STAT22000 Autumn 2013 Lecture 4

Yibi Huang

October 7, 2013

Scatterplots

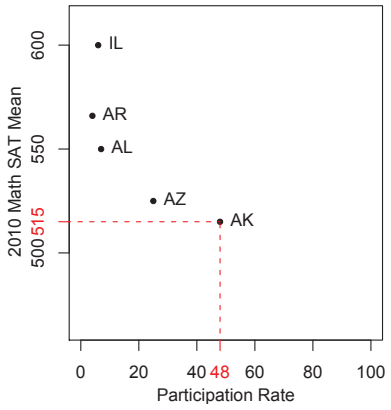
(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
$\vdots$
(x_n, y_n)

A **scatter plot** shows the relationship between two numerical variables measured on the same cases (e.g. same person, same company, etc).

The values of one variable are on the x-axis, and the values of the other are on the y-axis. Each individual is represented by a point in the graph.

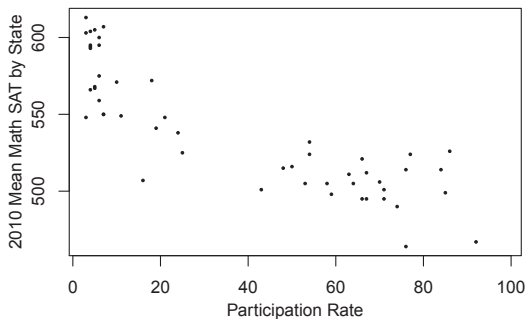
2010 Mean Math SAT Scores[a]

State	Participation Rate[b]	Math
AK	48	515
AL	7	550
AR	4	566
AZ	25	525
CA	50	516
CO	18	572
CT	84	514
DC	76	464
DE	71	495
FL	59	498
⋮	⋮	⋮
IL	6	600
⋮	⋮	⋮
WY	5	567



[a] Source: CollegeBoard
[b] The percentage of high school graduates in the class of 2010 who took the SAT

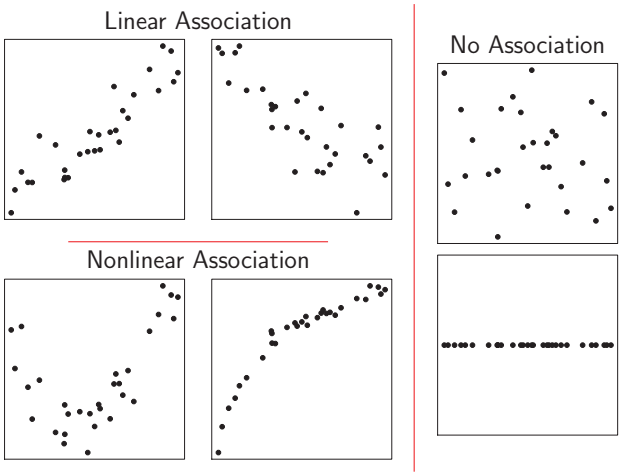

```
> A = read.table("SAT2010.csv",h=T)
> A[1:2,]                 # to show the first two rows of the data
     State Abbr Region Percent X2010R X2010M X2010W
1   Alaska   AK      U      48    518    515    491
2  Alabama   AL     SE       7    556    550    544
> plot(A$Percent,A$X2010M,xlab="Participation Rate",
  ylab="2010 Mean Math SAT by State")
```



What to Look in a Scatter Plot?

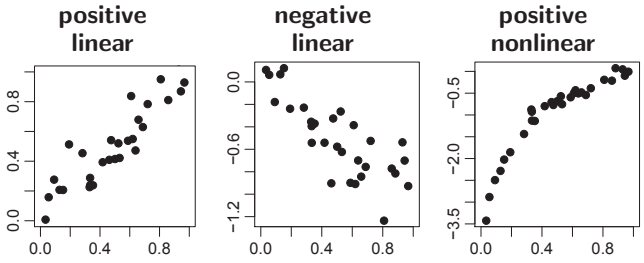
1. What is the overall pattern?
   - What is the *form* of the relationship? (linear, curved, clustered ...)
   - What is the *direction* of the relationship? (positive association, negative association)
   - What is the *strength* of the relationship? (strong, weak, ...)
2. Are there any deviations from the overall pattern? An point that falls outside the overall pattern is an **outlier**.

Form of an Association



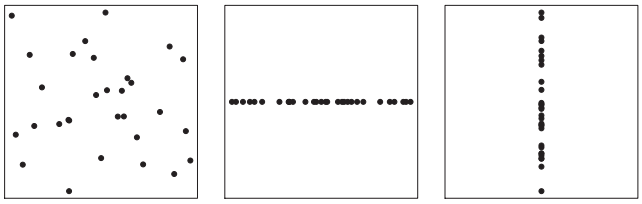
Positive and Negative Association

Positive / Negative association: High values of one variable tend to occur together with high / low values of the other variable.



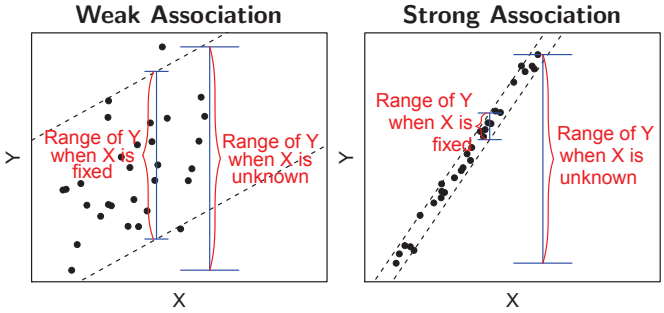
No Association

No relationship: X and Y vary independently. Knowing X tells you nothing about Y.



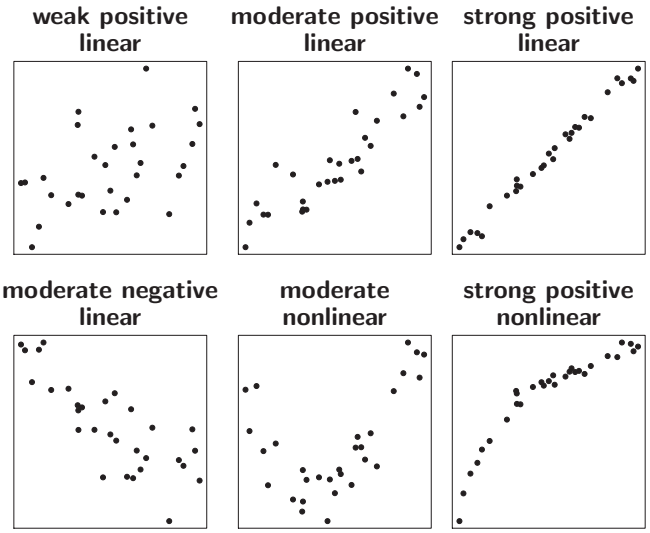
Strength of the Association (1)

The **strength** of the relationship between the two variables can be seen by how much variation, or **scatter**, there is around the main form.



Large spread of Y when X is known

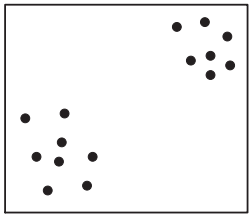
Small spread of Y when X is known



Clusters (1)

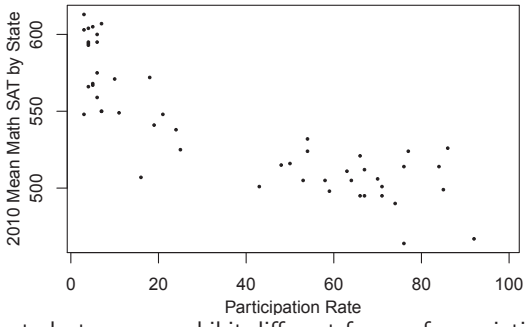
Sometimes points in a scatter plot form **more than one clusters**.

Clusters in a graph suggests that data may *describe several distinct kinds of individuals*.



Clusters (2)

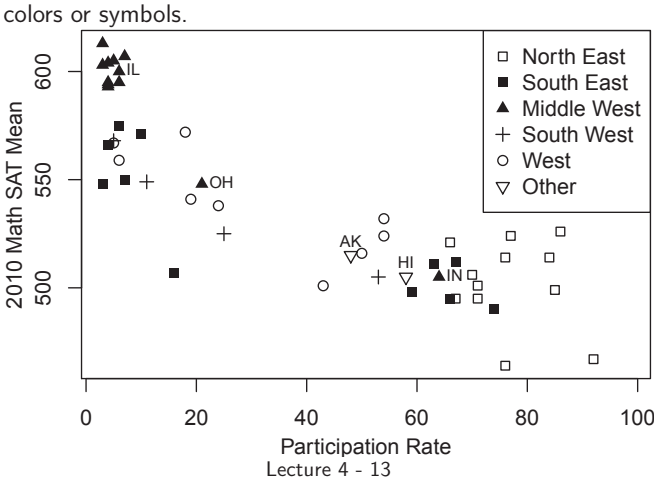
The mean MSAT score and participation rate scatter plot contains at least two clusters.



Different clusters may exhibit different forms of association. E.g., the left cluster above shows a moderate negative association, while the right cluster shows no (or a very weak) association.

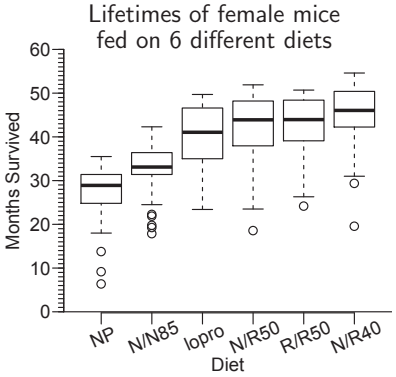
Adding Categorical Variables to Scatter Plots

Points in different categories can be marked with different colors or symbols.



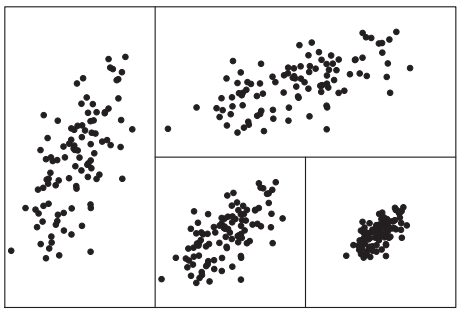
If One of the Variables Is Categorical...

The two variables in a scatter plot have to be *both numerical*. If one is categorial, and the other is numerical, one can use a **side-by-side box plot** to examine their relation.



Be careful in your interpretation: This is NOT a positive association, because diet is not numerical.

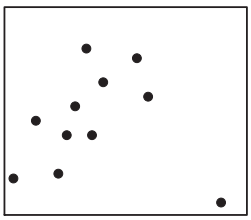
Which of the 4 scatter plots below shows the strongest association? The weakest?



While judging the strength of associations, our eyes can be fooled by changing the plotting scales or the space around the cloud of points. We need a more objective approach.

Outliers

In a scatterplot, **outliers** are points that fall outside of the overall pattern of the relationship.



2.2 Correlation

Outline

- The correlation coefficient "r"
- r does not distinguish between x and y
- r has no units of measurement
- r ranges from -1 to $+1$
- Influential points

The Correlation Coefficient "r"

Recall the sample SD of variables X and Y respectively are

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

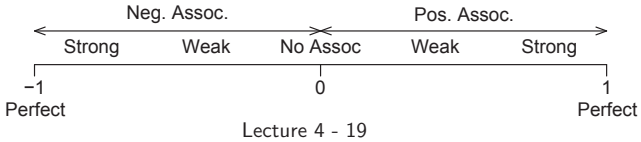
The **correlation coefficient** r (or simply, **correlation**) is defined as:

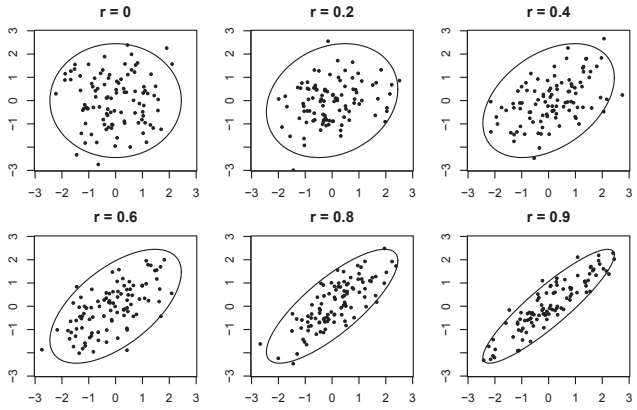
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\underbrace{\left(\frac{x_i - \bar{x}}{s_x}\right)}_{z\text{-score of } x_i}\underbrace{\left(\frac{y_i - \bar{y}}{s_y}\right)}_{z\text{-score of } y_i}.$$

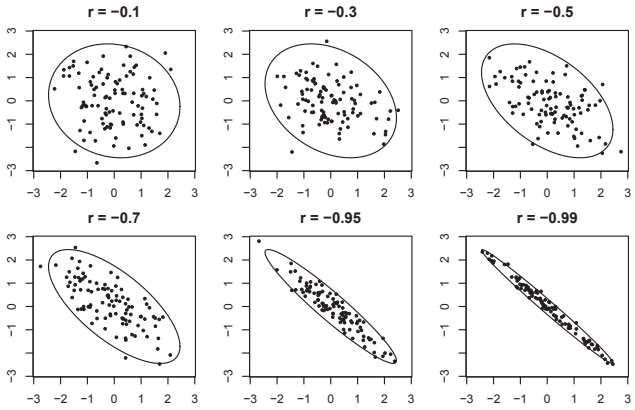
Correlation r Ranges From -1 to $+1$

Correlation r is a measure of the *direction* and *strength* of the **linear** relationship between two quantitative variables. "r" always lies between -1 and 1; the strength increases as you move away from 0 to either -1 or 1.

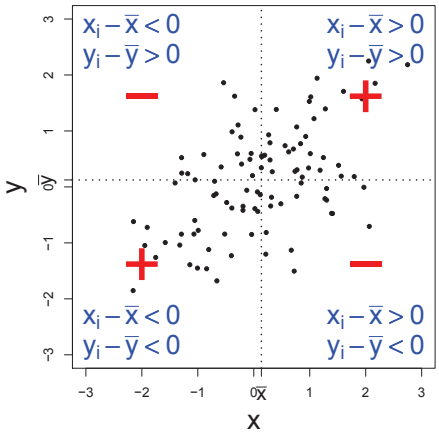
- $r > 0$: positive association
- $r < 0$: negative association
- $r \approx 0$: very weak linear relationship
- large $|r|$: strong linear relationship
- $r = -1$ or $r = 1$: *only* when all the data points on the scatterplot lie exactly along a straight line







Why r Measures the Strength of a Linear Relationship?



What is the sign of $\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$??

Here $r > 0$; more positive contributions than negative.

Correlation r Has No Unit

$$r = \frac{1}{n-1}\sum_{i=1}^{n} \underbrace{\left(\frac{x_i - \bar{x}}{s_x}\right)}_{z\text{-score of } x_i} \underbrace{\left(\frac{y_i - \bar{y}}{s_y}\right)}_{z\text{-score of } y_i}.$$

Recall that after standardization, the z-score of neither x_i nor y_i has unit.

- So r is unit free.
- So we can compare r between data sets, where variables are measured in different units or when variables are different. E.g. we may compare the

  r between [swim time and pulse],

  with the

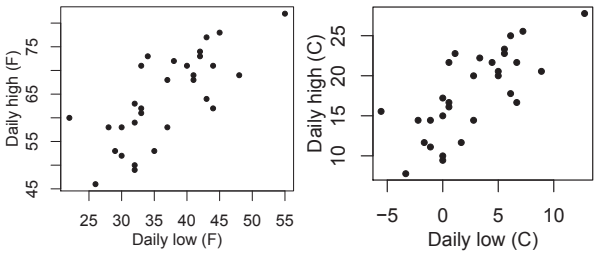
  r between [swim time and breathing rate].

Correlation r Has No Unit (2)

Changing the units of variables does not change the correlation coefficient r, because we get rid of all our units when we standardize (get z-scores).

E.g., no matter the temperatures are recorded in $^\circ F$, or $^\circ C$, the correlations are both $r = 0.74$ because

$$C = \frac{5}{9}(F - 32).$$

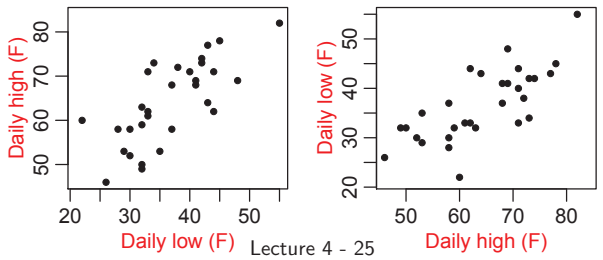


"r" Does Not Distinguish x & y

Sometimes one use the X variable to predict the Y variable. In this case, X is called the *explanatory variable*, and Y the *response*. The correlation coefficient r does not distinguish between the two. It treats x and y symmetrically.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Swapping the x-, y-axes doesn't change r (both $r = 0.74$.)



Linear Transformation of x & y Does Not Change "r"

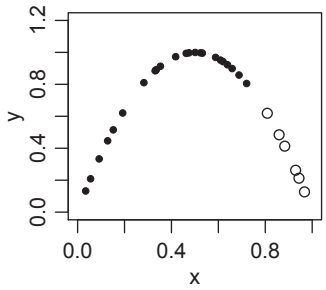
Correlation does not depend on the units of measurement. Translation and scaling have no effect on the correlation. For $x'_i = a + b\,x_i$ and $y'_i = c + d\,y_i$, $i = 1, 2, \ldots, n$

$$r_{x'y'} = \text{sign}(cd) \times r_{xy}$$

In other words, $r_{x'y'} = \pm r_{xy}$, linear transformation does not change the absolute value of correlation.

Correlation r Describes Linear Relationships Only

The scatter plot below shows a *perfect nonlinear* association. All points fall on the quadratic curve $y = 1 - 4(x - 0.5)^2$.

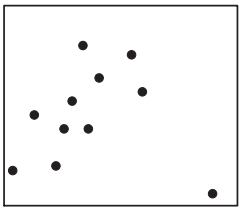


r of all black dots $= 0.803$,
r of all dots $= -0.019$.
(black + white)

No matter how strong the association,
the r of a curved relationship is NEVER 1 or -1.
It can even be 0, like the plot above.

Correlation Is VERY Sensitive to Outliers

Sometimes a single outlier can change r drastically.



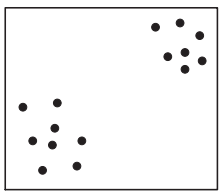
For the plot on the left,

$$r = \begin{cases} 0.0031 & \text{with the outlier} \\ 0.6895 & \text{without the outlier} \end{cases}$$

Outliers that may remarkably change the form of associations when removed are called **influential points**.

Remark: Not all outliers are influential points.

When Data Points Are Clustered ...



In the plot above, each of the two clusters exhibits a weak negative association ($r = -0.336$ and -0.323).
But the whole diagram shows a moderately strong positive association ($r = 0.849$).

- This is an example of the **Simpson's paradox**.

An overall r can be misleading when data points are clustered. Cluster-wise r's should be reported as well.

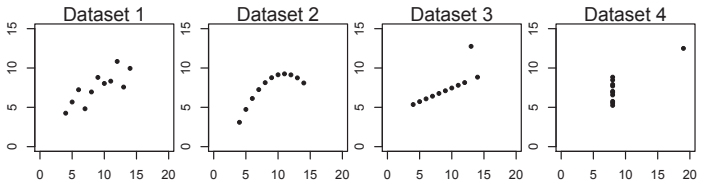
Always Check the Scatter Plots (1)

The 4 data sets below have identical $\bar{x}$, $\bar{y}$, s_x, s_y, and r.

	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.96	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.75	13	12.76	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.10	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.10	4	5.36	19	12.50
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Ave	9	7.5	9	7.5	9	7.5	9	7.5
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

How about their scatter plots?

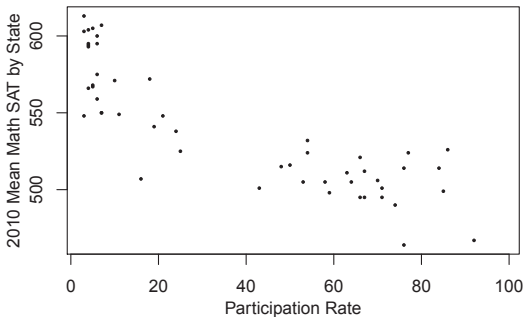
Always Check the Scatter Plots (2)



- In Dataset 2, y can be predicted exactly from x. But $r < 1$, because r only measures **linear** association.
- In Dataset 3, r would be 1 instead of 0.82 if the outlier were actually on the line.
- In Dataset 4, r would be 0 if the outlier were removed.

The correlation coefficient can be misleading in the presence of outliers, multiple clusters, or nonlinear association.

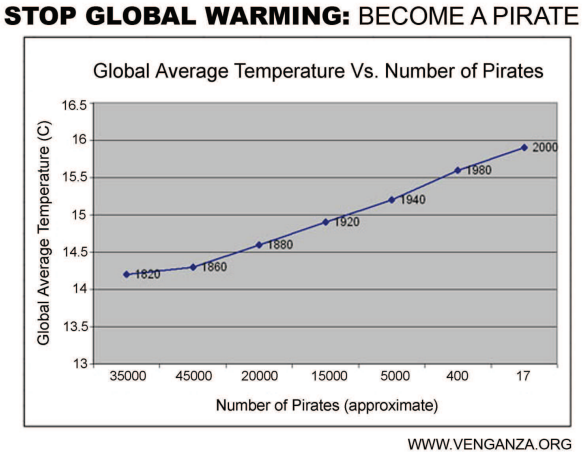
The R Command to Find r



The R-command to find the correlation of two variables: `cor()`.

```
> cor(A$Percent,A$X2010M)
[1] -0.8547424
```


Correlation Indicates Association, *Not* Causation



Brainstorm on Correlation

- Why do both variables have to be quantitative?
- Why doesn't a tight fit to a horizontal line imply a strong correlation?
- If the law requires women to marry only men 2 years older than themselves, what is the correlation of the ages between husband and wife?