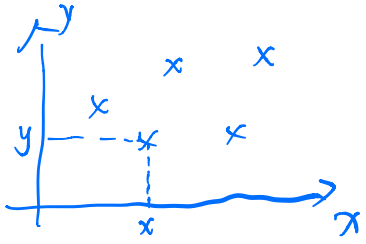
Correlation and Regression

2 Group of Data, X and Y, interested in relationship between them.

1. Scatter Plot.

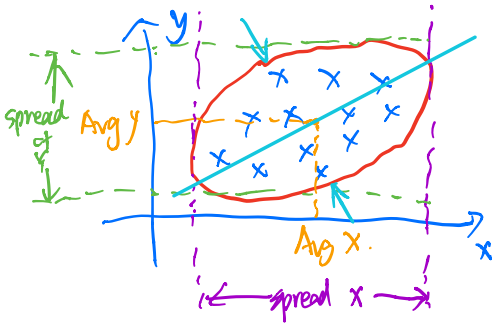
A scatterplot or scatter diagram is a 2 dimensional plot of data. The horizontal dimension is called x, and the vertical dimension is called Y.

Each point on a scatterplot shows two values, x value and Y value.



Scatter plot only shows association, but association does not mean causation.

Summarize a scatter plot.



Football shape, consider the center and spread. Spread has two types: spread on x axis and spread on y axis. Association.

mean of X, mean of Y gives the center.

SD of x gives spread on x, SD of Y gives spread on Y.

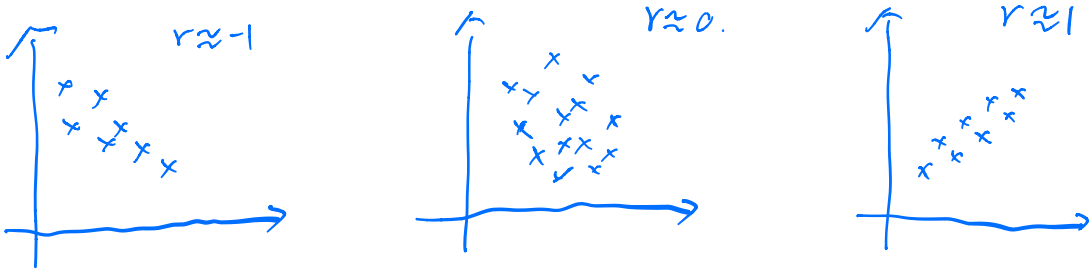
Strength of association is measured by correlation coefficient r.

Properties:

1. $-1 \le r \le 1$.

2. r is close to 1 or -1, the data are close to a line.

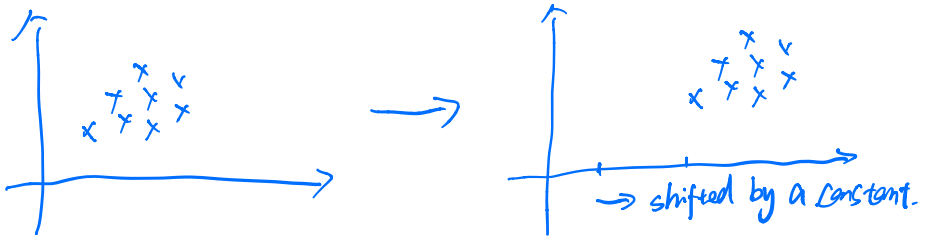
   r is close to 0, the data are not close to a line.



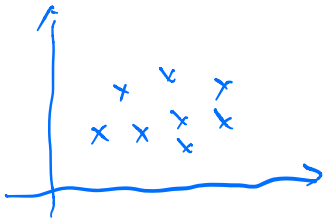
Similar pictures can be found in textbook page: 127, 129, 142.

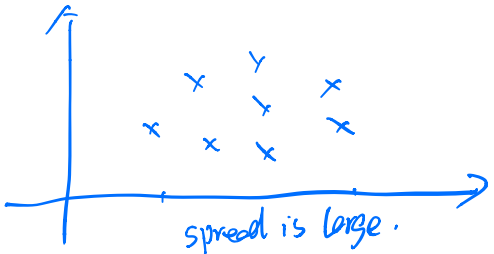
3. The correlation between x and y is the same as correlation between y and x

4. Invariant under addition. Add a constant to every one of x or y doesn't change r.



5. Invariant under multiplication: if all of the x or y value is multiplied by a constant, r doesn't change.





spread is large.
but closeness to a line remain the same.

Computation: $r = \frac{\text{cov}(x,y)}{(\text{SD of } x) \times (\text{SD of } y)}$ (The textbook has another way on page 132-134)

$\text{cov}(x,y) = (\text{avg of } xy) - (\text{avg of } x) \times (\text{avg of } y)$.

Example: (Same as Example 1 on page 132)

x:	1	3	4	5	7
y:	5	9	7	1	13

avg of x: $\frac{1+3+4+5+7}{5} = 4$.

avg of y: $\frac{5+9+7+1+13}{5} = 7$

SD of x: $\sqrt{\frac{(1-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(7-4)^2}{5}}$

$= 2$.

SD of y: $\sqrt{\frac{(5-7)^2+(9-7)^2+(7-7)^2+(1-7)^2+(13-7)^2}{5}}$

$= 4$.

avg of xy: $\frac{1\times5+3\times9+4\times7+5\times1+7\times13}{5}$

$= \frac{5+27+28+5+91}{5} = 31.2$.

$r = \frac{31.2-4\times7}{2\times4} = \frac{3.2}{8} = 0.4$.

Regression:

The regression method describes how one variable depends on another.

The regression line for y on x estimates the average value for y corresponding to each value of x.

In a regression problem, we are interested in predict the value of y^* for a given x^*.

In this kind of problems, you will usually be given avg of x, avg of y, SD of x, SD of y and r, as well as a new value of x, call it x^*, then you are asked to predict the coresponding y^*. There are two ways to do it.

1. Matching Z-score.

Step 1: compute Z score of x^* by: $Z_x = \frac{x^* - \text{Avg } x}{SD_x}$.

Step 2: compute Z score of corresponding y^*: $Z_y = r \times Z_x$.

Step 3: prediction value: $y^* = \text{Avg of } y + Z_y \times SD_y$.

(You compute Z score, but you don't need Z table in this case)

Example: Age (x): 20, 21, 18, 22, 24. Income (y): 22, 20, 18, 14, 26. $/h.

Avg x: $\frac{20+21+18+22+24}{5} = 21$. SD of x: $\sqrt{\frac{(20-21)^2+(14-21)^2+(18-21)^2+(22-21)^2+(24-21)^2}{5}} = 2$.

Avg y: $\frac{22+20+18+14+26}{5} = 20$ SD of Y: $\sqrt{\frac{(22-20)^2+(20-20)^2+(18-20)^2+(14-20)^2+(26-20)^2}{5}} = 4$

Avg. xy: $\frac{20\times22+21\times20+18\times18+22\times14+24\times26}{5} = 423.2$.

$r = \frac{423.2-21\times20}{2\times4} = \frac{3.2}{8} = 0.4$

what is your best predict income if age is 25?

Step1: $Z_x = \frac{25-21}{2} = \frac{4}{2} = 2$.

Step2: $Z_y = 0.4\times2 = 0.8$.

Step3: $y^* = 20+0.8\times4 = 23.2$. predict Y^* as 23.2 $/h.

Simpler formula. $y^* =$ Avg of $Y + Z_Y \times SD_Y$

$=$ Avg of $Y + r\times Z_x \times SD_Y$

$=$ Avg of $Y + r\times\left(\frac{x^*-\text{Avg of } x}{SD_x}\right)\times SD_Y$

$=$ Avg of $Y + r\times\frac{SD_x}{SD_x}\times(x^*-\text{Avg of } x)$.

using this formula: $y^* = 20 + 0.4\times\frac{4}{2}\times(25-21) = 20+0.4\times2\times4 = 23.2$ $/h.

Method 2: use regression line.

$y^* =$ slope $\times x^*$ + intercept.

Step1: Slope of the line: slope $= r\times\frac{SD_Y}{SD_x}$.

Step 2: intercept of the line: intercept = Avg of y − slope × Avg of x.

Step 3: plug in $y^* =$ slope $\times x^*$ + intercept.

Example: same question as before.

Slope: $r \times \frac{SD_y}{SD_x} = 0.4 \times \frac{4}{2} = 0.8$.

intercept: Avg y − slope × Avg of x = 20 − 0.8 × 21 = 3.2.

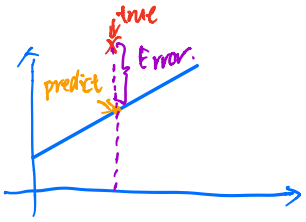
$y^* = 0.8 \times 25 + 3.2 = 23.2$. Connections between 2 methods:

y^* = slope × x^* + Avg of y − slope × Avg of x

= Avg of y + slope × (x^* − Avg of x)

= Avg of y + $r \times \frac{SD_y}{SD_x}$ × (x^* − Avg of x).

How to measure the quality of your prediction? use R.M.S. error



R.M.S error = $\sqrt{\frac{(\text{error \#1})^2 + (\text{error \#2})^2 + \cdots + (\text{error \#n})^2}{n}}$

= $\sqrt{1-r^2} \times$ SD of y.

Example: Same data as before.

regression line: $y = 0.8x + 3.2$.

x	20	21	18	22	24
y	22	20	18	14	26
predict	19.2	20	17.6	20.8	22.4

R.M.S. error = $\sqrt{\frac{(22-19.2)^2 + (20-20)^2 + (18-17.6)^2 + (14-20.8)^2 + (26-22.4)^2}{5}}$

= $\sqrt{\frac{67.2}{5}} \approx 3.666$

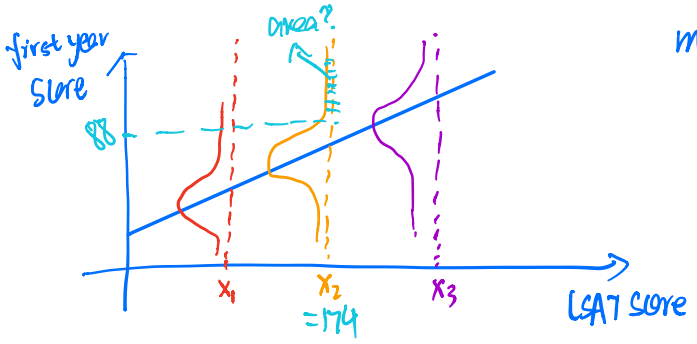
R.M.S. error $= \sqrt{1-0.4^2} \times 4 \approx 3.666$.

Another type of question: about percentage.

Example: Average LSAT score = 162, SD = 6.

Average first year score = 68, SD = 10, $r = 0.60$

Q: Among the student who scored 174 on LSAT, about what percentage had first year scores over 88?



mean, SD of new normal curve?

1. Step 1: slope, intercept, RMS error.

slope: $r \times \frac{SD_y}{SD_x} = 0.6 \times \frac{10}{6} = 1$.

intercept: Avg Y − slope × Avg $X = 68 - 1 \times 162 = -94$.

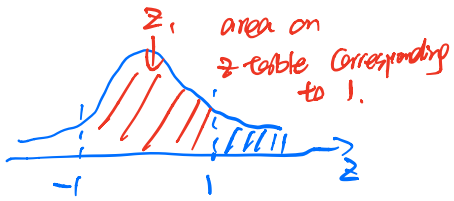
RMS error: $\sqrt{1-r^2} \times SD_y = \sqrt{1-0.6^2} \times 10 = 8$

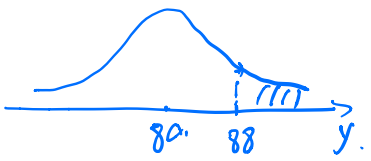
2. new mean: slope × x^* + intercept $= 1 \times 174 - 94 = 80$.

3. new SD: just RMS error, 8.

4. new z score: $\frac{88-80}{8}$ unit in z, 1 in z.

5. area: $\frac{50-\frac{z_1}{2}}{2} = \frac{50-\frac{68.27}{2}}{2} = 7.93\%$





Shadow area are the same.