Chapter 14: Linear Regression

Note Title 4/27/2017

(All of Chapter 14 combined)

14.1: Linear Equations

Slope-intercept form for a linear equation:

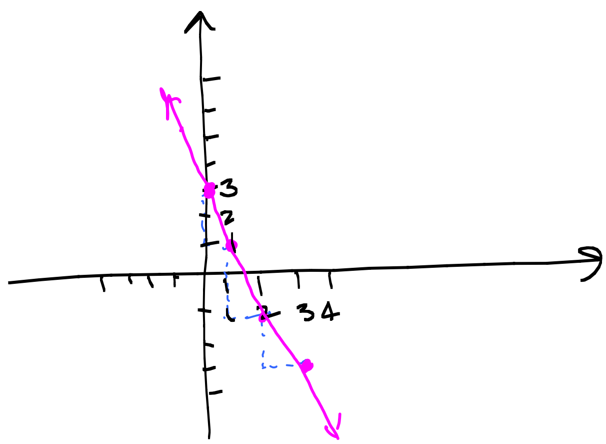
$$y = mx + b$$

$$m = \text{slope} = \frac{y_2 - y_1}{x_2 - x_1} = \text{"}\frac{\text{rise}}{\text{run}}\text{"}$$

b = y-intercept (where it intersects the y-axis)

Ex: Graph the line

$$y = -2x + 3$$



Slope: $m = -2 = -\frac{2}{1}$

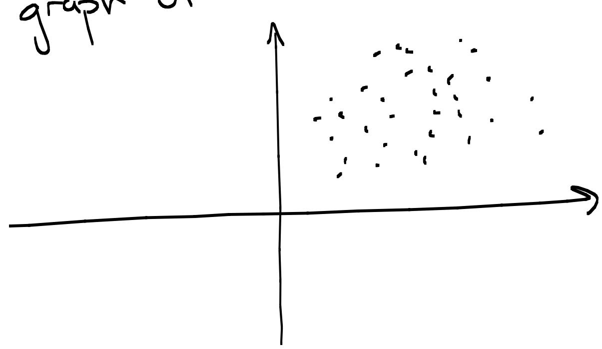
y intercept: 3 or $(0, 3)$

slope negative: $\searrow$

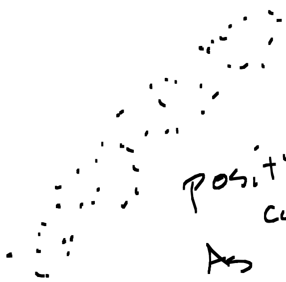
(positive slopes go uphill from left to right)

$\nearrow +$

Scatter Plot: graph of observed ordered pairs (x, y)



If there is a discernible pattern in the scatter plot, x can be used to predict y.

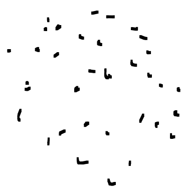


positive linear correlation

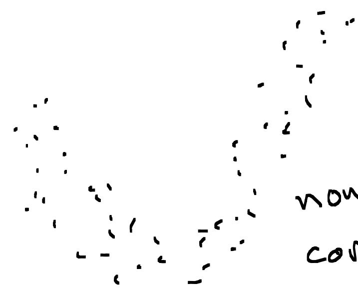
As $x\uparrow$, $y\uparrow$

negative linear correlation

As $x\uparrow$, $y\downarrow$



no correlation



nonlinear correlation

We will focus on data that has a linear correlation. We want to find a line that represents the data pattern.

In statistics, we use b_1 and b_0 instead of m and b.

$$y = b_1 x + b_0$$

slope: b_1

y-intercept: b_0

x is the <u>predictor</u> (<u>independent</u>) variable.

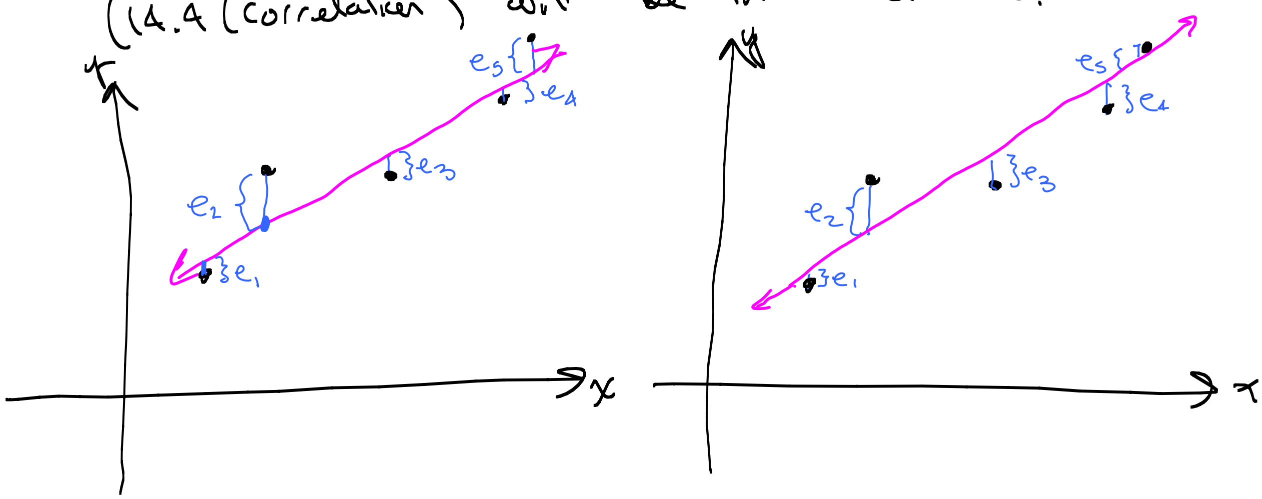
y is the <u>response</u> (<u>dependent</u>) variable

Why b_1 and b_0? Linear regression can be done with multiple predictor (independent) variables.

then you have: $y = b_4 x_4 + b_3 x_3 + b_2 x_2 + b_1 x_1 + b_0$

14.2: The Regression Equation

(14.4 (correlation) will be in here too.



Which line is better?

e_i: error = the signed vertical distance between data point (x_i, y_i) and the line.

The "best" line is the line that minimizes the squares of the errors.

This "best" line is called the <u>Least Squares Regression Line</u>.

$\hat{y}$ = predicted value of y

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

Actual y: $y = b_1 x + b_0 + e$ ← error term

We use $\hat{y}$ to distinguish it from the actual y's that are from the ordered pairs in our data set.

Example

suppose $x =$ # hours studied, $y =$ grade on exam, for this data set.

x	y
10	96
3	64
6	80
7	76
8	92
4	60

	x	y	xy	x^2	y^2
	10	96	960	100	9216
	3	64	192	9	4096
	6	80	480	36	6400
	7	76	532	49	5776
	8	92	736	64	8464
	4	60	240	16	3600
Sums	38	468	3140	274	37552

Σ means "sum" (sigma notation)

Conceptual Formulas

$$S_{xy} = \Sigma (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \Sigma (x_i - \bar{x})^2$$

$$S_{yy} = \Sigma (y_i - \bar{y})^2$$

Computational Formulas

$$S_{xy} = \Sigma (x_i y_i) - \frac{\Sigma x_i \Sigma y_i}{n}$$

$$S_{xx} = \Sigma (x_i^2) - \frac{(\Sigma x_i)^2}{n}$$

$$S_{yy} = \Sigma (y_i)^2 - \frac{(\Sigma y_i)^2}{n}$$

Slope of regression line: $b_1 = \dfrac{S_{xy}}{S_{xx}}$

Regression line is: $\bar{y} = b_1 \bar{x} + b_0$ use this to get b_0

Correlation Coefficient: (Section 14.4) $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

Correlation coefficient r (sometimes just called the correlation) is always between -1 and 1. It tells us how well the x's predict the y's and the direction of the relationship.

x	y	xy	x^2	y^2
10	96	960	100	9216
3	64	192	9	4096
6	80	480	36	6400
7	76	532	49	5776
8	92	736	64	8464
4	60	240	16	3600
Sums 38 Σx_i	468 Σy_i	3140 $\Sigma x_i y_i$	274 Σx_i^2	37552 Σy_i^2

note: n = 6

$$S_{xy} = \Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n} = 3140 - \frac{38(468)}{6} = 176$$

$$S_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = 274 - \frac{(38)^2}{6} = 33.3333$$

$$S_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} = 37552 - \frac{(468)^2}{6} = 1048$$

<u>Slope of Regression line:</u>

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{176}{33.3333} = 5.28$$

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{38}{6} = 6.3333 \quad , \bar{y} = \frac{\Sigma y_i}{n} = \frac{468}{6} = 78$$

Regression line: $\bar{y} = b_1\bar{x} + b_0$

$b_0 + 5.28(6.3333) = 78$

$78 = 5.28(6.3333) + b_0$

$b_0 = 78 - 5.28(6.3333) = 44.56$

next page

Regression Line:

$$y = b_1 x + b_0 \text{ , so}$$

$$\boxed{y = 5.28x + 44.56}$$

Use the regression line to predict values of y for x's between the boundaries of the original x-values

(so we can use it for prediction between 3 and 10)

Correlation coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{176}{\sqrt{33.33(1048)}} = 0.942$$

(a strong positive correlation)

Note Title 2/5/2015