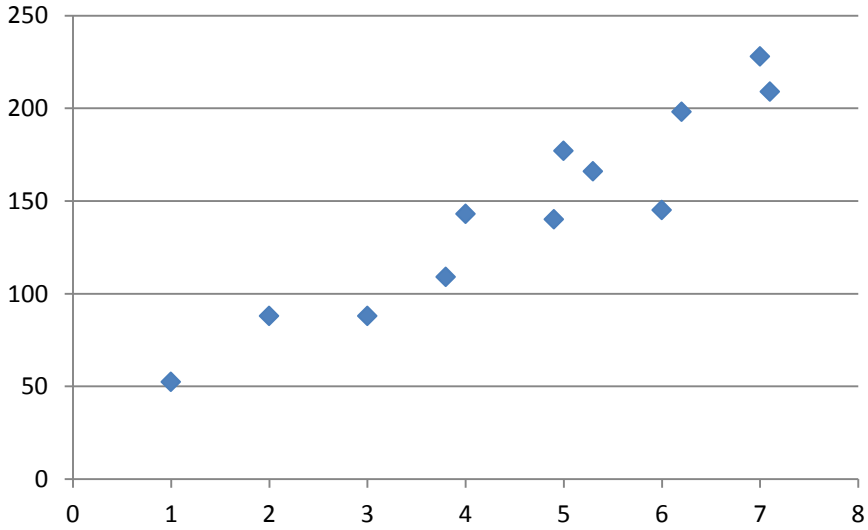
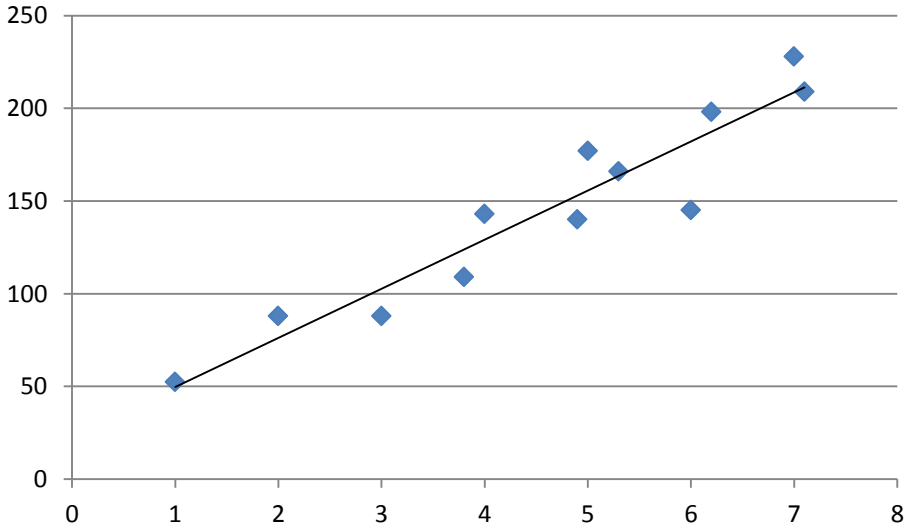
Linear Regression Analysis

Analysis of paired data and
using a given value of one variable to predict the value of the other





Linear Regression Analysis

Ex: The chirp rate of crickets is strongly correlated with the outside temperature. On 8 different days a statistician went outside and measured two quantities: the number of chirps a cricket made in a minute (x) and the outside temperature (y) . The data is in the table below.

X	Chirps in 1 minute	88	118	110	86	120	103	96	90
Y	Outside Temperature (˚F)	69.7	93.3	84.3	76.3	88.6	82.6	71.6	79.6

a) Find the linear correlation coefficient *r*

b) Find the equation of the least squares regression line

c) Predict the outside temperature if you hear a cricket chirping 140 times per minute

d) Find a 95% prediction interval for the outside temperature if a cricket is chirping 140 times per minute

Linear Regression Analysis

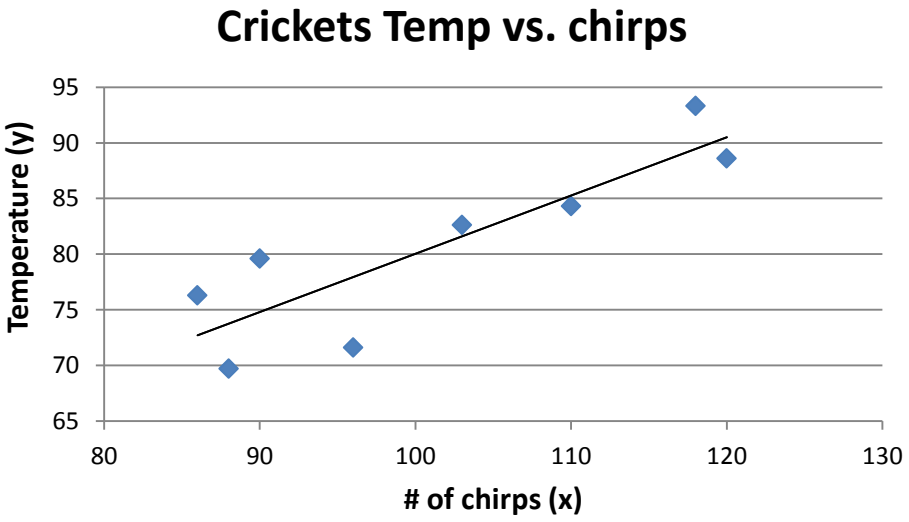
Ex (continued):

X	Chirps in 1 minute	88	118	110	86	120	103	96	90
Y	Outside Temperature (˚F)	69.7	93.3	84.3	76.3	88.6	82.6	71.6	79.6

Predictor Variable: X = Number of chirps in 1 minute

Response Variable: Y = Outside Temperature

Scatter Plot



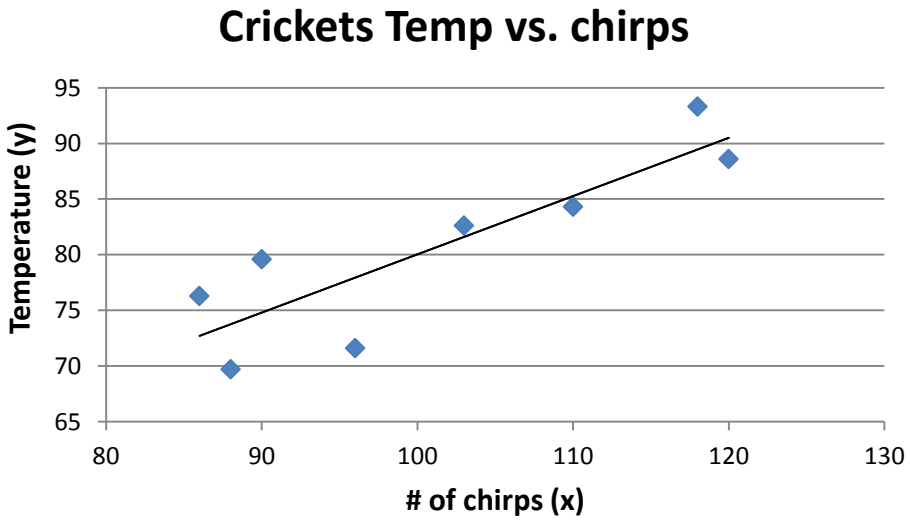
a) r = 0.8693

b) $\hat{y}$ = 0.5236x + 27.674

Linear Regression Analysis

Ex (continued):

X	Chirps in 1 minute	88	118	110	86	120	103	96	90
Y	Outside Temperature (˚F)	69.7	93.3	84.3	76.3	88.6	82.6	71.6	79.6



a) r = 0.8693

b) $\hat{y}$ = 0.5236x + 27.674

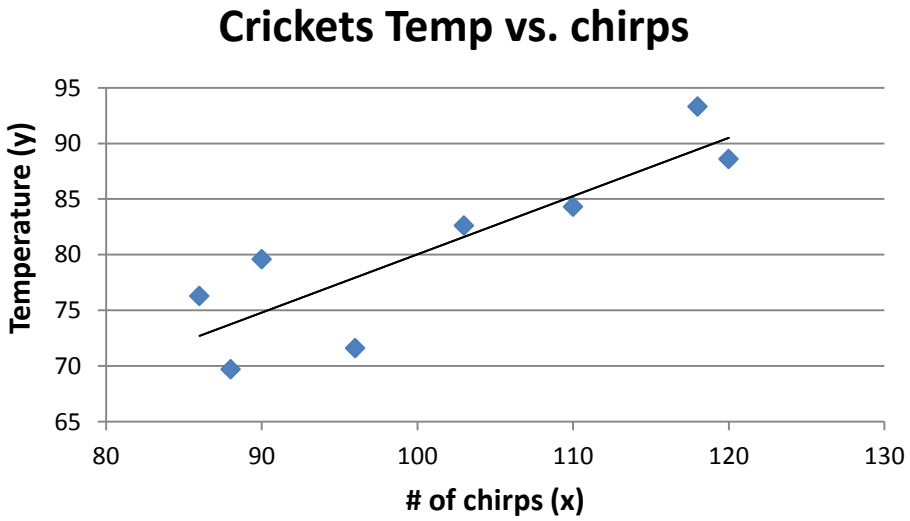
c) If a cricket is chirping at 140 chirps per minute, the outside temperature is about

$\hat{y}$ = 0.5236(140) + 27.674 = 101 ^{o}F

Linear Regression Analysis

Ex (continued):

X	Chirps in 1 minute	88	118	110	86	120	103	96	90
Y	Outside Temperature (˚F)	69.7	93.3	84.3	76.3	88.6	82.6	71.6	79.6



a) r = 0.8693

b) $\hat{y}$ = 0.5236x + 27.674

c) If x = 140 chirps / min,
$\hat{y}$ = 101 ° F

d) For a 95% prediction interval, the margin or error E is

E = 15.9 ˚F

and the prediction interval is $\hat{y} - E < y < \hat{y} + E$ or $85.1^{\circ}F < y < 116.9^{\circ}F$

The Linear Correlation Coefficient *r*

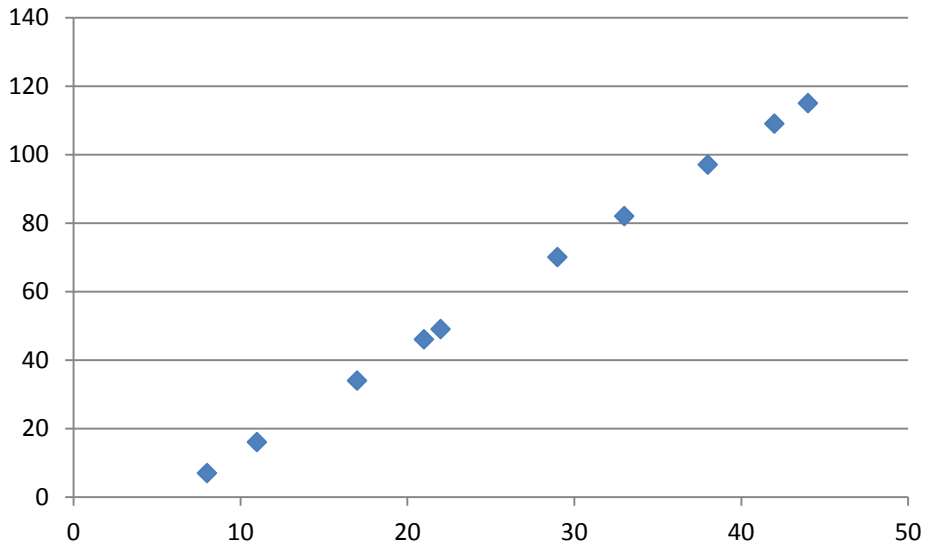
The Linear Correlation Coefficient *r*

- *r* is a number that can be calculated from paired data that tells you how close to a straight line the graph of the data is
- *r* is always between -1 and 1
- When *r* is 1 or -1 , the data form a perfect straight line
- When *r* is close to 1 or -1 , the graph of the data looks almost like a straight line, but with a little scatter
- The further away *r* is from -1 and 1 (i.e. closer to 0), the less the data looks like a straight line. It can look scattered all over the place, or can form a neat shape (but not a line)
- If *r* is positive, the data trends upward. That means:

  when x increases, y increases the slope of the "line" is positive
- If *r* is negative, the data trends downward. That means:

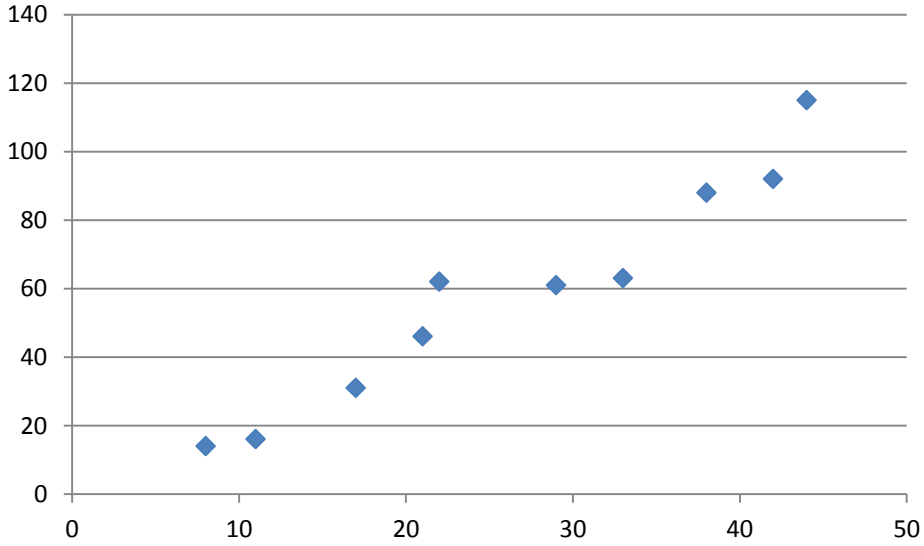
  when x increases, y decreases the slope of the "line" is negative

The Linear Correlation Coefficient *r*



r is positive when x ↑ , y ↑

r = 1

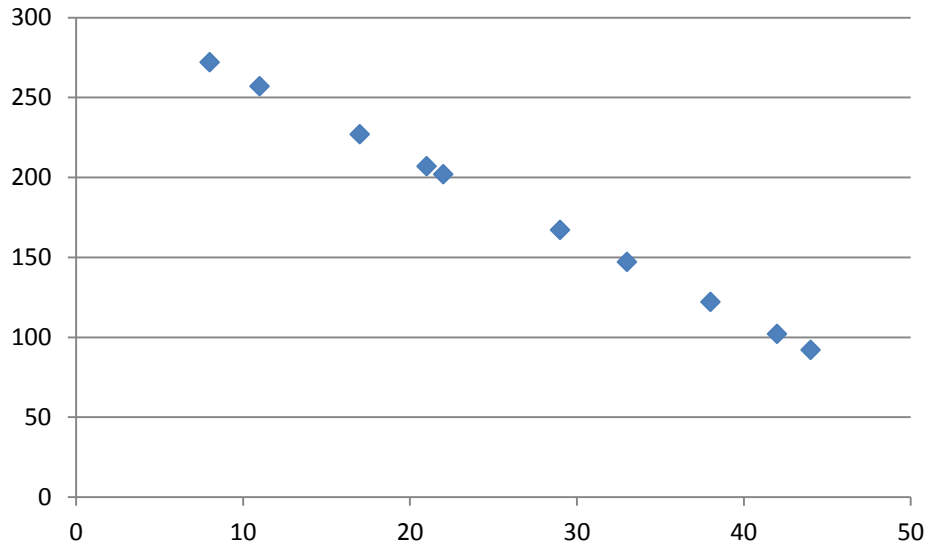


r is positive when x ↑ , y ↑

r is close to 1

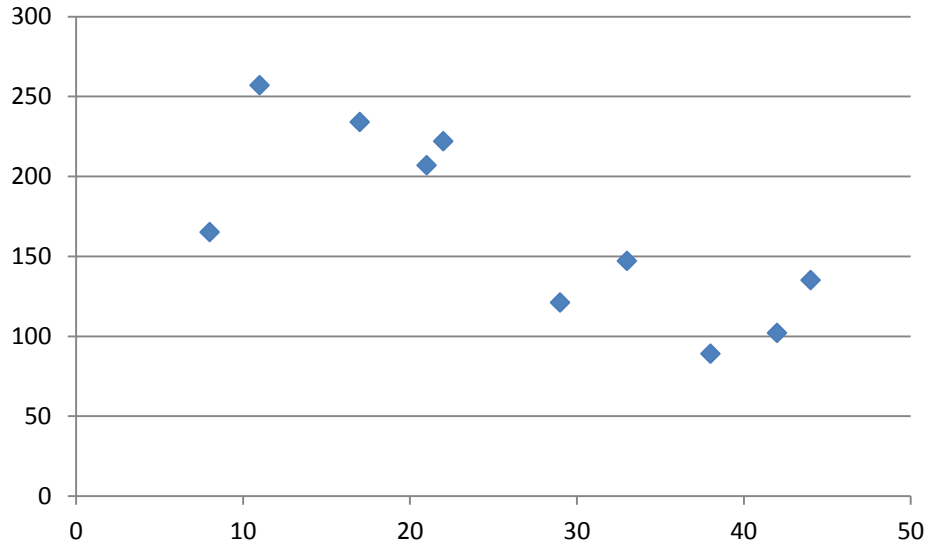
(actual *r* = 0.9698)

The Linear Correlation Coefficient *r*



r is negative when x ↑ , y ↓

$r = -1$

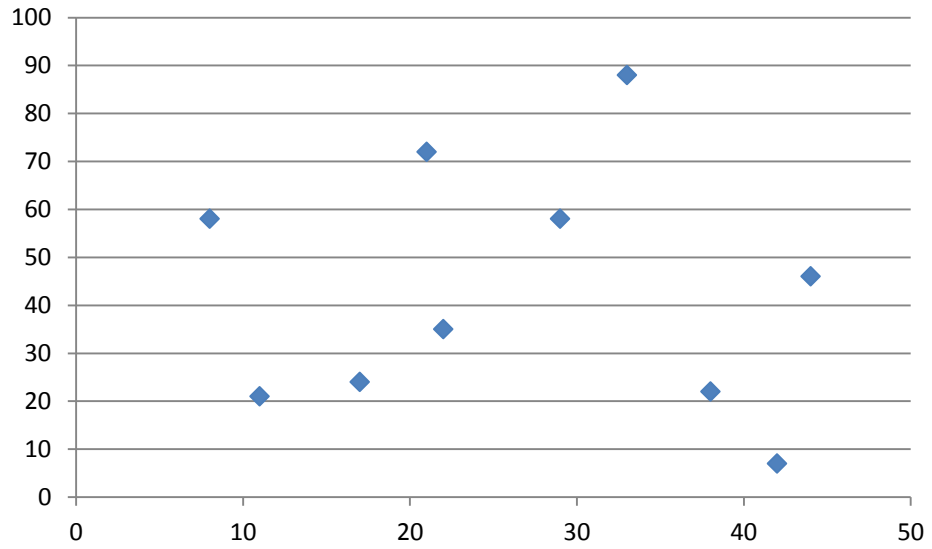


r is negative when x ↑ , y ↓

r is close to -1

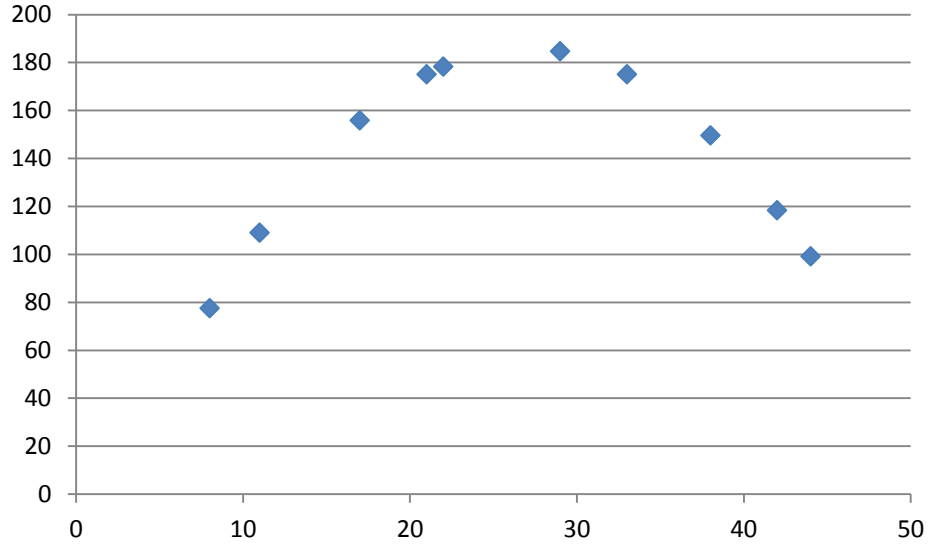
(actual $r = -0.7705$)

The Linear Correlation Coefficient *r*



r is close to 0 when x ↑ , y ???

(actual r = -0.1010)



r is close to 0 when x ↑ , y ???

(actual r = 0.1183)

The Linear Correlation Coefficient *r*

The formula for *r* is:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n\left(\sum x^2\right) - (\sum x)^2}\sqrt{n\left(\sum y^2\right) - (\sum y)^2}}$$

(more on this later…)

The Least-Squares Regression Line

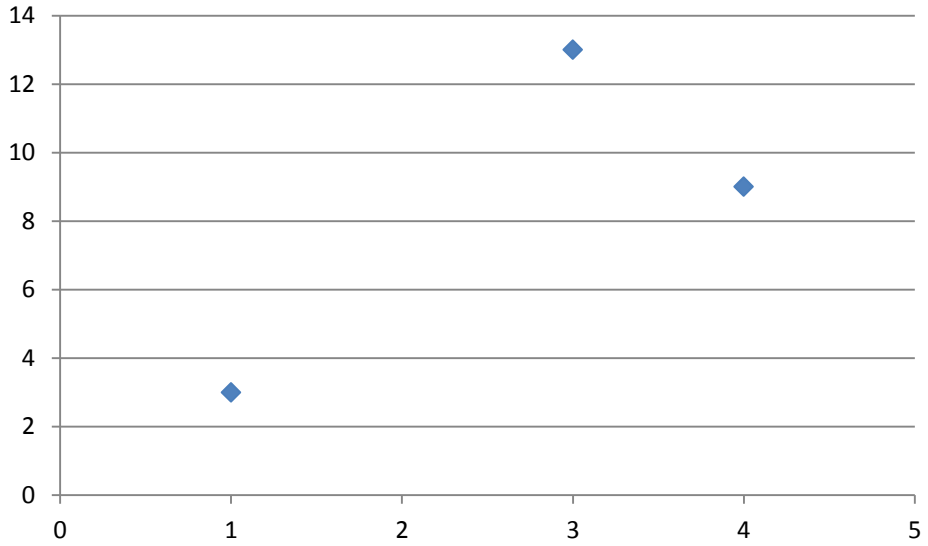
The Least-Squares Regression Line

Goal: We want to find a line that is as close as possible to all the data points

Every line has a "total error" and we are looking for the line with the smallest "total error"

Ex: Data

x	1	3	4
y	3	13	9



Let's see which line is "closer" to this data. Our choices in this example are

$$\hat{y} = -2x + 13$$

or

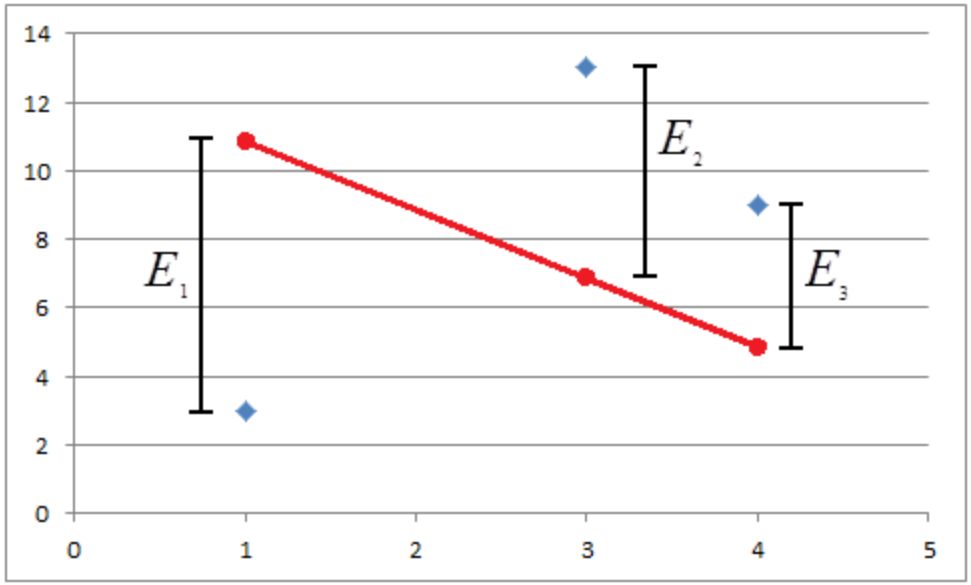
$$\hat{y} = 2x + 2$$

The Least-Squares Regression Line

Ex (continued): Data

x	1	3	4
y	3	13	9

$\hat{y} = -2x + 13$



$E_1 = y - \hat{y} = [3] - [-2(1) + 13] = -8$

$E_2 = y - \hat{y} = [13] - [-2(3) + 13] = 6$

$E_3 = y - \hat{y} = [9] - [-2(4) + 13] = 4$

What is the "total error"?

$E_1 + E_2 + E_3$?

No because we don't want negative numbers to cancel out positive ones

$|E_1| + |E_2| + |E_3|$?

No because absolute values are hard to deal with in math

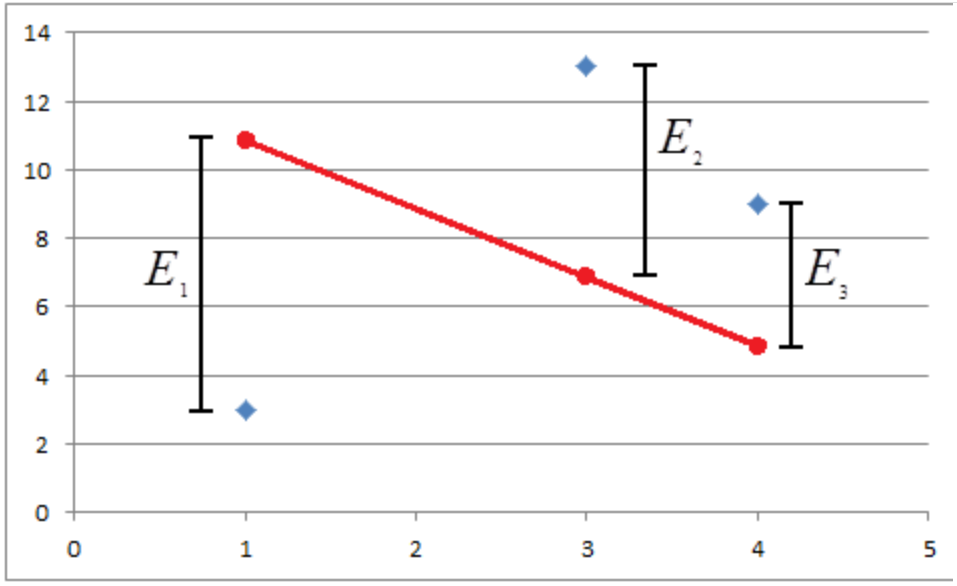
So we'll use $E_1^2 + E_2^2 + E_3^2$

The Least-Squares Regression Line

Ex (continued): Data

x	1	3	4
y	3	13	9

$\hat{y} = -2x + 13$



$E_1 = y - \hat{y} = [3] - [-2(1) + 13] = -8$

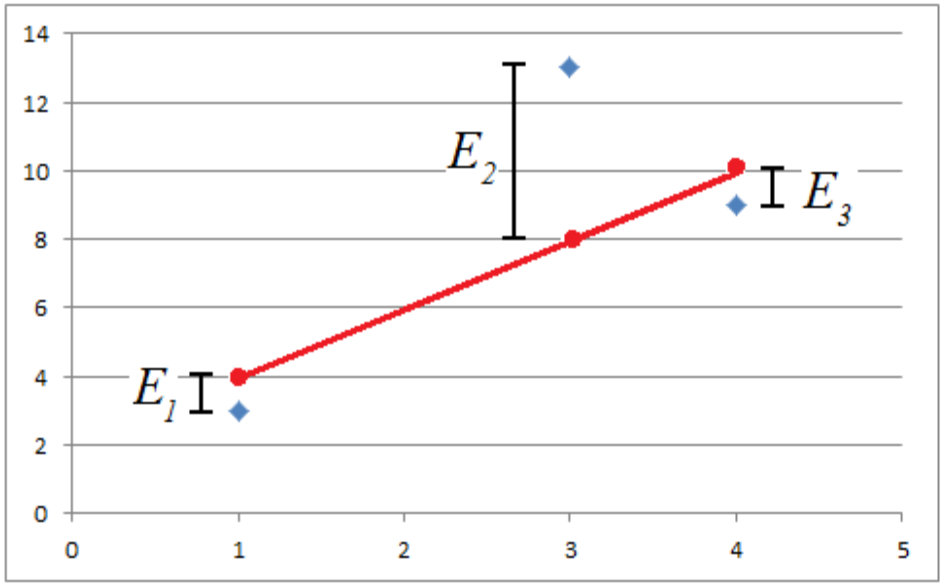
$E_2 = y - \hat{y} = [13] - [-2(3) + 13] = 6$

$E_3 = y - \hat{y} = [9] - [-2(4) + 13] = 4$

Total error =

$E_1^2 + E_2^2 + E_3^2 = (-8)^2 + (6)^2 + (4)^2 = 116$

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



$E_1 = y - \hat{y} = [3] - [2(1) + 2] = -1$

$E_2 = y - \hat{y} = [13] - [2(3) + 2] = 5$

$E_3 = y - \hat{y} = [9] - [2(4) + 2] = -1$

Total error =

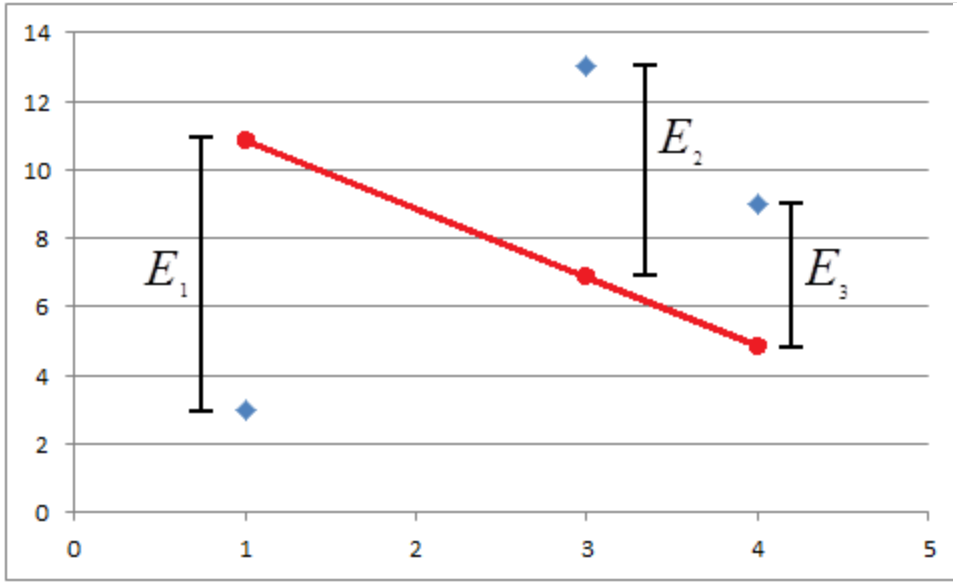
$E_1^2 + E_2^2 + E_3^2 = (-1)^2 + (5)^2 + (-1)^2 = 27$

The Least-Squares Regression Line

Ex (continued): Data

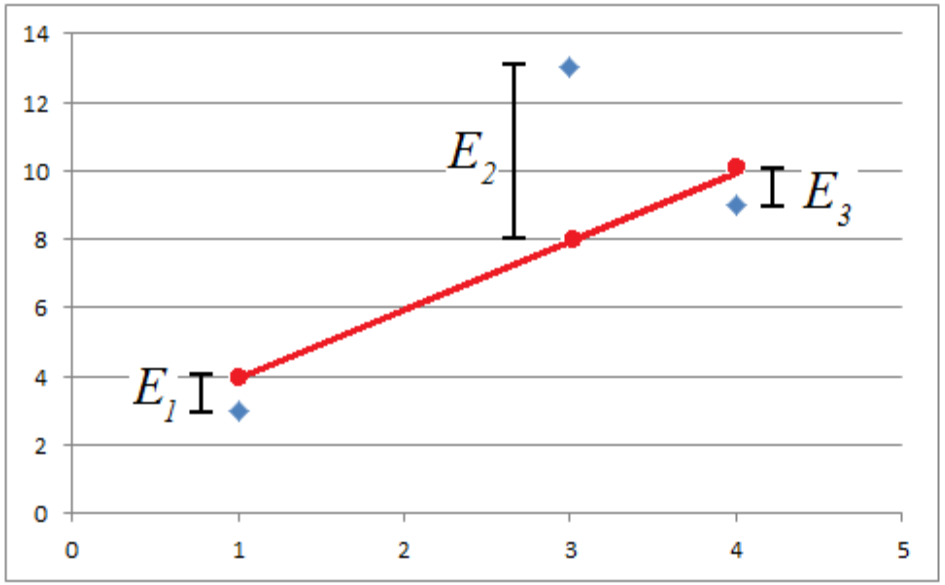
x	1	3	4
y	3	13	9

$\hat{y} = -2x + 13$



Total error = 116

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



Total error = 27

So the line $\hat{y} = 2x + 2$ is "closer" to the data than the line $\hat{y} = -2x + 13$

Is there an even "closer" line???

The Least-Squares Regression Line

Using calculus, it can be shown that given some data, there is a line that is "closer" to the data than any other line. This line is called the Least-Squares Regression Line (LSRL)

Formula for the Least-Squares Regression Line

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

$$b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} \qquad b_0 = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

(more on this later…)

Prediction Intervals

Prediction Intervals

The Least-Squares Regression Line can be used to predict the value of *y* given the value of *x*. The predicted value of *y* is called $\hat{y}$

The problem: $\hat{y}$ is a point estimate because it is a 1 number guess of the value of *y*. Instead, we want an interval estimate (a confidence interval). When finding confidence intervals for problems involving paired data, they are called Prediction Intervals.

As usual, to find a prediction interval, we need the margin or error formula *E*, and

$$\hat{y} - E < y < \hat{y} + E$$

Summary of Formulas

Correlation Coefficient *r*:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n\left(\sum x^2\right) - (\sum x)^2}\sqrt{n\left(\sum y^2\right) - (\sum y)^2}}$$

LSRL:

$$b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} \qquad b_0 = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

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

Prediction Interval (use *t* – distribution):

$$df = n - 2 \qquad s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\left(\sum x^2\right) - \left(\sum x\right)^2}} \qquad \hat{y} - E < y < \hat{y} + E$$

A Detailed Linear Regression Problem

A Detailed Linear Regression Problem

Ex: In order to study global warming, the data below was taken over different years (CO_2 levels in parts per million and temperature in °C)

CO_2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

a) Find the linear correlation coefficient r

b) Find the equation of the least squares regression line

c) Predict the temperature of the Earth if the CO_2 levels reach 400 parts per million

d) Construct a 90% prediction interval for the temperature of the Earth if the CO_2 levels reach 400 parts per million

A Detailed Linear Regression Problem

Ex (continued):

CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

Before we do any of the parts of this problem, we need to find

$$\Sigma x \qquad \Sigma x^2 \qquad \Sigma y \qquad \Sigma y^2 \qquad \Sigma xy \qquad \bar{x}$$

because these appear in many of the formulas.

A Detailed Linear Regression Problem

Ex (continued):

CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

$$\Sigma x = 314+317+\ldots+369 = 3377$$

$$\Sigma x^2 = (314)^2+(317)^2+\ldots+(369)^2 = 1143757$$

$$\Sigma y = 13.9+14+\ldots+14.4 = 141.7$$

$$\Sigma y^2 = (13.9)^2+(14)^2+\ldots+(14.4)^2 = 2008.39$$

$$\Sigma xy = (314)(13.9)+(317)(14)+\ldots+(369)(14.4) = 47888.6$$

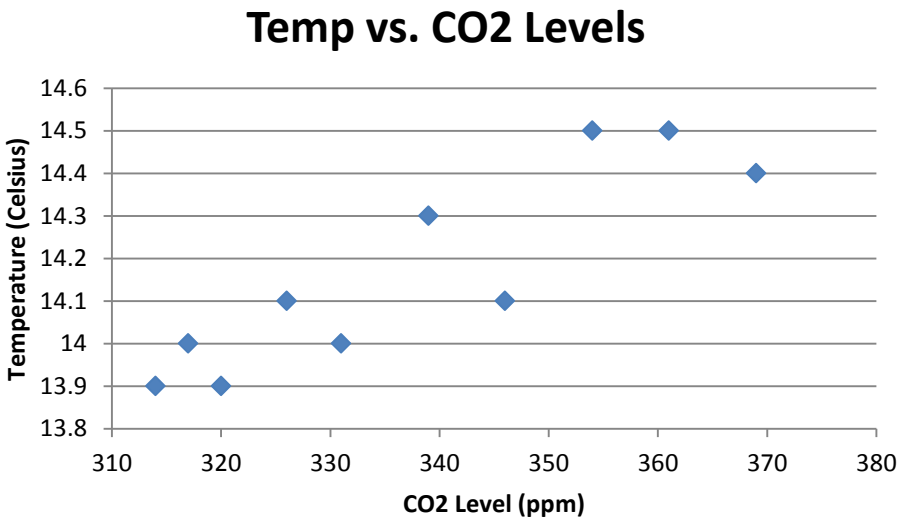
$$\bar{x} = \frac{\Sigma x}{n} = \frac{3377}{10} = 337.7$$

A Detailed Linear Regression Problem

Ex (continued):

CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

Before we calculate r , here is a graph of the data



What do you think r is?

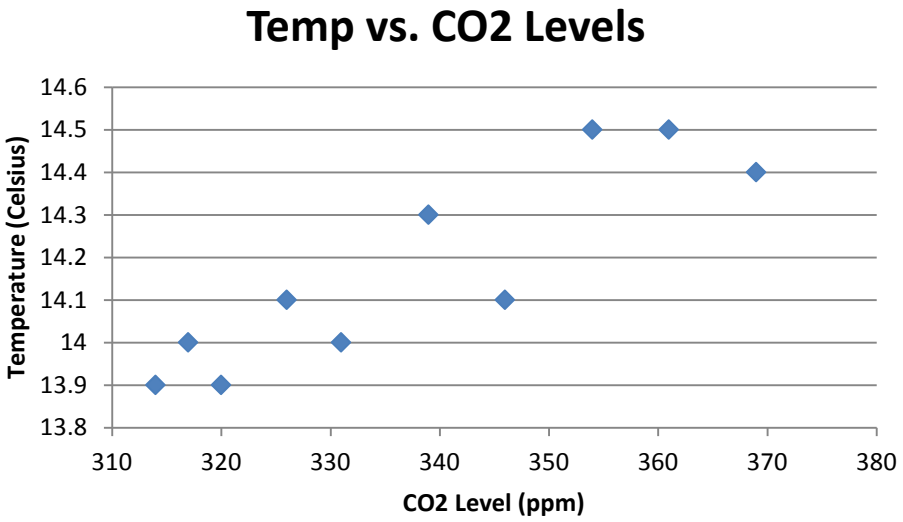
r is positive when x ↑ , y ↑

r is close to 1

A Detailed Linear Regression Problem

Ex (continued):

CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4



a) To calculate r , plug into the formula for r

$$r = 0.8920$$

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n\left(\sum x^2\right) - (\sum x)^2}\sqrt{n\left(\sum y^2\right) - (\sum y)^2}}$$

$$= \frac{(10)(47888.6) - (3377)(141.7)}{\sqrt{(10)(1143757) - (3377)^2}\sqrt{10(2008.39) - (141.7)^2}}$$

A Detailed Linear Regression Problem

Ex (continued):

CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

b) To find the equation of the regression line, plug into the equations for b_1 and b_0

$$b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(10)(47888.6) - (3377)(141.7)}{(10)(1143757) - (3377)^2}$$

$$= 0.0109$$

$$b_0 = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$= \frac{(141.7)(1143757) - (3377)(47888.6)}{(10)(1143757) - (3377)^2} = 10.483$$

$\hat{y} = b_1 x + b_0$ so $\hat{y} = 0.0109x + 10.483$

A Detailed Linear Regression Problem

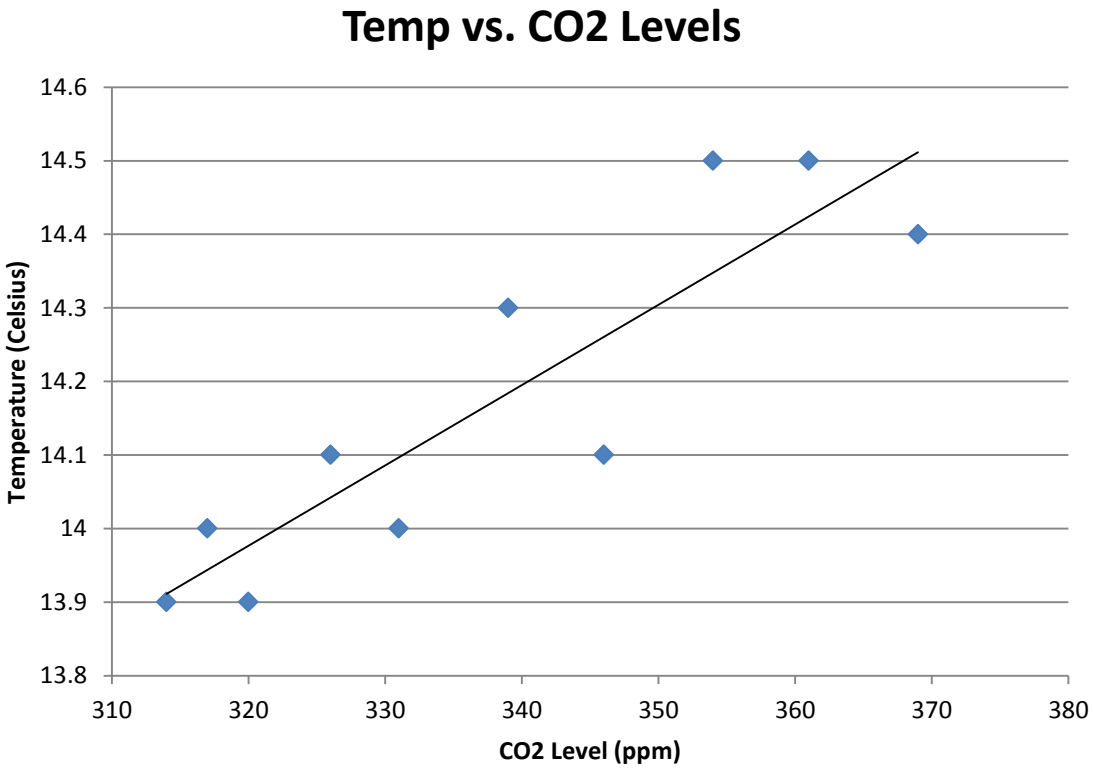
Ex (continued):

CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

a) $r = 0.8920$

b) $\hat{y} = 0.0109x + 10.483$

Here's what's going on



A Detailed Linear Regression Problem

Ex (continued):

CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

c) The whole point of the regression line $\hat{y} = 0.0109x + 10.483$ is to make predictions. To make the prediction, just plug in the given value of *x* to predict the *y* value

When the Earth's CO2 level reaches 400 ppm (that is called X_0), the best one number point estimate prediction for *y* is

$$\hat{y} = 0.0109(400) + 10.483 = 14.843\,^{o}C$$

A Detailed Linear Regression Problem

Ex (continued):

CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

a) $r = 0.8920$ b) $\hat{y} = 0.0109x + 10.483$ c) $\hat{y} = 14.843\,^{o}C$

d) As usual, to find a prediction interval, we need to find E. To find E we need to find $t_{\alpha/2}$, s_e, and plug everything into the formula for E.

$$\alpha = 1 - conf.\,level$$

$$= 1 - 0.90 = 0.10$$

$$t_{\alpha/2} = 1.860 \quad \text{(from table)}$$

$$\alpha/2 = 0.10/2 = 0.05$$

$$df = n - 2 = 10 - 2 = 8$$

A Detailed Linear Regression Problem

Ex (continued):

CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

a) $r = 0.8920$ b) $\hat{y} = 0.0109x + 10.483$ c) $\hat{y} = 14.843\,^{o}C$

d) $t_{\alpha/2} = 1.860$ $s_e = ?$ E = ?

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

$$= \sqrt{\frac{2008.39 - (10.483)(141.7) - (0.0109)(47888.6)}{10-2}}$$

$$= 0.347$$

A Detailed Linear Regression Problem

Ex (continued):

CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

a) $r = 0.8920$ b) $\hat{y} = 0.0109x + 10.483$ c) $\hat{y} = 14.843\,^{o}C$

d) $t_{\alpha/2} = 1.860$ $s_e = 0.347$ E = ?

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}}$$

$$= (1.860)(0.347)\sqrt{1 + \frac{1}{10} + \frac{10(400 - 337.7)^2}{10(1143757) - (3377)^2}}$$

$$= 0.970$$

A Detailed Linear Regression Problem

Ex (continued):

CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

a) $r = 0.8920$ b) $\hat{y} = 0.0109x + 10.483$ c) $\hat{y} = 14.843\,^{\circ}C$

d) $t_{\alpha/2} = 1.860$ $s_e = 0.347$ $E = 0.970$

… and the 90% prediction interval is

$$\hat{y} - E < y < \hat{y} + E$$

$$14.843 - 0.970 < y < 14.843 + 0.970$$

$$13.873\,^{\circ}C < y < 15.813\,^{\circ}C$$

A Detailed Linear Regression Problem

Ex (continued):

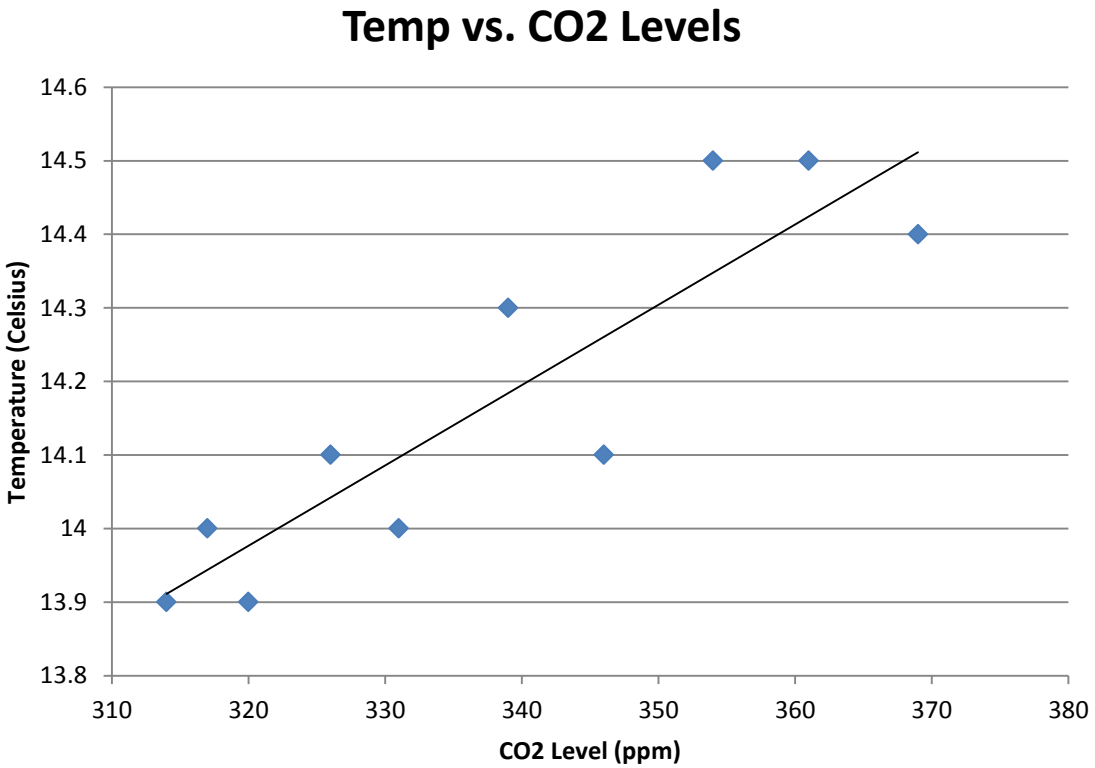
CO2 x	314	317	320	326	331	339	346	354	361	369
Temperature y	13.9	14	13.9	14.1	14	14.3	14.1	14.5	14.5	14.4

a) $r = 0.8920$

b) $\hat{y} = 0.0109x + 10.483$

c) $\hat{y} = 14.843\,^{\circ}C$

d) $13.873\,^{\circ}C < y < 15.813\,^{\circ}C$



Homework

Homework

Sec. 4.1: #'s 1-43 odd

Sec. 4.2: #'s 1-31 odd

Sec. 14.2: #'s 1-17 all

Thanks everyone, it’s been a fun semester… ☺ ☹