FST 2-3 Notes
TOPIC: Linear Regression & Correlation
GOAL
Discuss data which, when graphed, shows a roughly linear pattern of growth. Explain how to use technology to find an equation for the line of best fit and to determine the closeness of fit, as measured by the linear correlation coefficient.

SPUR Objectives

D Identify properties of regression lines and of the correlation coefficient.

F Find and interpret linear regression and models.

I Use scatterplots and residual plots to draw conclusions about linear models for data.

Vocabulary

method of least squares
line of best fit, least squares line, regression line
center of mass
correlation coefficient
perfect correlation
strong correlation
weak correlation

Linear Regression

Refers to finding the line of best fit by using the method of least squares.
least squares line
regression line

Properties

- Only 1 line of best fit for data set
- Contains the center of mass of the data $(\bar{x}, \bar{y})$ whose coordinates are the mean of the x-values and the mean of the y-values
- Slope & y-intercept computed from the data points

** Use the data on Pg 4 Curb weight vs Highway MPG

skip → The gold medal winning times for the men's 100-meter dash are listed below for the last 20 Summer Olympic Games.

skip this data

City	Year (L1)	Winning Time(s) (L2)
Beijing	2008	9.69
Athens	2004	9.85
Sydney	2000	9.87
Atlanta	1996	9.84
Barcelona	1992	9.96
Seoul	1988	9.92
Los Angeles	1984	9.99
Moscow	1980	10.25
Montreal	1976	10.06
Munich	1972	10.14
Mexico City	1968	9.95
Tokyo	1964	10.0
Rome	1960	10.2
Melbourne	1956	10.5
Helsinki	1952	10.4
London	1948	10.3
Berlin	1936	10.3
Los Angeles	1932	10.3
Amsterdam	1928	10.8
Paris	1924	10.6

a) Find a **best-fit linear model** for the relationship between the year *r* and the winning time *t*.

Step1: Enter data in to L1 and L2

STAT #1

Step 2: Create a scatterplot (STAT PLOT) of the data (ZOOM 9) 2nd Y=

Step 3: Find the line of best fit; LinReg(ax + b)

$\boxed{\text{STAT} \rightarrow \text{Calc \#4}}$

$y = -4.221212931x + 44.25674593$

Step 4: Graph the line of best fit

Go to Y= enter above equation

Then Hit Graph

b) Find the center of mass of the data. $(\bar{x}, \bar{y})$

STAT → Calc #1 1 var Stats L1

STAT → Calc #1 1 var Stats L2

c) Verify that the center of mass is on the line. $(\overset{\bar{x}}{4.325}, \overset{\bar{y}}{26})$

$\bar{x} = 4.325$

$\bar{y} = 26$

$y = -4.22121293(4.325) + 44.25674593$

$Y = 26$

d) Find the sum of squared residuals for the linear regression. STAT #1 Go to L3, Place cursor on L3, hit enter, 2nd STAT (LIST) ↓ RESID enter enter. STAT → Calc #1 1 var Stats L3 Σx^2 = Sum of squared residuals

$\Sigma x^2 = 60.52$

e) The 1940 and 1944 Summer Olympic Games were cancelled because of World War II. According to the regression line, what would the winning times have been in those years?

skip

Diagnostics Check

On your calculators, run the linear regression again for our data we used in class yesterday.

Does your screen display....

a =
b =

or

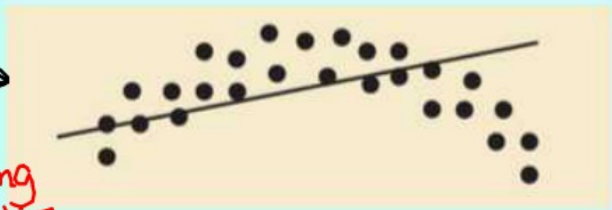
a =
b =
r^2 =
r =

If you have a & b *only*, you need to do the following:

2nd-CATALOG (last row) $\boxed{2nd}$ $\boxed{0}$
arrow down to DiagnosticOn
enter
enter again to DONE

run the linear regression again to verify that you do now get r^2 and r

Correlation



-1 Strong negative correlation

0 No correlation weak

1 Strong positive correlation

Correlation Coefficient $= r$

- Short for linear correlation
- denoted by the letter r
- measures the strength of the linear relation
- $-1 \leq r \leq 1$
- Will calculate using graphing calculator

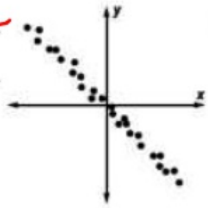
Positive correlation: as x increases, y increases (+ slope)

Negative correlation: as x increases, y decreases (- slope)

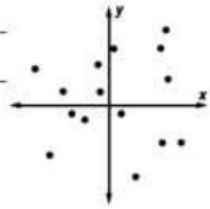
No correlation = no relationship

In 2–5 a dot plot is given. a. State whether the correlation coefficient of the line of best fit is positive, negative, or approximately zero. b. State whether the correlation is strong, medium, or weak.

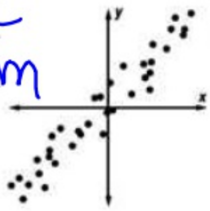
2. a. negative
b. Strong



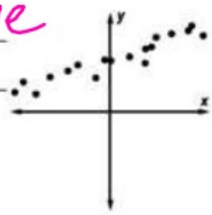
3. a. Zero
b. weak



4. a. positive
b. medium



5. a. positive
b. ~~medium~~ strong



1. Consider the table at the right that relates curb weight of certain 2008 vehicles and their estimated highway mpg.

L1 Curb Weight (000 lb)	L2 Highway mpg
2.6	34
6.8	18
5.7	23
4.1	22
3.5	28
2.5	37
3.4	30
6.0	16

a. Use a statistics utility to find a line of best fit for this data.

$y = -4.2212x + 44.2567$

b. Find the correlation coefficient. $r = {}^{-}.92$

c. Multiple choice. The correlation coefficient can best be described as

A weakly postive
B moderately positive
C strongly positive
D weakly negative
E moderately negative
F strongly negative close to -1

d. Describe in words what the correlation coefficient means in this context.

As the curb weight of certain vehicles ↑, the highest MPG ↓

e) Find the sum of squared residuals

$\Sigma x^2 = 60.52$