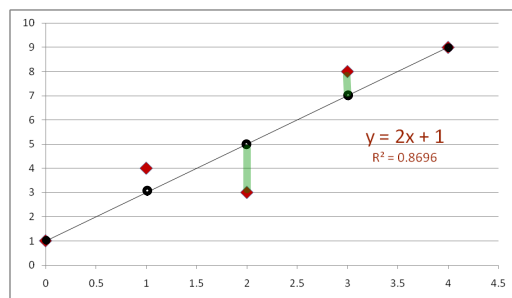


14.3 Coefficient of Determination

Best fit line

How useful is the regression line in helping to make predictions?

Is there any measure we can use to help us assess how good a prediction it is?



Coefficient of Determination, r^2 ★

Compare the variation in the predicted values to the variation in the observed values ♦

Class Notes: Prof. G. Battaly, Westchester Community College, NY

[Statistics Home Page](#)

[Class Notes](#)

[Homework](#)

14.3 Coefficient of Determination

Coefficient of Determination, r^2

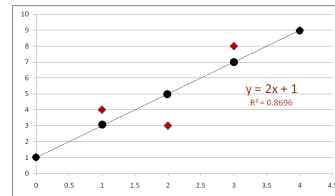
The proportion of variation in the observed values (x,y) explained by the variation in the response variable (x,y)

*response
observed*

The larger this proportion is, the better the response variable is as a predictor of the observed variable.

Since observed variation is the total variation, this proportion has a value between 0 and 1.

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



Class Notes: Prof. G. Battaly, Westchester Community College, NY

[Statistics Home Page](#)

[Class Notes](#)

[Homework](#)

14.3 Coefficient of Determination

Coefficient of Determination, r^2

The proportion of variation in the observed values explained by the variation in the response variable

The larger this proportion is, the better the response variable is as a predictor of the observed variable.

$$r^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}^i - \bar{y})^2}{\sum (y - \bar{y})^2} \frac{\text{response}}{\text{observed}}$$

or $r^2 = \frac{S_y^2}{S_y^2}$

remember

$$S_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

Class Notes: Prof. G. Battaly, Westchester Community College, NY

[Statistics Home Page](#)

[Class Notes](#)

[Homework](#)



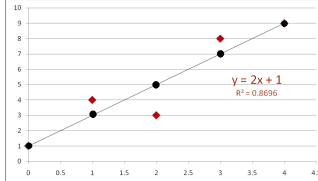
14.3 Coefficient of Determination

$$r^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

response variation / *observed*

previous example; for coefficient of determination

| x | y | \hat{y} 2x+1 | $(y - \bar{y})^2$ SST | $(\hat{y} - \bar{y})^2$ SSR |
|------|----|-------------------|--------------------------|--------------------------------|
| 0 | 1 | 1 | | |
| 4 | 9 | 9 | | |
| 3 | 8 | 7 | | |
| 1 | 4 | 3 | | |
| 2 | 3 | 5 | | |
| 10 | 25 | 25 | | |
| mean | 2 | 5 | r^2 | 0.870 |



Coefficient of Determination = 40/46 = 0.870

87.0% of variation in observations of y is explained by the regression equation $\hat{y} = 2x + 1$

Class Notes: Prof. G. Battaly, Westchester Community College, NY

[Statistics Home Page](#)

[Class Notes](#)

[Homework](#)

14.3 Coefficient of Determination

$$\frac{40}{46} = \frac{20}{23} \quad r^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

response variation / *observed*

for coefficient of determination

| x | y | \hat{y} 2x+1 | $(y - \bar{y})^2$ SST | SSR |
|----|----|-------------------|--------------------------|---------------------|
| 0 | 1 | 1 | $(-5)^2 = 25$ 16 | $(-4)^2 = 16$ 16 |
| 4 | 9 | 9 | $(4-5)^2 = 1$ 16 | $(4-5)^2 = 1$ 16 |
| 3 | 8 | 7 | $(8-5)^2 = 9$ 9 | $(7-5)^2 = 4$ 4 |
| 1 | 4 | 3 | $(4-5)^2 = 1$ 1 | $(-2)^2 = 4$ 4 |
| 2 | 3 | 5 | $(3-5)^2 = 4$ 4 | $(5-5)^2 = 0$ 0 |
| 10 | 25 | 25 | 46 | 40 |
| 2 | 5 | 5 | r^2 | 0.87 |

$= \frac{40}{46}$

Coefficient of Determination = 40/46 = 0.870

87.0% of variation in observations of y is explained by the regression equation $\hat{y} = 2x + 1$

Class Notes: Prof. G. Battaly, Westchester Community College, NY

[Statistics Home Page](#)

[Class Notes](#)

[Homework](#)

14.3 Coefficient of Determination

$SSE = SST - SSR$

error betw obs and response
total error in y
error in response

Total Error = Error explained by R + Error not Explained
 $SST = SSR + SSE$

Therefore, when SSE, the error between the observed y and the regression equation, is a minimum, the equation is a better predictor, because SSR is a higher proportion of SST, or the regression equation explains more of the variation in the original data.

| x | y | 2x+1 | diff | diff ² |
|----|----|------|------|-------------------|
| 0 | 1 | 1 | 0 | 0 |
| 4 | 9 | 9 | 0 | 0 |
| 3 | 8 | 7 | 1 | 1 |
| 1 | 4 | 3 | 1 | 1 |
| 2 | 3 | 5 | -2 | 4 |
| 10 | 25 | 25 | 0 | 6 |

Class Notes: Prof. G. Battaly, Westchester Community College, NY

[Statistics Home Page](#)

[Class Notes](#)

[Homework](#)



14.3 Coefficient of Determination

$SSE = SST - SSR = 46 - 40 = 6$

error betw obs and response
total error in y
error in response

least squares

| x | y | 2x+1 | SST | SSR | SSE |
|----|----|------|----------------|------|-----|
| 0 | 1 | 1 | 16 | 16 | 0 |
| 4 | 9 | 9 | 16 | 16 | 0 |
| 3 | 8 | 7 | 9 | 4 | 1 |
| 1 | 4 | 3 | 1 | 4 | 1 |
| 2 | 3 | 5 | 4 | 0 | 4 |
| 10 | 25 | 25 | 46 | 40 | 6 |
| 2 | 5 | 5 | r ² | 0.87 | |

s

3.391

3.162

$r = 3.162/3.391 = 0.9325; r^2 = 0.869$

Class Notes: Prof. G. Battaly, Westchester Community College, NY

[Statistics Home Page](#)

[Class Notes](#)

[Homework](#)



14.3 Coefficient of Determination

Find on Calculator:

STAT / TESTS / LinRegTTest

xlist: L1

Ylist: L2

2 tailed, left tailed, right tailed

calculate

Output:

$y=ax+b$

2 tailed, rt tailed, lf tailed

t=

p=

df=

a=

b=

s=

$r^2 = .$

r =

$y = 2x + 1$

$r^2 = 0.86956$

$r = 0.9325$

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page

©Gertrude Battaly, 2014

Class Notes

Homework

from 14.2 The Regression Equation

A random sample of custom homes for sale include the following information: a size of x hundred sq. ft selling at y thousand. Predict the price of a home that is 2600 ft.

| x | y |
|----|-----|
| 26 | 540 |
| 27 | 555 |
| 33 | 575 |
| 29 | 577 |
| 29 | 606 |
| 34 | 661 |
| 30 | 738 |
| 40 | 804 |
| 22 | 496 |

Y1(L1) gets predicted values, \hat{y} , if x is in L1

STAT/CALC/
LinReg(ax+b) L1, L2, Y1

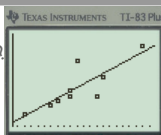
LinReg
 $y=ax+b$
 $a=15.89351852$
 $b=140.0833333$

calculate regression line

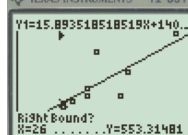
Thus, $y = 15.894x + 140.083$
or increase of \$15,894 for each additional square foot

STATPLOT/Plot1/ON/
scattergram
L1
L2
ZOOM/ STAT

does data look linear?



Use function $Y=$ to predict
/ GRAPH
/2nd CALC
/ value / 62 enter



Result: $y = 553.315$ Understand prediction
[from substitution $y = 15.894(26) + 140.083$]
Conclude: A home of 2600 sq. ft will cost \$553,315.

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page

Class Notes

Homework

14.3 Coefficient of Determination



A random sample of custom homes for sale include the following information: a size of x hundred sq. ft selling at \$ y thousand. Predict the price of a home that is 2600 ft.

Find: r^2 ; interpret; is the regression equation useful in making predictions?

Use LinRegTTest to get r^2

| x | y |
|-----|-----|
| 26 | 540 |
| 27 | 555 |
| 33 | 575 |
| 29 | 577 |
| 29 | 606 |
| 34 | 661 |
| 30 | 738 |
| 40 | 804 |
| 22 | 496 |

Class Notes: Prof. G. Battaly, Westchester Community College, NY



14.3 Coefficient of Determination



A random sample of custom homes for sale include the following information: a size of x hundred sq. ft selling at \$ y thousand. Predict the price of a home that is 2600 ft.

Find: r^2 ; interpret; is the regression equation useful in making predictions?

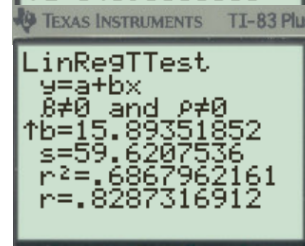
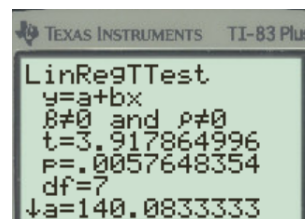
Use LinRegTTest to get r^2

| x | y |
|-----|-----|
| 26 | 540 |
| 27 | 555 |
| 33 | 575 |
| 29 | 577 |
| 29 | 606 |
| 34 | 661 |
| 30 | 738 |
| 40 | 804 |
| 22 | 496 |

$r^2 = 0.68679$

68.7% of the variation in home prices is explained by the regression equation

The regression equation is useful for predictions, with the understanding that there are other contributing factors that help to determine the price.



Class Notes: Prof. G. Battaly, Westchester Community College, NY

