AP Stats
Unit 02 – Bivariate Data
Day 2 Notes

Name Key

Correlation: measures the direction and strength of the linear relationship between 2 quantitative variables.

- standardized value, no units
- r = correlation

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

← I will not make you do this! :

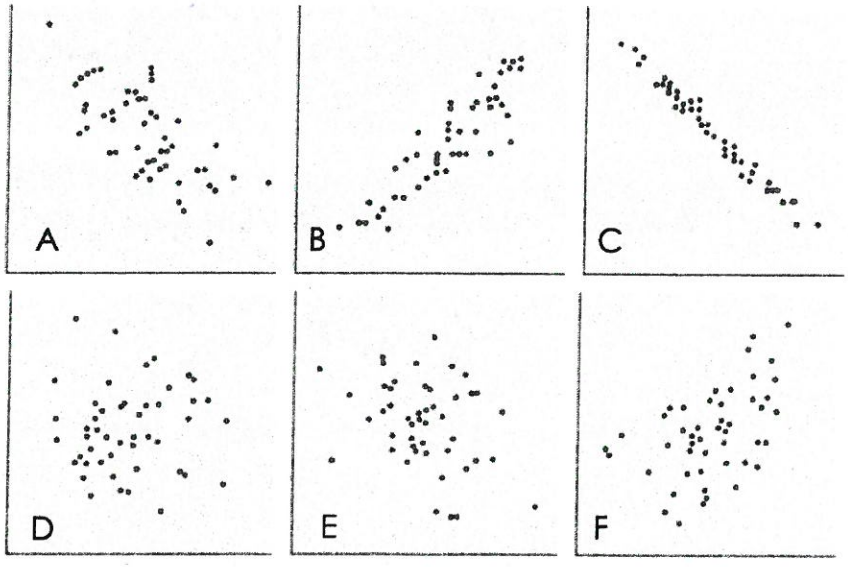
Facts about correlation:

1. makes no distinction between x and y
2. both variables must be quantitative
3. has no units
4. +r = positive correlation (slope)
   -r = negative correlation (slope)
5. always $-1 \leq r \leq 1$ with 0 = no association, +1 strong+, -1 strong-
6. only used for linear data (no curves!)
7. not resistant. can easily be affected by outliers/influential points.

Example:
Match the correlation values of r with the appropriate graph.

1. r = -0.99 C
2. r = -0.7 A
3. r = -0.3 E
4. r = 0 D
5. r = 0.5 F
6. r = 0.9 B

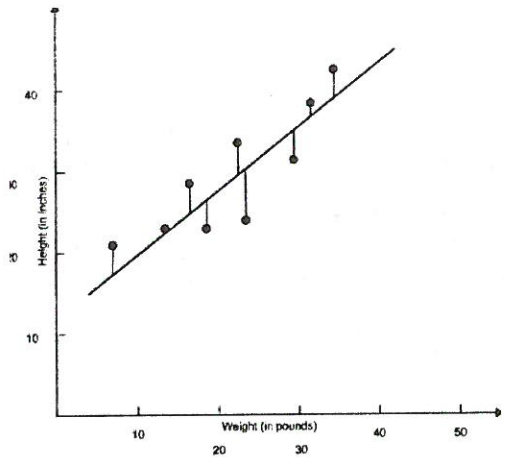
*applet



Regression Line: a straight line that describes how a response variable (y) changes as an explanatory variable (x) changes.

→ used to predict a value of y for a given value of x

→ must be created from explanatory and response variables.

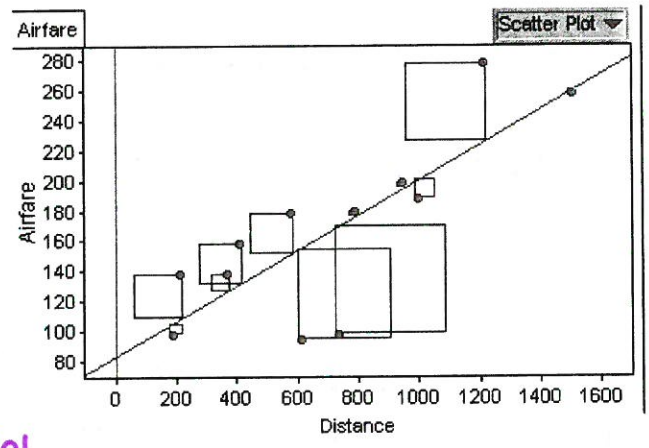


Least-Squares Regression Line (LSRL):

→ a method for finding a line through a scatterplot that summarizes the relationship between 2 variables.

→ the LSRL of y on x is the line that makes the sum of the squared residuals as small as possible.

- As the plots to the right illustrate, the LSRL makes the squares of the vertical distance as small as possible.



* mathematical model (equation)

Equation for the LSRL: $\hat{y} = a + bx$

↑ "y hat" is the predicted value of y for a given value of x.

With slope $b = r\frac{s_y}{s_x}$ and intercept $a = \bar{y} - b\bar{x}$

Where:
- r = correlation
- s_x = standard deviation of explanatory variable (x)
- s_y = standard deviation of response variable (y)
- $\bar{x}$ = mean of explanatory variable (x)
- $\bar{y}$ = mean of response variable (y)

R^2- the coefficient of determination: the percent of variation in the values of y that are accounted for by the LSRL of y on x.

- $r^2 = 1 - \frac{\Sigma \text{residuals}^2}{\Sigma (y_i - \bar{y})^2}$ or...
- literally take r and square it. ☺

Facts about the least-squares regression:

1. determine which variable is the explanatory and which is the response.

2. has a close connection between correlation and slope.

3. LSRL always passes through the point $(\bar{x}, \bar{y})$ and uses $\bar{x}$, s_x, $\bar{y}$, s_y, and r to create an equation.

4. r describes the strength of a straight line relationship.

Example:
Ninth grade students at BHS go on a backpacking trip each fall. Students are divided into hiking groups with 8 people by selecting names from a hate. Before leaving students and their backpacks were weighed. Here are the data from one hiking group in a recent year:

x	**Body**	120	187	109	103	131	165	158	116
y	**Backpack**	26	30	26	24	29	35	31	28

Find the following piece of information:

a) Correlation Coefficient (calculator) $r = 0.7947$

b) R^2 and what it means in context $R^2 = 0.6315$ or 63.15% of the variance in backpack weight is accounted for by the LSRL of backpack weight on body weight.

c) LSRL equation (calculator) $\hat{y} = 16.26 + 0.091x$

x = body weight
y = backpack weight

OR

$\widehat{\text{backpack weight}} = 16.26 + 0.091(\text{body weight})$

Example: Does Fidgeting Keep you Slim?

Some people don't gain weight even when they overeat. Perhaps fidgeting and other "non-exercise activity" (NEA) explains why- some people may spontaneously increase non-exercise activity when fed more. Researchers deliberately overfed 16 healthy young adults for 8 weeks. They measured fat gain (in kilograms) as the response variable and change in energy use (in calories) from activity other than deliberate exercise-fidgeting, daily living, and the like- as the explanatory variable. Here are the data:

NEA (cal)	-94	-57	-29	135	143	151	245	355	392	473	486	535	571	580	620	690
Fat Gain (kg)	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

a) Find the mean of the explanatory and response variables (label them $\bar{x}$ *and* $\bar{y}$)

$\bar{x} = 324.75$ cal

$\bar{y} = 2.3875$ kg

b) Find the standard deviation of the explanatory and response variables (label them s_x and s_y)

$S_x = 257.66$ cal

$S_y = 1.1389$ kg

c) The correlation r = -.78. Using the formula, create a LSRL equation.

$a = \bar{y} - b\bar{x}$ $\quad b = r\frac{S_y}{S_x} = -0.78 \cdot \frac{1.1389}{257.66} = -0.003448$

$a = 2.3875 + 0.003448(324.75)$

$a = 3.5072$

fat gain $= 3.5072 - 0.003448$ (NEA)

d) Do people with larger increases in NEA tend to gain less fat? Explain your reason.

yes. the slope is negative which means that as NEA increases, fat gain decreases.

e) Determine the value of R^2 and explain what it means in context.

$r = -0.78$

$R^2 = (-0.78)^2 = 60.84\%$. of the variance in fat gain is accounted for by the LSRL of fat gain on NEA. :)