4.1 and 4.2: Correlation and Linear Regression

Review: Linear equations

Slope-intercept form: $y = m x + b$

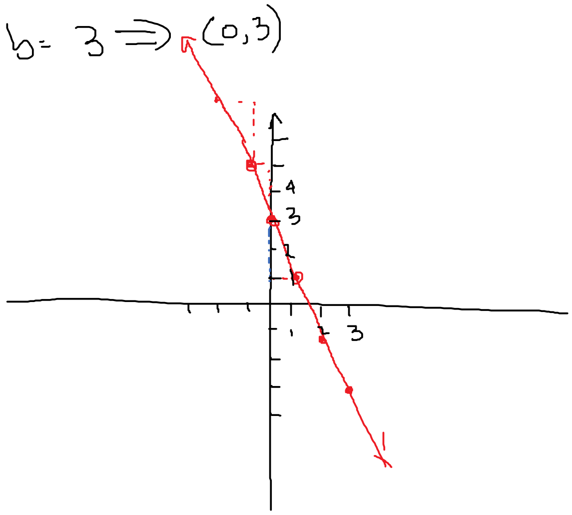
$m = \text{slope} = \dfrac{y_2 - y_1}{x_2 - x_1} = \dfrac{\text{"rise"}}{\text{run}}$

$b = \text{y-intercept}$

Ex: Graph the line $y = -2x + 3$

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

y-intercept: $b = 3 \Rightarrow (0, 3)$



Positive slopes:

Negative Slopes:

In statistics, we usually 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 predictor (independent) variable

y is the response (dependent) variable

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

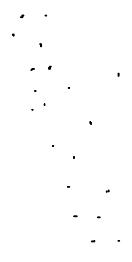
Then you have:

$y = b_3 x_3 + b_2 x_2 + b_1 x_1 + b_0$

<u>Scatter plot</u>: graph of observed points (x, y)

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

positive linear correlation
as $x \uparrow$, $y \uparrow$ also



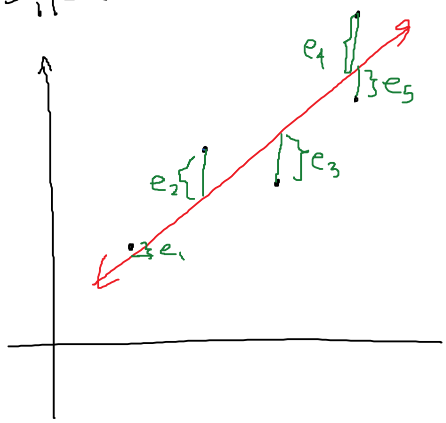
negative linear correlation
as $x \uparrow$, $y \downarrow$

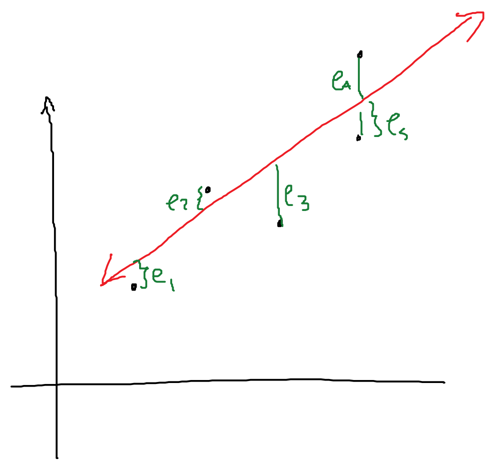
non-linear correlation

no correlation

The Regression Equation

Suppose we draw a line through a scatterplot





Which line is "better"?

e: error = the signed vertical distance between the data point and the line.

The "best" line: we choose the line that minimizes the sum of the squares of the errors.

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

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

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

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

Example: Suppose x = # hours studied, y = exam grade. for this data set

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

Conceptual Formulas

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

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

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

Computational Formulas

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

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

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

Slope of Regression line: $b_1 = \frac{S_{xy}}{S_{xx}}$

Regression line is: $\bar{y} = b_1 \bar{x} + b_0$ (goes through $(\bar{x}, \bar{y})$). use this to get b_0

Correlation Coefficient: $r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

The 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.

1 and -1: perfect prediction (all points on line)

1: positive correlation. $x \uparrow \Rightarrow y \uparrow$

-1: negative correlation $x \uparrow \Rightarrow y \downarrow$

0: no prediction.

Back to example:

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	37 552
$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum y^2$

$n=6$

Computational Formulas

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

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

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

$$S_{xy} = \sum(xy) - \frac{\sum x \sum y}{n} = 3140 - \frac{38(468)}{6} = 176$$

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

$$S_{yy} = \sum(y^2) - \frac{(\sum y)^2}{n} = 37\,552 - \frac{(468)^2}{6} = 1048$$

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

Slope of Regression Line

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

Regression Line: $\bar{y} = b_1 \bar{x} + b_0$ (so $(\bar{x}, \bar{y})$ is on line)

$$78 = 5.28(6.3333) + b_0$$

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

Regression line: $y = b_1 x + b_0$, so $\boxed{y = 5.28x + 44.56}$

positive slope.

Calculate correlation coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{176}{\sqrt{33.3333(1048)}}$$

$$\boxed{r = 0.942}$$ correlation

strong positive correlation

$r^2 = (0.942)^2 = 0.88672 \Rightarrow$ this means that 88.7% of the difference in y is explained by x