14.2 The Regression Equation

GOALS:

1. Use a scatter diagram to display data for 2 variables visually. Does the relationship look linear?
2. Understand that the least squares, or best fit line, is the line that is closest to all the data points.
3. The best fit regression equation has the smallest possible vertical distance between the y values of the points (observed) and the regression line (predicted).
4. Recognize the equations for the slope and y-intercept of the regression line.
5. Use the calculator to compute the regression line.

Study Ch. 14.2, #35, 37, 51, 59-63

[# 35-39, 47 (by hand), 51-59 (by calc)]

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

Statistics Home Page Class Notes Homework Best fit line

14.2 The Regression Equation

χ^2 enabled qualitative comparisons:

- given distribution compared to expected
- dependent or independent

What about **quantative variables** that seem to vary relative to each other.

- Cost of used cars: age vs price
- Height and weight
- PCBs and survival in fish
- Wing chord and mass in Saw-whets

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

Statistics Home Page Class Notes Homework

14.2 The Regression Equation

Example

Given:
Scuba divers have maximum dive times they cannot exceed when diving to different depths. The data below shows the maximum dive times in minutes for different depths.

Depth(ft)	MaxTime(min)
50	80
60	55
70	45
80	35
90	25
100	22

Find:
If the divers need to investigate a find that is 62 feet deep, how long can they safely remain at work with each dive?

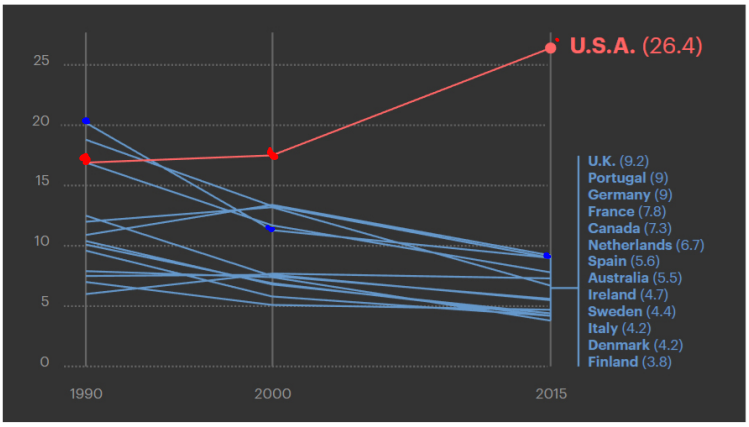
Question:
Can we use a mathematical model to predict the maximum time to safely dive 62 feet? If so, what model? How reliable will it be?

One answer:
Plot the points. If they appear to be approximately linear, use linear regression: generate the best fit line and use it to predict a time. If data is normal, use t-test to find reliability.

14.2 The Regression Equation

Example of a line graph - **NOT** what we want!

Maternal Mortality Is Rising in the U.S. As It Declines Elsewhere



Per 100,000 live births. Source: "Global, regional, and national levels of maternal mortality, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015," The Lancet. Note: Only data for 1990, 2000 and 2015 was made available in the journal.

1. Only shows 3 points of the 25 years.
2. Does show trends, but **does not allow a reliable prediction.**
3. **Need a mathematical model.**

A **REGRESSION LINE is a mathematical model** that enables a reliable prediction when assumptions are met.

© G. Battaly 2019

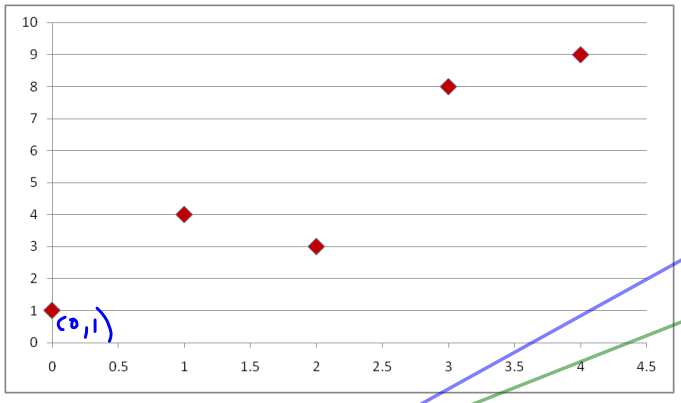
14.2 The Regression Equation

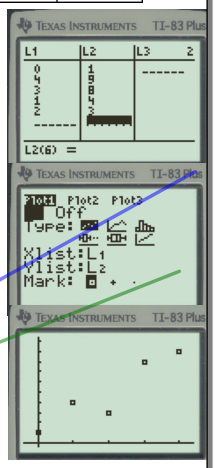
Different, simpler example for demonstration

Note: The 2 lines presented here are not the best fit. They are intended to provide a comparison and show why the best fit is the best of many possible lines.

1. Plot the points as a SCATTERGRAM
2. Try to fit a line to the data.
3. **Which line is best?**
4. Will any line work?
5. How do find the best fitting line?

x	y			
0	1			
4	9			
3	8			
1	4			
2	3			





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

Statistics Home Page Class Notes Homework

14.2 The Regression Equation

exploring one possibility

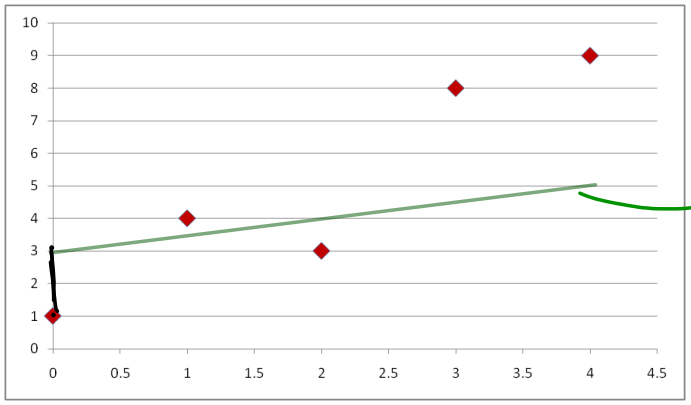
Try 2 possible lines: $\hat{y} = 0.5x + 3$

$y = 1.5x + 1$

x	y	$\hat{y} = \frac{1}{2}x + 3$	$y - \hat{y}$	$(y - \hat{y})^2$
0	1			
4	9			
3	8			
1	4			
2	3			

Without a direct approach, we might explore a few possible lines.

But, we need a way to compare them to decide which is a better model.



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

Statistics Home Page Class Notes Homework

14.2 The Regression Equation

exploring one possibility

Try 2 possible lines: $\hat{y} = 0.5\,x + 3$

$y = 1.5\,x + 1$

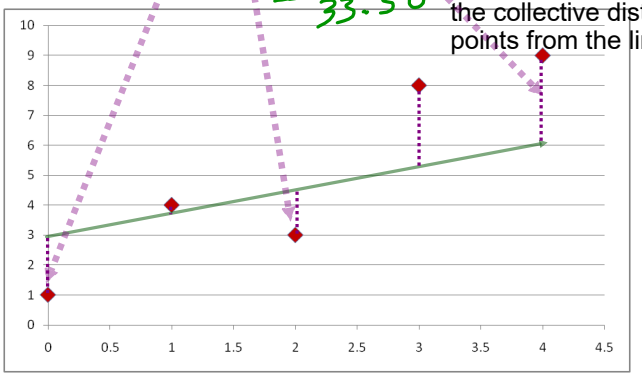
x	y	$\hat{y} = \frac{1}{2}x+3$	$y-\hat{y}$	$(y-\hat{y})^2$
0	1	3	-2	4
4	9	5	4	16
3	8	4.5	3.5	12.25
1	4	3.5	0.5	0.25
2	3	4	-1	1
				33.50

$\left(\frac{7}{2}\right)^2 = \frac{49}{4}$

$\frac{3}{2}x+1$

x	y	.5x+3	diff	diff^2
0	1	3	-2	4
4	9	5	4	16
3	8	4.5	3.5	12.25
1	4	3.5	0.5	0.25
2	3	4	-1	1
10	25	20	5	33.5

sums of squares: represents the collective distance of the points from the line



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

Statistics Home Page Class Notes Homework

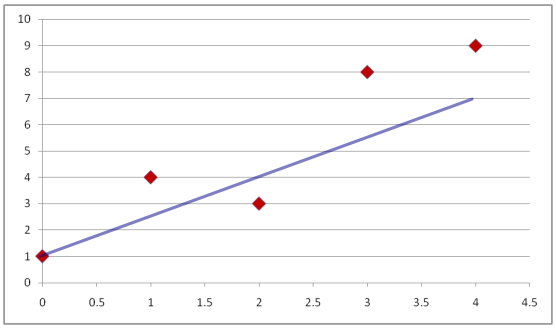
14.2 The Regression Equation

exploring another possibility

Try 2 possible lines: $y = 0.5\,x + 3$

$\hat{y} = 1.5\,x + 1$

x	y	$\hat{y}$ $\frac{3}{2}x+1$	$y-\hat{y}$	$(y-\hat{y})^2$
0	1			
4	9			
3	8			
1	4			
2	3			



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

Statistics Home Page Class Notes Homework

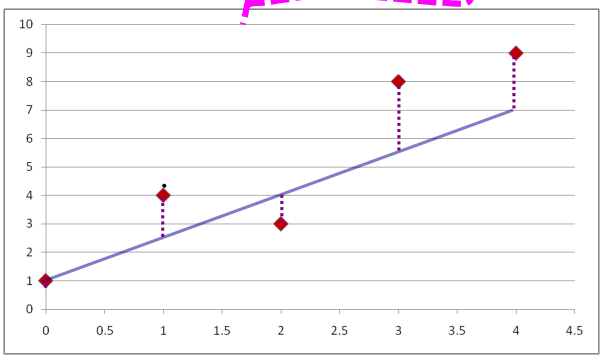
14.2 The Regression Equation

Try 2 possible lines: $y = 0.5\,x + 3$ and $y = 1.5\,x + 1$

x	y	$\hat{y} = \frac{3}{2}x+1$	$y-\hat{y}$	$(y-\hat{y})^2$
0	1	1	0	0
4	9	7	2	4
3	8	5.5	2.5	6.25
1	4	2.5	1.5	2.25
2	3	4	-1	1
				13.50

x	y	1.5x+1	diff	diff^2
0	1	1	0	0
4	9	7	2	4
3	8	5.5	2.5	6.25
1	4	2.5	1.5	2.25
2	3	4	-1	1
10	25	20	5	13.5

sums of squares: represents the collective distance of the points from the line



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

Statistics Home Page Class Notes Homework

14.2 The Regression Equation

2 possible lines: $y = 0.5\,x + 3$ Sums of squares: 33.5

$y = 1.5\,x + 1$ 13.5

Which is a better fit? 33.5 vs 13.5?

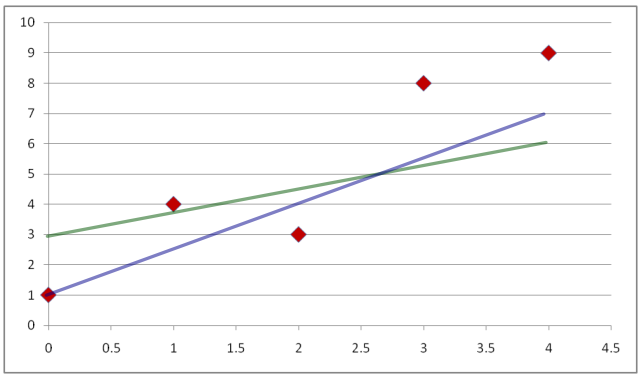
BUT, **really want the very best fit**

smallest sum of squares

or

LEAST SQUARES

x	y			
0	1			
4	9			
3	8			
1	4			
2	3			



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

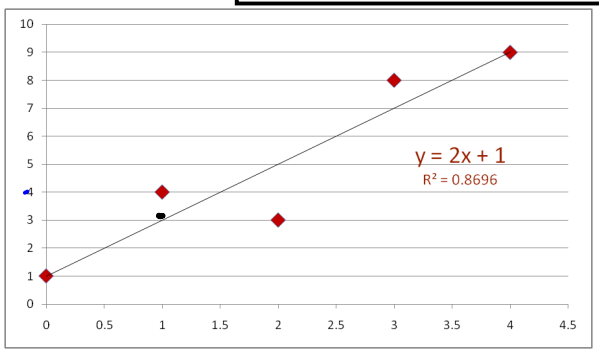
Statistics Home Page Class Notes Homework

14.2 The Regression Equation

Note: y = 2x+1 is the best fit line for this data. The sum of squares is smaller than for all other lines that might seem to fit the data.

Best fit line

x	y	2x+1	diff	diff^2
0	1			
4	9			
3	8			
1	4			
2	3			
10	25			

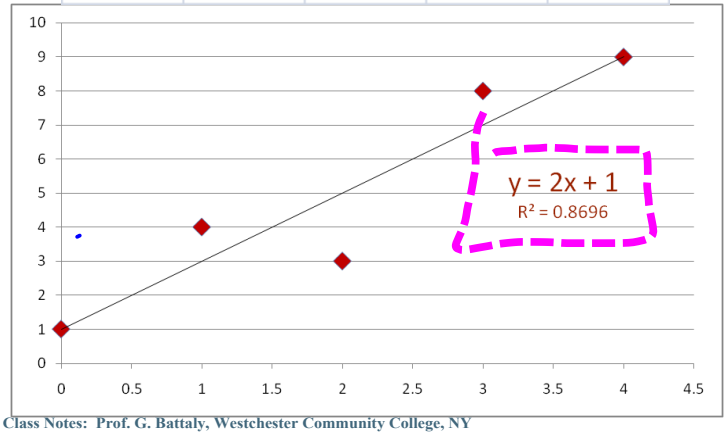


Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework

14.2 The Regression Equation

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

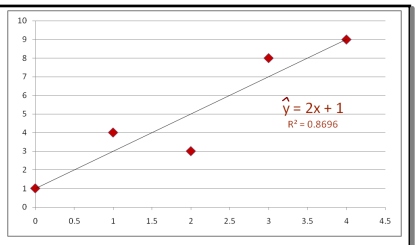
14.2 The Regression Equation

How to find the best fit line?

$\hat{y} = b_1 x + b_0$

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{\sum(x-\bar{x})^2} = \frac{\sum xy - (\sum x \sum y)/n}{\sum x^2 - (\sum x)^2/n}$$

Math that results in **least squares** line.



$$b_0 = \bar{y} - b_1 \bar{x}$$

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - (\sum x)^2/n}{n-1}}$$

x	y	x^2	xy
0	1		
4	9		
3	8		
1	4		
2	3		
10	25		

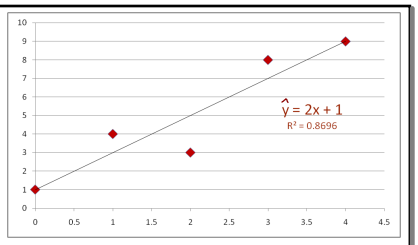
Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework

14.2 The Regression Equation

How found?

$\hat{y} = b_1 x + b_0$



$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum xy - (\sum x \sum y)/n}{\sum x^2 - (\sum x)^2/n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$\hat{y} = 2x + 1$

x	y	x^2	xy
0	1	0	0
4	9	16	36
3	8	9	24
1	4	1	4
2	3	4	6
10	25	30	70

$\bar{x} = 2$ $\bar{y} = 5$

$$b_1 = \frac{\sum xy - (\sum x \sum y)/n}{\sum x^2 - (\sum x)^2/n} = \frac{70 - \frac{10(25)}{5}}{30 - \frac{10^2}{5}} = \frac{70-50}{30-20} = \frac{20}{10} = 2$$

$b_0 = 5 - 2(2) = 1$

$\hat{y} = 2x + 1$

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

Statistics Home Page Class Notes Homework

14.2 The Regression Equation

Find on Calculator:

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

Note: Use STAT/CALC/LinReg(ax+b) L1, L2, Y1
to place the regression equation in Y1 for graphing and prediction

To get Y1, select:
VARS/Y-VARS/1:Function

Output:

a = 2

b = 1

143 $r^2 = .869565...$

144 $r = .9325...$

$\hat{y} = ax + b$

$\hat{y} = 2x + 1$

To get r and r^2 as output from LinReg:
Catalog (2nd/0)
Turn on DiagnosticOn by
selecting it, and entering to execute

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

Statistics Home Page Class Notes Homework

14.2 The Regression Equation

Given:
Scuba divers have maximum dive times they cannot exceed when diving to different depths. The data below shows the maximum dive times in minutes for different depths.

Depth(ft)	MaxTime(min)
50	80
60	55
70	45
80	35
90	25
100	22

Find:
If the divers need to investigate a find that is 62 feet deep, how long can they safely remain at work with each dive?

Use linear regression to predict:

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

Statistics Home Page Class Notes Homework

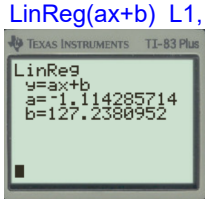
14.2 The Regression Equation

Given:
Scuba divers have maximum dive times they cannot exceed when diving to different depths. The data below shows the maximum dive times in minutes for different depths.

Depth(ft)	MaxTime(min)
50	80
60	55
70	45
80	35
90	25
100	22

Find:
If the divers need to investigate a find that is 62 feet deep, how long can they safely remain at work with each dive? Use linear regression to predict:

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

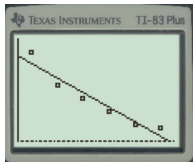


calculate regression line

Thus, y = -1.11x +127.24
or 1.11 min less for each additional foot of depth

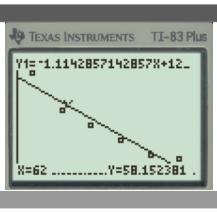
STATPLOT/Plot1/ON/
scattergram
L1
L2
ZOOM/ STAT

does data look linear?



Use function to predict

Y=
/ GRAPH
/2nd CALC
/ value / 62 enter



Result: y = 58.152 Understand prediction
[from substitution y = -1.11**(62)** +127.24]
Conclude: can stay under water for 58 min
if diving to depth of 62 feet

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

Statistics Home Page Class Notes Homework

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 sq ft.

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

Statistics Home Page Class Notes Homework

14.2 The Regression Equation

Terms:

Predictor variable size

Response variable price

extrapolation - outside range of data, generally not valid

outlier - far from regression line

influential observation - if left out changes dramatically

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

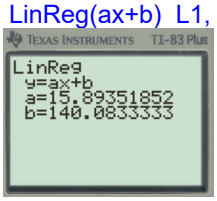
Statistics Home Page Class Notes Homework

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

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

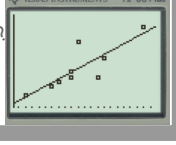


calculate regression line

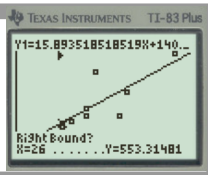
Thus, y = 15.894x +140.083
or increase of $15,894 for each additional square foot

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

does data look linear?



Use function to predict
Y=
/ 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