Statistical Reasoning
Scatterplots and Regression key
Name:

Date: Class:

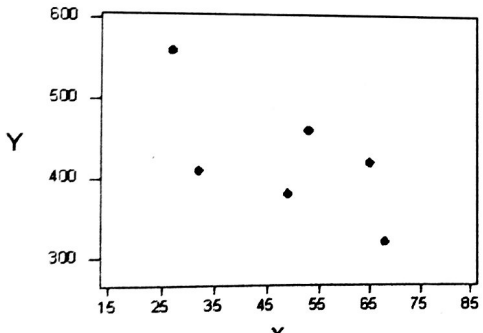
Analyzing a Linear Relationship: Regression

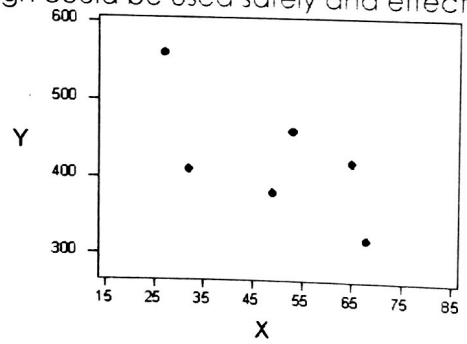
- So far we've used the scatterplot to describe the relationship between two quantitative variables, and with linear relationships we have supplemented the scatterplot strength with the correlation (r).
  - The correlation characterizes the linear relationship's strength and direction.

- After plotting the data in a scatterplot, and deciding if the correlation is strong enough to continue, we may further describe the data using a regression line line (the line drawn that best fits the data).
  - Regression can be used to make predictions.
  - Extrapolation - Using the regression line to predict outside the range of the data (i.e. a model of height vs. age of pre-teens doesn't extend to adults— growth curve doesn't continue). Use common sense.

- The least-squares regression line is the line that has the smallest sum of squared residuals (the deviations of the data points from the line, in the vertical direction).
  - It is the line that best fits all the scatter plot dots.

Example: Below is a scatterplot that shows a linear relationship between the age of a driver and the maximum distance at which a highway sign was legible. Suppose a government agency wanted to predict the maximum distance at which the sign would be legible for 60-year-old drivers, and thus make sure that the sign could be used safely and effectively.





The best line has the smallest variations between the predicted line and the actual data points. The deviations between the scatterplot's y-value and the regression line's y value is called a residual. Points that are below the predicted regression line are negative and points that are above the predicted regression line are positive. Getting a numeric quantity to compare the various lines proves difficult because often the positive and negative values cancel each other out and leave a sum of zero. Squaring the residuals removes all negatives and the lowest sum of squared residuals represents the regression line with least deviations.

This line is called the least-squares regression line, **and, as we'll see, it fits the linear pattern of the data very well.**

+ 2nd - 0

EXAMPLE: DiagnasticON

Hybartasuarus is an extinct beast having claws like a tiger and the strength of a bear. Only six fossil specimens are known. Because these specimens differ greatly in size, some scientists think they are completely different animals rather than the same animal at different ages.

If the fossils belong to the same species and differ in size because some are younger than others, there should be a strong positive linear relationship between the lengths of the femur and humerus bones. A weak linear relationship would suggest the fossils belong to different species.

Below is the data on the lengths in centimeters of the femur and the humerus for the five specimens.

Femur	38	56	59	64	74
Humerus	41	63	70	72	84

Step 1: Make a scatterplot of the data. **Step 2:** Perform a LIN REG **Step 3:** Draw your prediction line on your scatterplot (optional)

① 2nd stat plot
② window
③ Graph

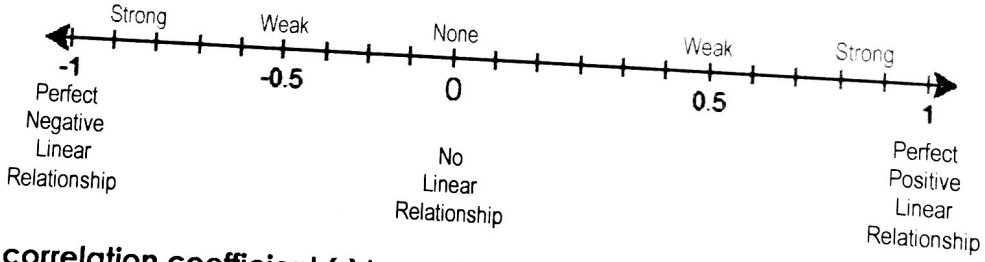
Stat
Calc
#4

a. What is the equation of the regression line?

$y = \underset{a}{1.2} \cdot x + \underset{b}{-3.66}$

b. Find the correlation coefficient (**r**). **r** = 0.99

c. Describe the regression line.



The **correlation coefficient (r)** is used to measure the how strongly the regression line predict the data. an **r** value that is closer to 1 or –1 is stronger than values closers to 0.

Shape: linear
Direction: positive
Strength: strong

d. Use your equation to predict the humerus length for a femur that is 85 cm. Is this a reasonable prediction? Why or why not?

$y = 1.2(85) - 3.66$

$= 98.34$ cm

10.4 Regression

Example 1:
Below are the data obtained in a study of age and systolic blood pressure of six randomly selected subjects.

Age (x)	Pressure (y)
43	128
48	120
56	135
61	143
67	141
70	152

a) Construct a scatterplot of the data.

b) Find the Least Squares Regression Line (LSRL). Identify the variables and the slope and y-intercept and interpret them in context.

$y = .96x + 81.05$ slope = .96 For each year, BP raises .96.

x = age y = pressure y-int = 81.05 at age 0, the BP is 81.05.

c) Find and interpret the correlation coefficient.

$r = .897$ The data have a strong, positive correlation.

d) Predict the Pressure for a person of age 75. Do you feel comfortable doing this?

$y = .96(75) + 81.05 = 153.05$

yes - strong correlation

Example 2:
Below are the data obtained in a study of the number of absences and the final grades of seven randomly selected students from a statistics class.

Number of Absences (x)	Final Grade (y)
6	82
2	86
15	43
9	74
12	58
5	90
8	78

e) Construct a scatterplot of the data.

f) Find the Least Squares Regression Line (LSRL). Identify the variables and the slope and y-intercept and interpret them in context.

$y = -3.62x + 102.49$ slope $= -3.62$ For each day absent, avg drops 3.62 pts.

$x =$ # absences $y =$ grade y-int $= 102.49$ 0 absences $= 102.49$ avg

g) Find and interpret the correlation coefficient.

$r = -.94$ The data have a very strong negative correlation

h) Predict the grade for a student who has missed 20 days. Do you feel comfortable doing this?

$y = -3.62(20) + 102.49 = \boxed{30.09}$

Yes - very strong r

Example 3:

Below are the data obtained in a study on the number of hours that nine people exercise each week and the amount of milk (in ounces) that each person consumes per week.

Hours (x)	Amount of Milk (y)
3	48
0	8
2	32
5	64
8	10
5	32
10	56
2	72
1	48

i) Construct a scatterplot of the data.

j) Find the Least Squares Regression Line (LSRL). Identify the variables and the slope and y-intercept and interpret them in context.

$y = .45x + 37.29$ slope $= .45$ For each hour exercised, milk consumption rises .45 oz,

$x =$ hours $y =$ milk y-int $= 39.29$ Someone who doesn't exercise, consumes 39.29 oz milk,

k) Find and interpret the correlation coefficient.

$r = .067$

The data are not really correlated.

l) Predict the amount of milk consumed by someone who exercises 12 hours per week. Do you feel comfortable doing this?

$y = .45(12) + 37.29 = 44.69$

NO - b/c not correlated