CORRELATION & REGRESSION

Chapter 10

CORRELATION

10-2

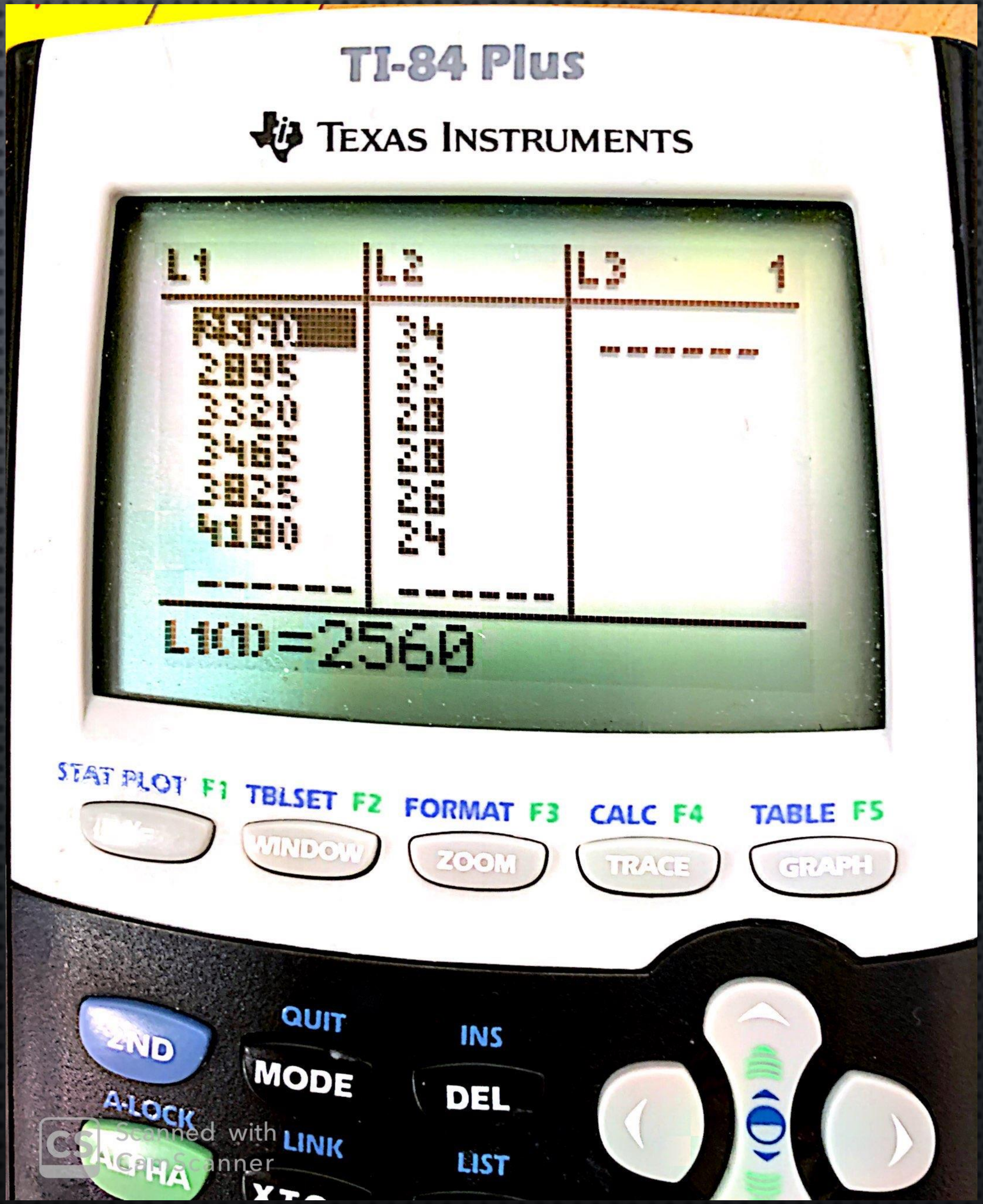
Bivariate Data

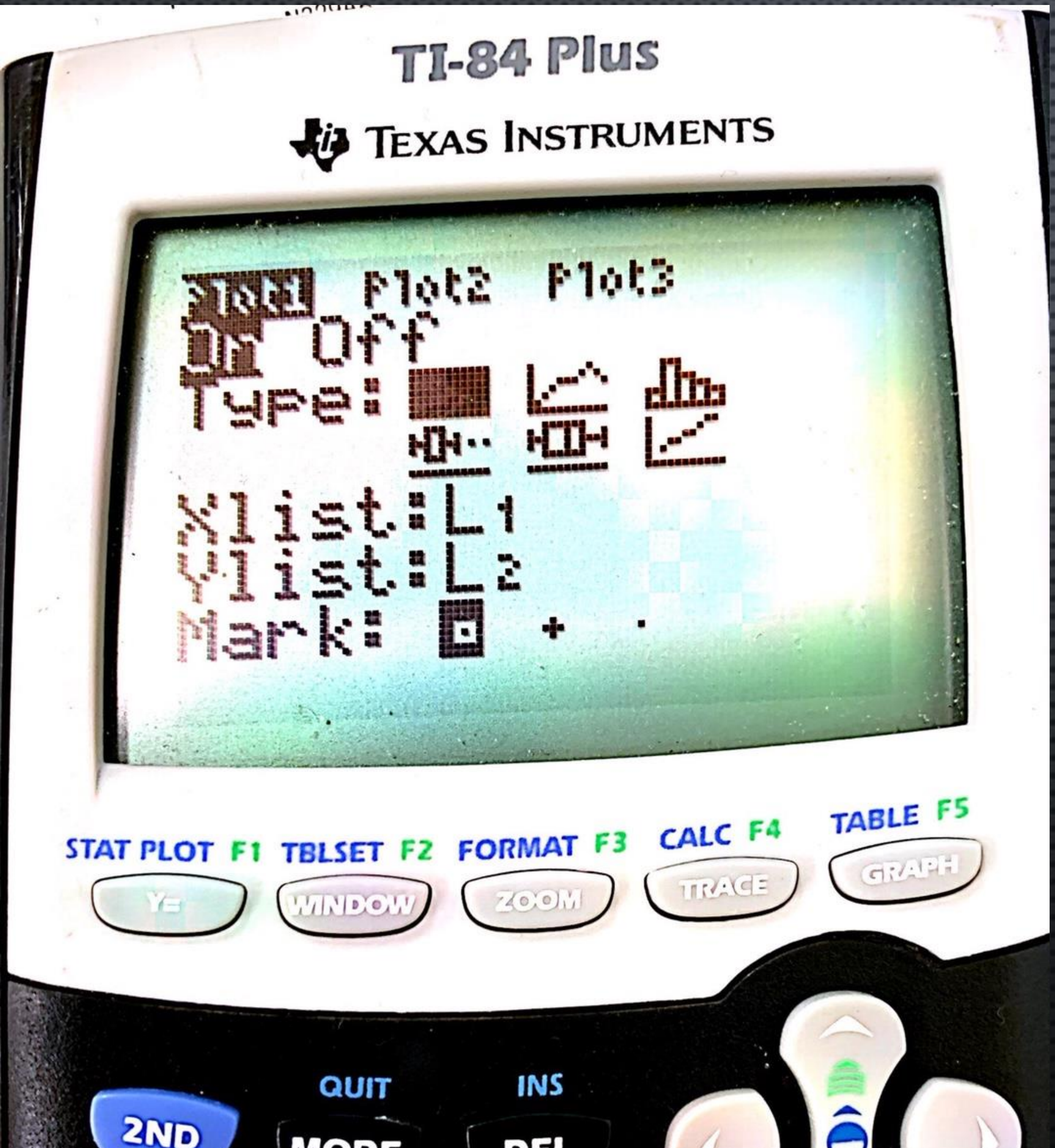
- Data on each of two variables, where each value of one of the variables is paired with a value of the other variable.
- Appendix B data set 14 (p754)

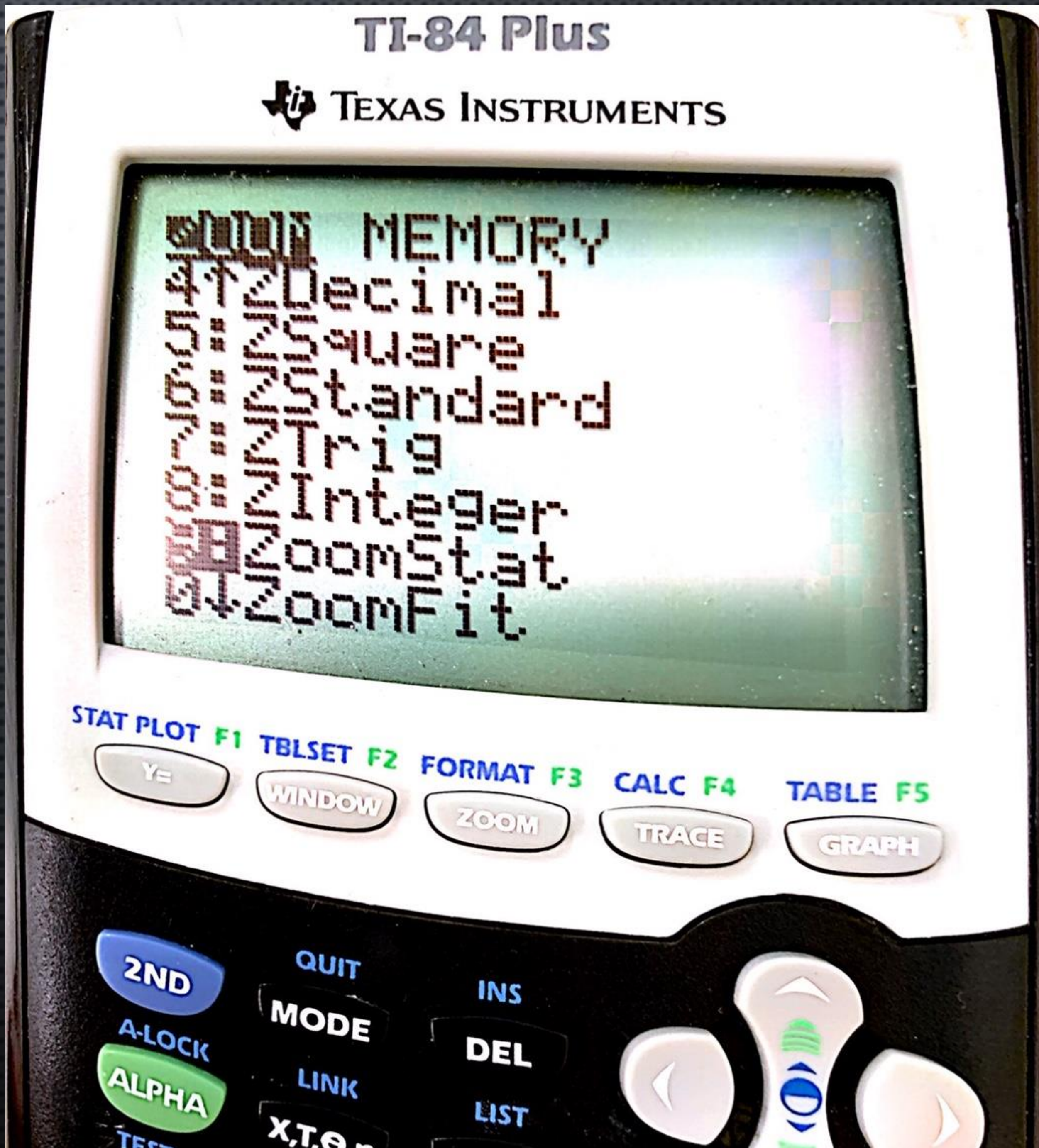
Car weight (lbs)	Highway MPG
2560	34
2895	33
3320	28
3465	28
3835	26
4180	24

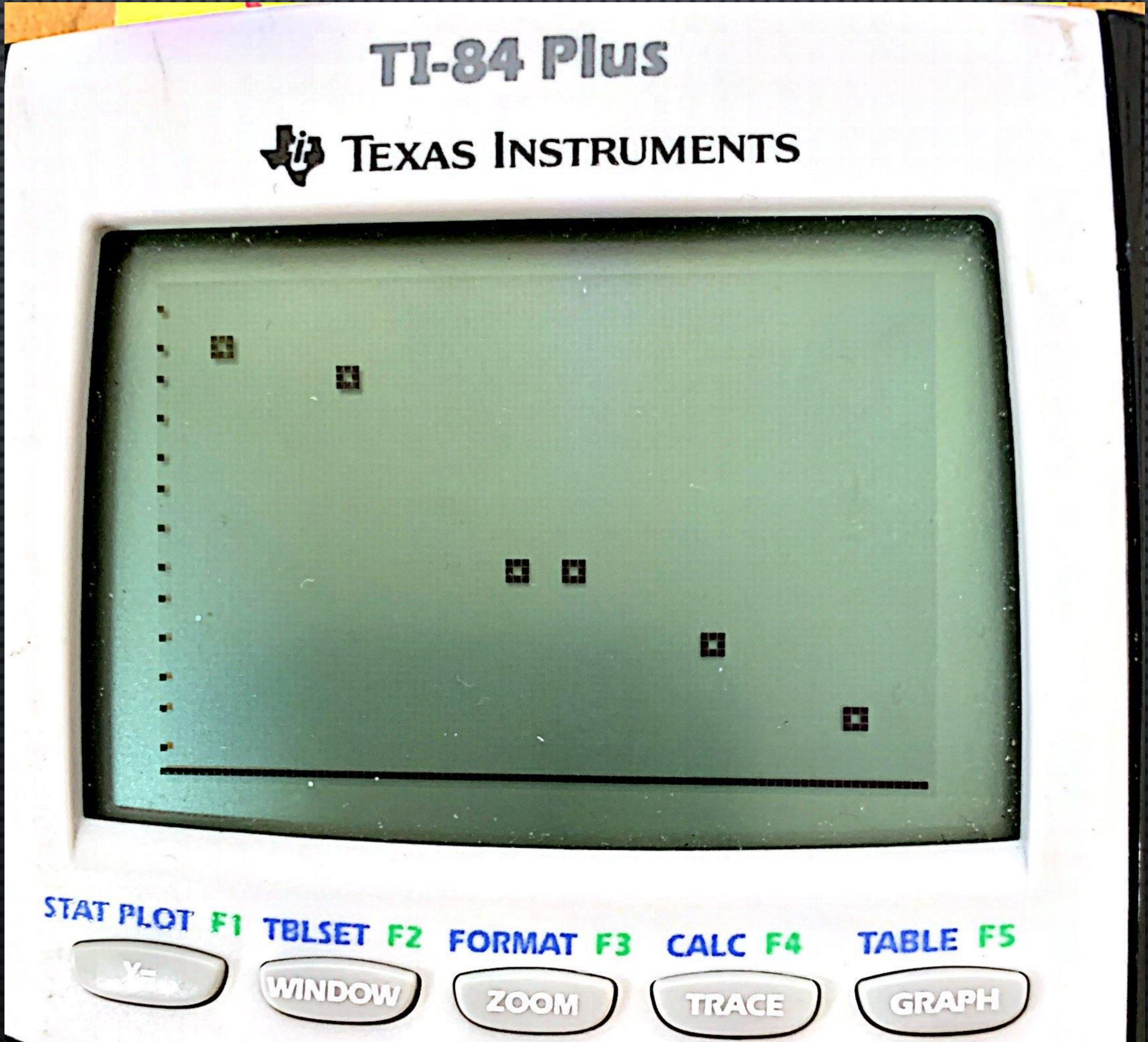
Scatterplot

- Use the statistical software on your calculator
- List Editor
- Stat Plot
- Zoom Stat




TI-84 Plus
TEXAS INSTRUMENTS
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Ylist:L2
Mark: + ·
STAT PLOT F1
TBLSET F2
FORMAT F3
CALC F4
TABLE F5
Y=
WINDOW
ZOOM
TRACE
GRAPH
2ND
QUIT
INS


TI-84 Plus
TEXAS INSTRUMENTS
ZOOM MEMORY
4↑ZDecimal
5:ZSquare
6:ZStandard
7:ZTrig
8:ZInteger
9:ZoomStat
0↓ZoomFit
STAT PLOT F1
TBLSET F2
FORMAT F3
CALC F4
TABLE F5
Y=
WINDOW
ZOOM
TRACE
GRAPH
2ND
QUIT
MODE
INS
DEL
A-LOCK
ALPHA
LINK
LIST


TI-84 Plus
TEXAS INSTRUMENTS
STAT PLOT F1
TBLSET F2
FORMAT F3
CALC F4
TABLE F5
Y=
WINDOW
ZOOM
TRACE
GRAPH

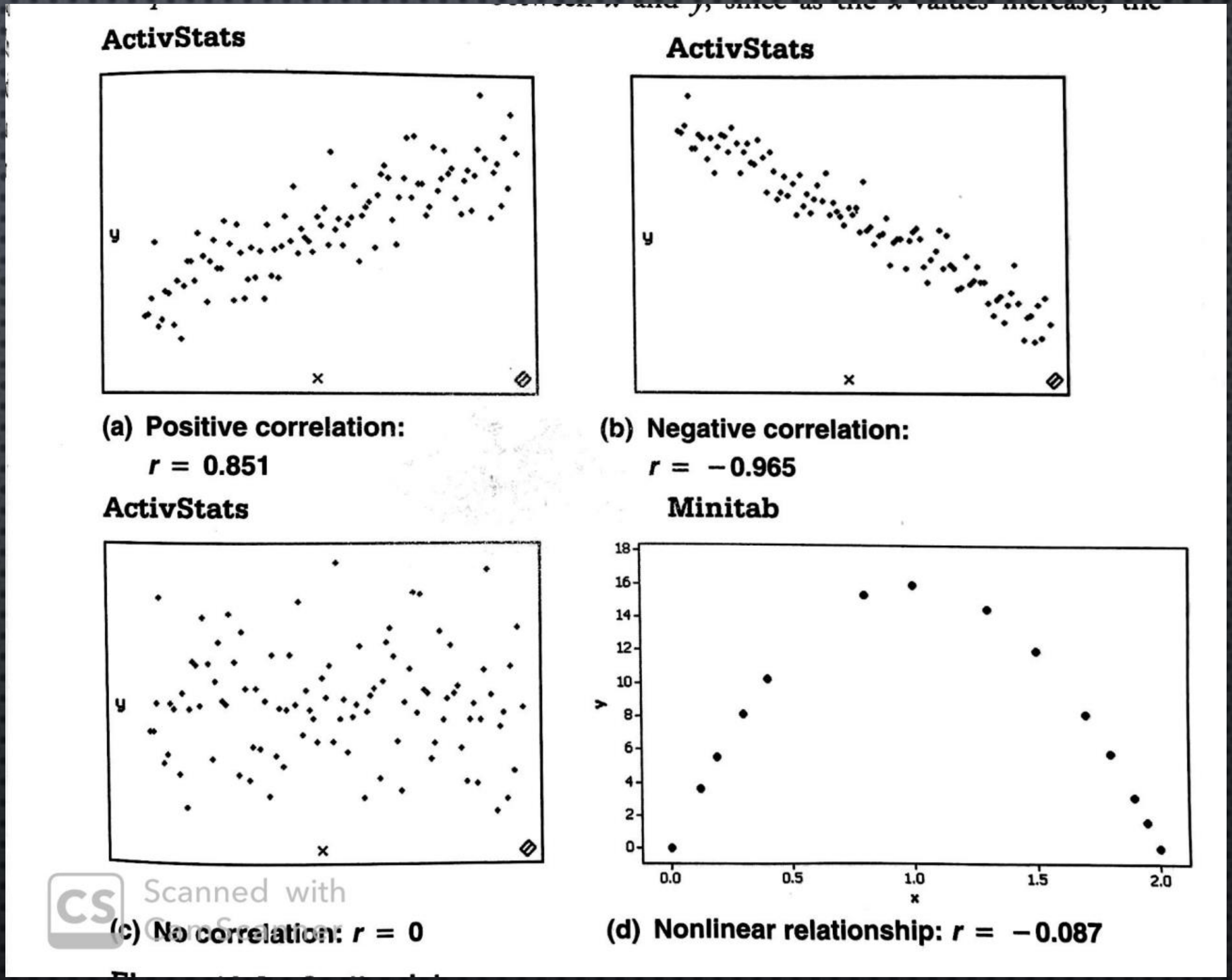
CORRELATION

- **Correlation** exists between two variable when the values of one variable are associated with the values of the other variable.
- **Linear Correlation** exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

Scatterplots

Linear correlation coefficient, r, measures the strength of the linear relationship

$$-1 \leq r \leq 1$$



Measuring Correlation Coefficient

- 3 methods
  - Sums of the squares & squares of the sums
  - Z-scores
  - technology

CORRELATION COEFFICIENT FORMULA 10-1 (P499)

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) \ - (\sum y)^2}}$$

	x (car weight)	y (MPG)	x^2	y^2	xy
	2560	34	6553600	1156	87040
	2895	33	8381025	1089	95535
	3320	28	11022400	784	92960
	3465	28	12006225	784	97020
	3825	26	14630625	676	99450
	4180	24	17472400	576	100320
sums	**20245**	**173**	**70066275**	**5065**	**572325**

$$r = \frac{6(572325) - (20245)(173)}{\sqrt{6(70066275) - (20245)^2}\sqrt{6(5065) - (173)^2}}$$

$$r = -0.982$$

CORRELATION COEFFICIENT FORMULA 10-2 (P499)

$$r = \frac{\sum(z_x z_y)}{n-1}$$

z_x DENOTES THE z SCORE FOR AN INDIVIDUAL SAMPLE VALUE x

z_y DENOTES THE z SCORE FOR AN INDIVIDUAL SAMPLE VALUE y

	x (car weight)	zx	y (MPG)	zy	zx*zy
	2560	-1.3737361	34	1.318027211	-1.8106215
	2895	-0.8084936	33	1.06292517	-0.8593682
	3320	-0.0913949	28	-0.212585034	0.0194292
	3465	0.1532623	28	-0.212585034	-0.0325813
	3825	0.760687	26	-0.722789116	-0.5498163
	4180	1.3596753	24	-1.232993197	-1.6764704
mean	3374.16667		28.833333		**-4.9094285**
st dev	592.666		3.92		

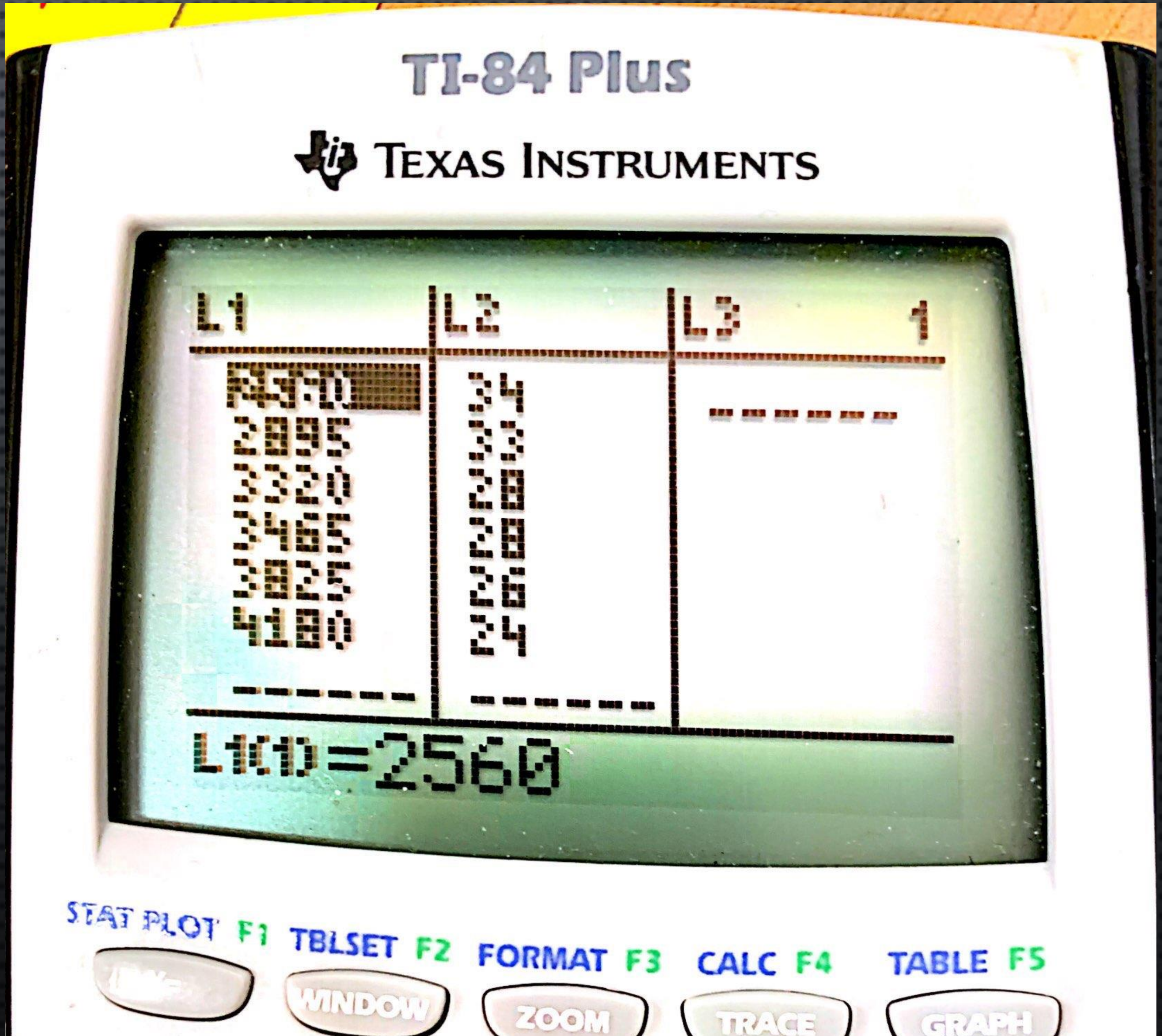
$$r = \frac{\sum(z_x z_y)}{n-1}$$

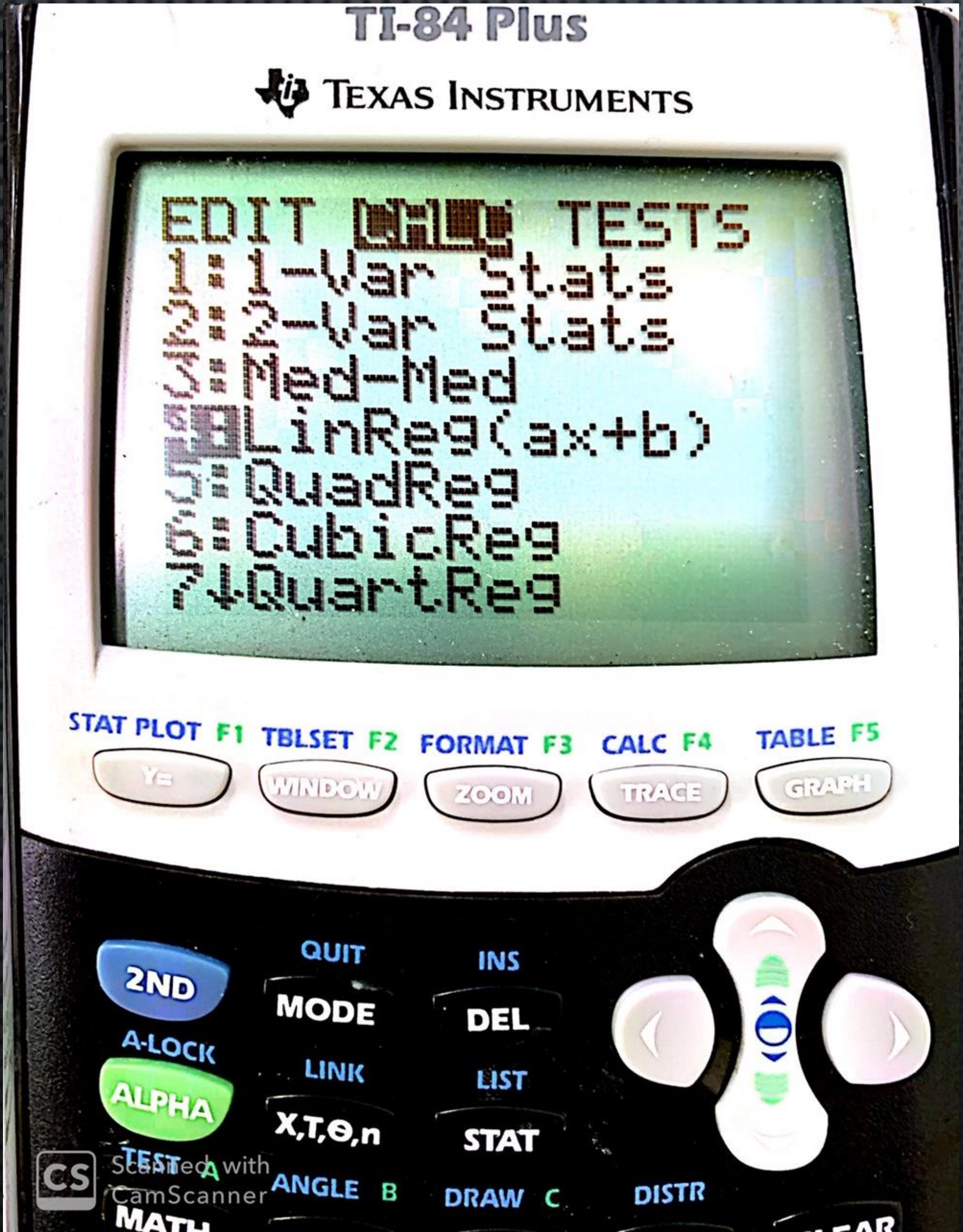
$$r = \frac{-4.909}{5}$$

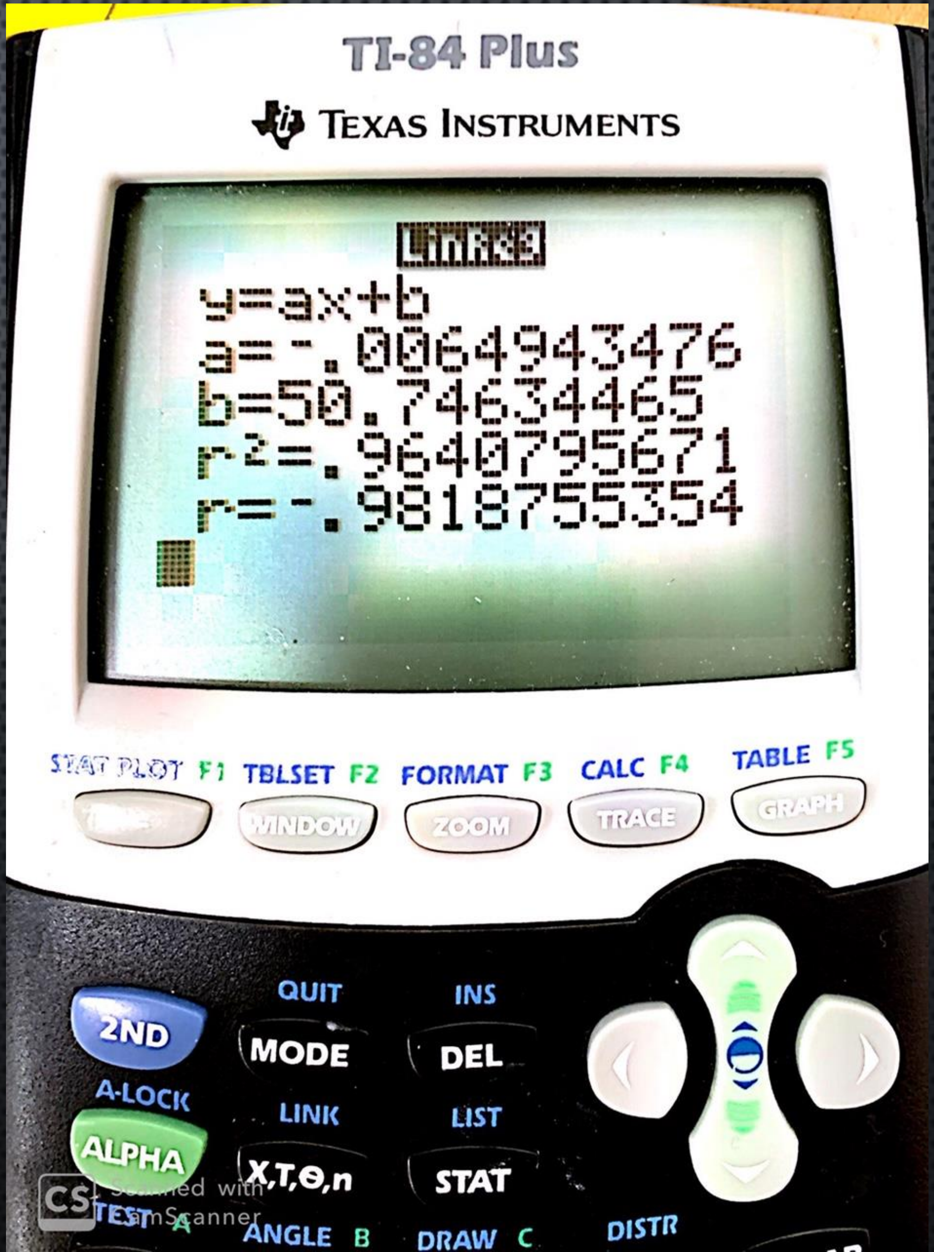
$$r = -0.9818$$

USING TECHNOLOGY

- TI-84
- List Editor
- Stat
  - Calc
  - LinRegression

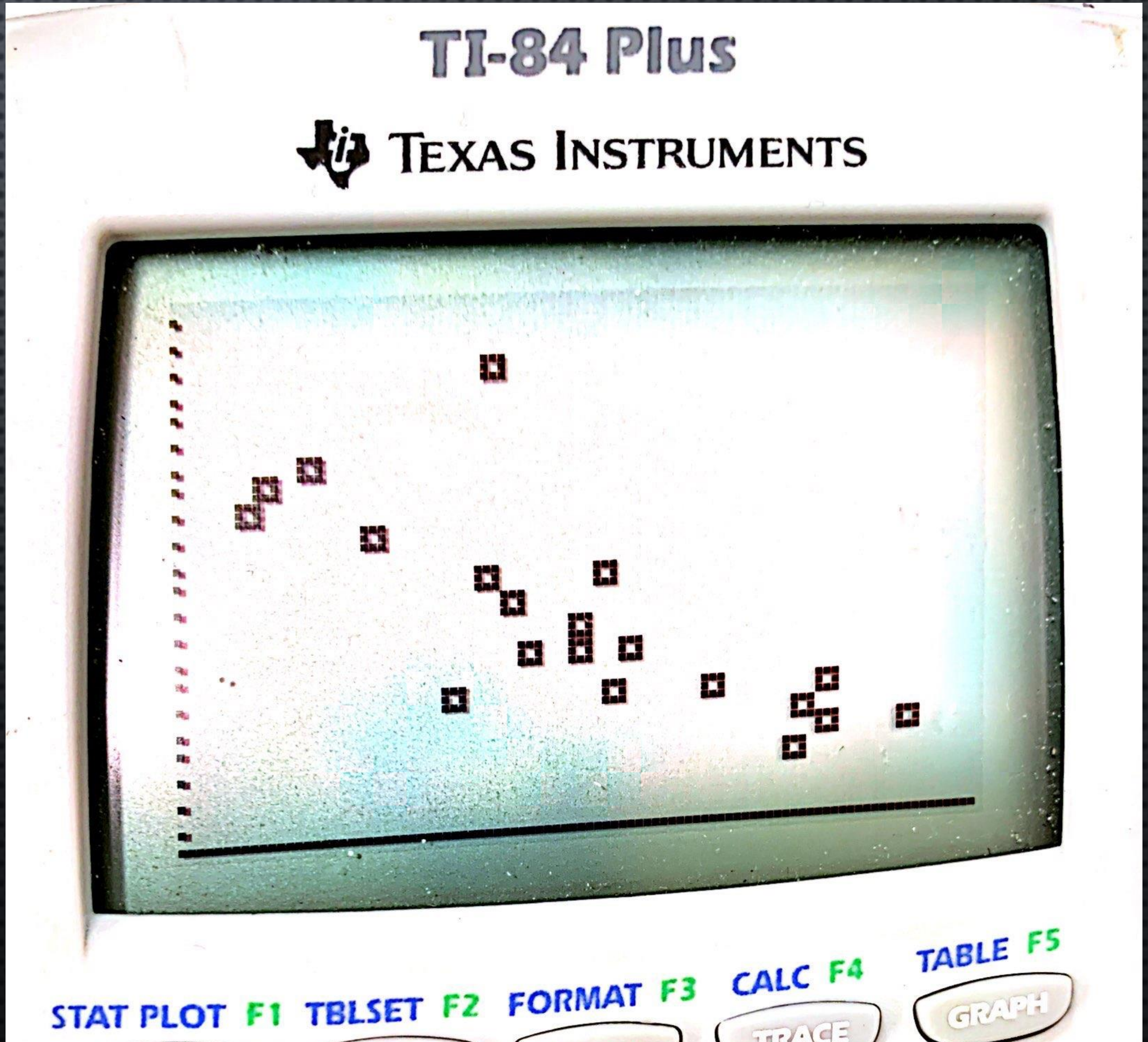



TI-84 Plus
TEXAS INSTRUMENTS
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg

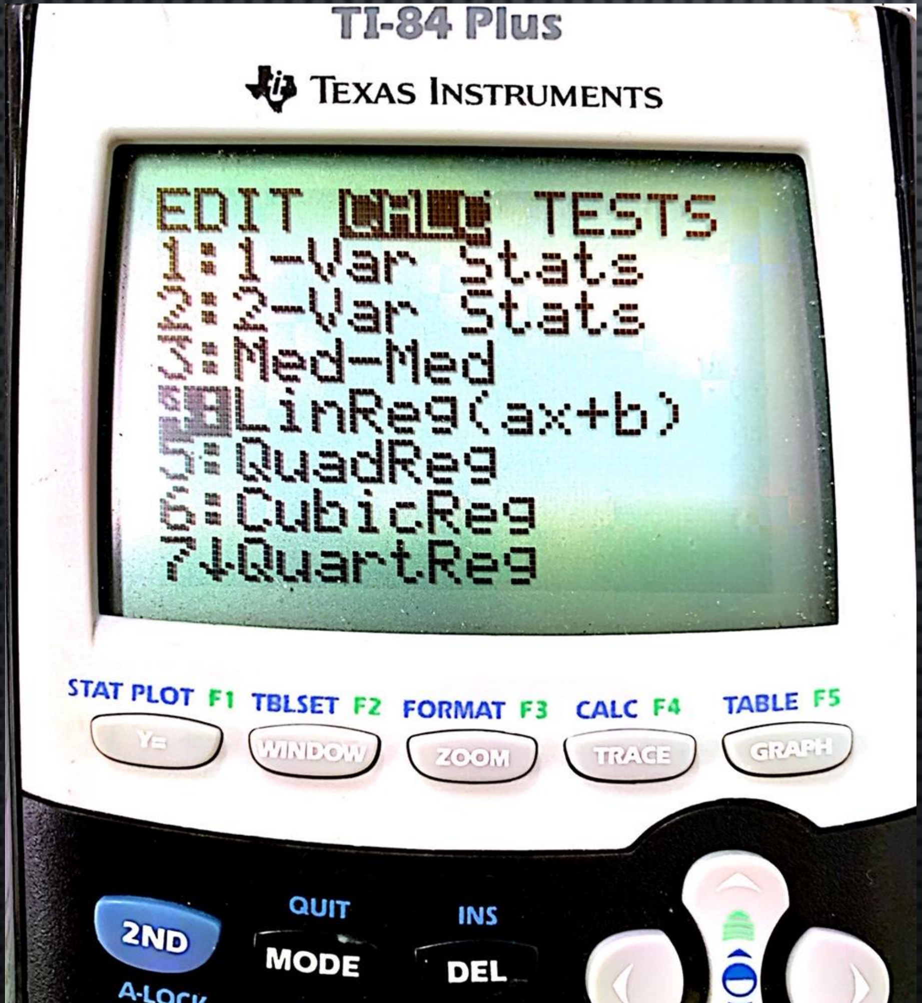

TI-84 Plus
TEXAS INSTRUMENTS
LinReg
y=ax+b
a=-.006494347б
b=50.74634465
r²=.9640795671
r=-.9818755354

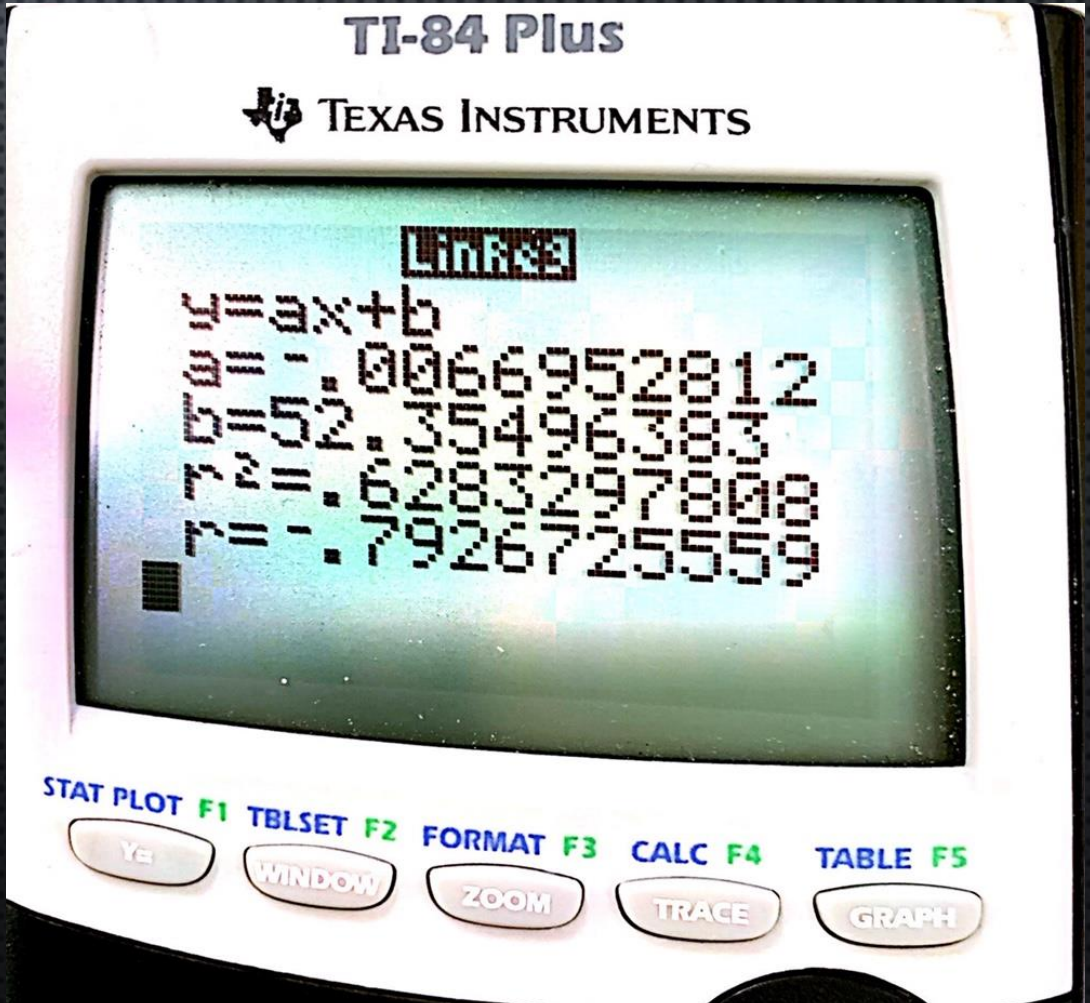
CORRELATION FOR LARGE DATA SETS

- Appendix B data set 14 (p754)
- What if we use all of the data in data set 14
- $n = 21$



CORRELATION FOR LARGE DATA SETS





SMALL V. LARGE DATA SETS

$$n = 6$$

$$r = -0.9818$$

$$n = 21$$

$$r = -0.7927$$

Small v. Large Data Sets

- Are we suggesting that the larger data set has a weaker Linear relationship?
- Use hypothesis testing to test the claim of a linear correlation between two variables

Hypothesis Testing for Linear Correlation

- r – the sample correlation coefficient
- ρ (rho) – The population correlation coefficient
- $H_0: \rho = 0$ (there is no linear correlation)
- $H_A: \rho \neq 0$ (there is a linear correlation)

THE TEST STATISTIC

- $t = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$

- Critical Value can be found in table A-3
- $n - 2$ degrees of freedom

HYPOTHESIS TEST

$H_0: \rho = 0$

$H_A: \rho \neq 0$

$r = -0.7927$

$n = 21$

$df = 19$

$\alpha = 0.05$

$t_{\alpha/2} = 2.093$

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

$$t = \frac{-0.7927}{\sqrt{\frac{1 - (-0.7927)^2}{21 - 2}}}$$

$$t = -5.66$$

ALTERNATIVE METHOD

- INSTEAD OF USING THE TEST STATISTIC, COMPARE THE SAMPLE CORRELATION COEFFICIENT TO THE CRITICAL VALUES OF THE PEARSON CORRELATION COEFFICIENT r
- TABLE A-6 ON PAGE 732
- $r = -0.7927$
- $n = 21$

TO TEST $H_0: \rho = 0$ AGAINST $H_A: \rho \neq 0$, REJECT H_0 IF THE ABSOLUTE VALUE OF r IS GREATER THAN THE CRITICAL VALUE IN THE TABLE.

$H_0: \rho = 0$
$H_A: \rho \neq 0$

$r = -0.7927$
$n = 21$

$0.7927 > 0.444$
$reject\ H_0$

Table A-6 Critical Values of the Pearson Correlation Coefficient r

n	$\alpha = .05$	$\alpha = .01$
4	.950	.990
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330
70	.236	.305
80	.220	.286

HOMEOWRK

- P513 #13-16, 24, 29

Additional Things About Correlation

Properties of r

- The value of r is always between -1 and 1 inclusive
- The value of r is not affected by the choice of x or y
- r only measures the strength of a linear relationship
- r is very sensitive to outliers

Common errors

- Correlation does not mean causation
- Errors arise when data is based on averages or rates
- r is only a test for linear correlation. Just because paired data are not related linearly does not mean that they aren't related in some other way.

r^2 The Proportion of Variation

- $r = -0.7927$
- $r^2 = 0.6284$
- This means that 62.84% of the variation in the MPG can be explained by the linear relationship

Linear Regression

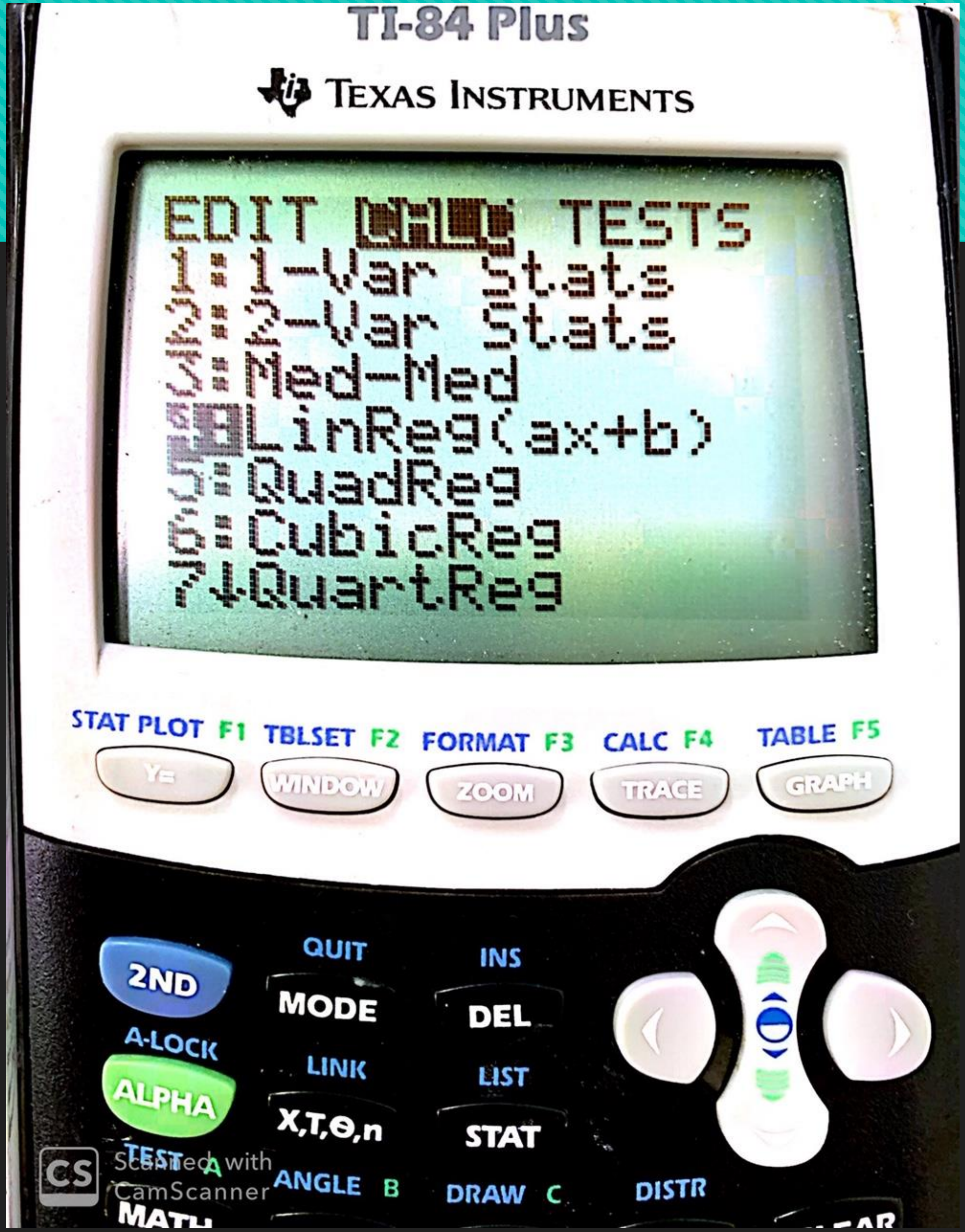
10-3

Regression line

- The straight line that “best” fits a set of bivariate data
- $\hat{y} = b_0 + b_1 x$
  - x - predictor variable, independent variable
  - $\hat{y}$ - response variable, or dependent variable
  - b_0 - y-intercept
  - b_1 - slope

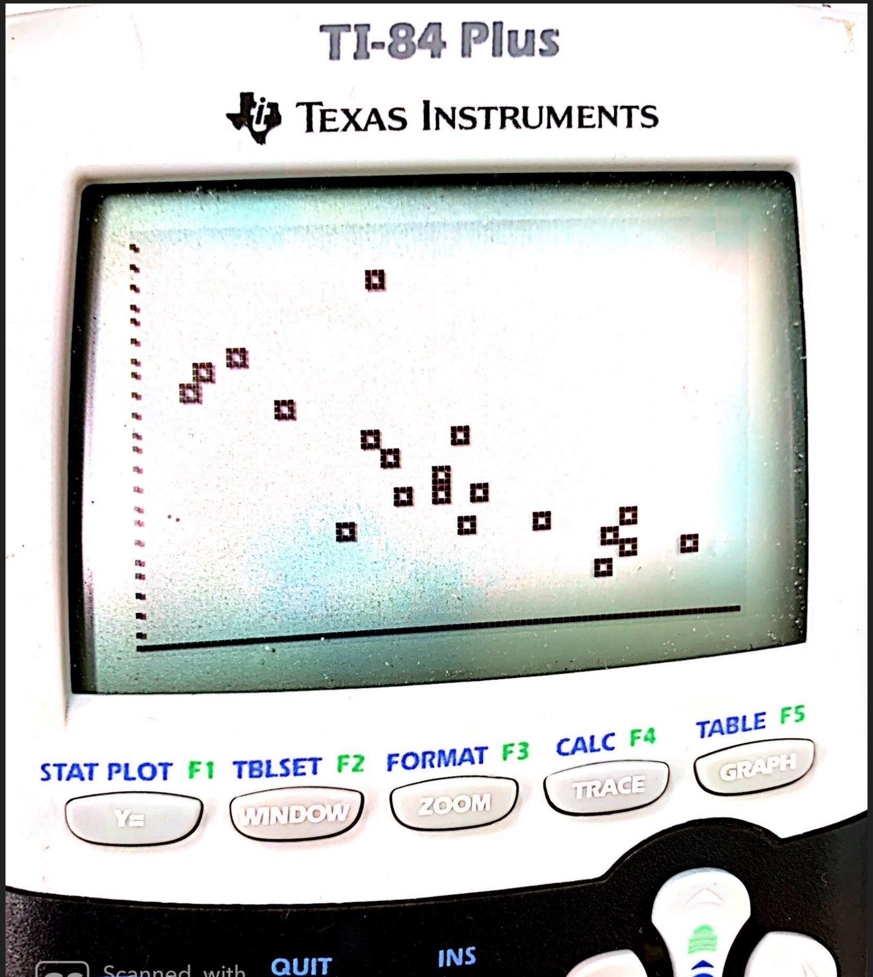
Linear Regression

- Use the linear regression function in the calculator to find the equation for the line of best fit

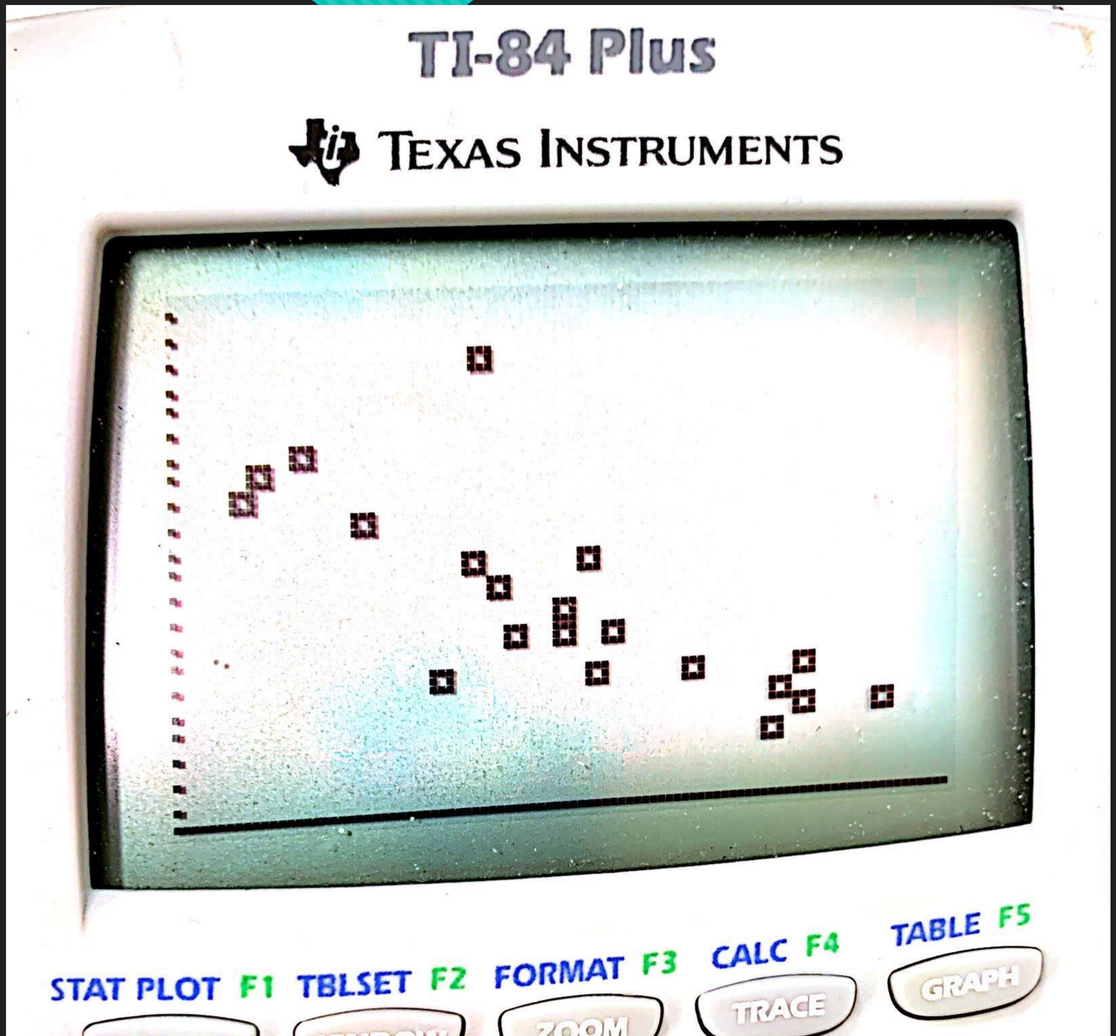


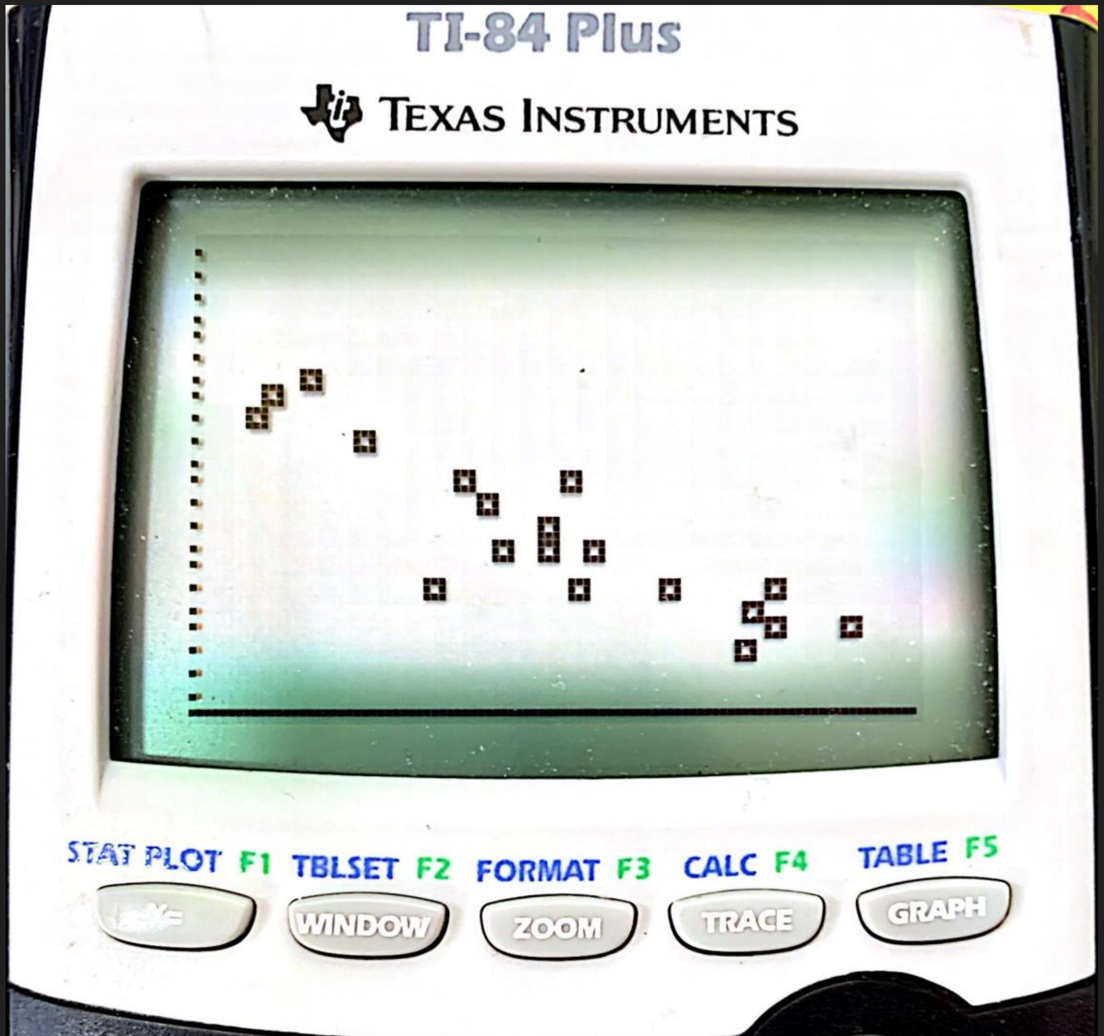
Requirement check

- Data is assumed to be simple random sample
- A scatterplot suggests a linear pattern
- There is a linear correlation
- There are no outliers



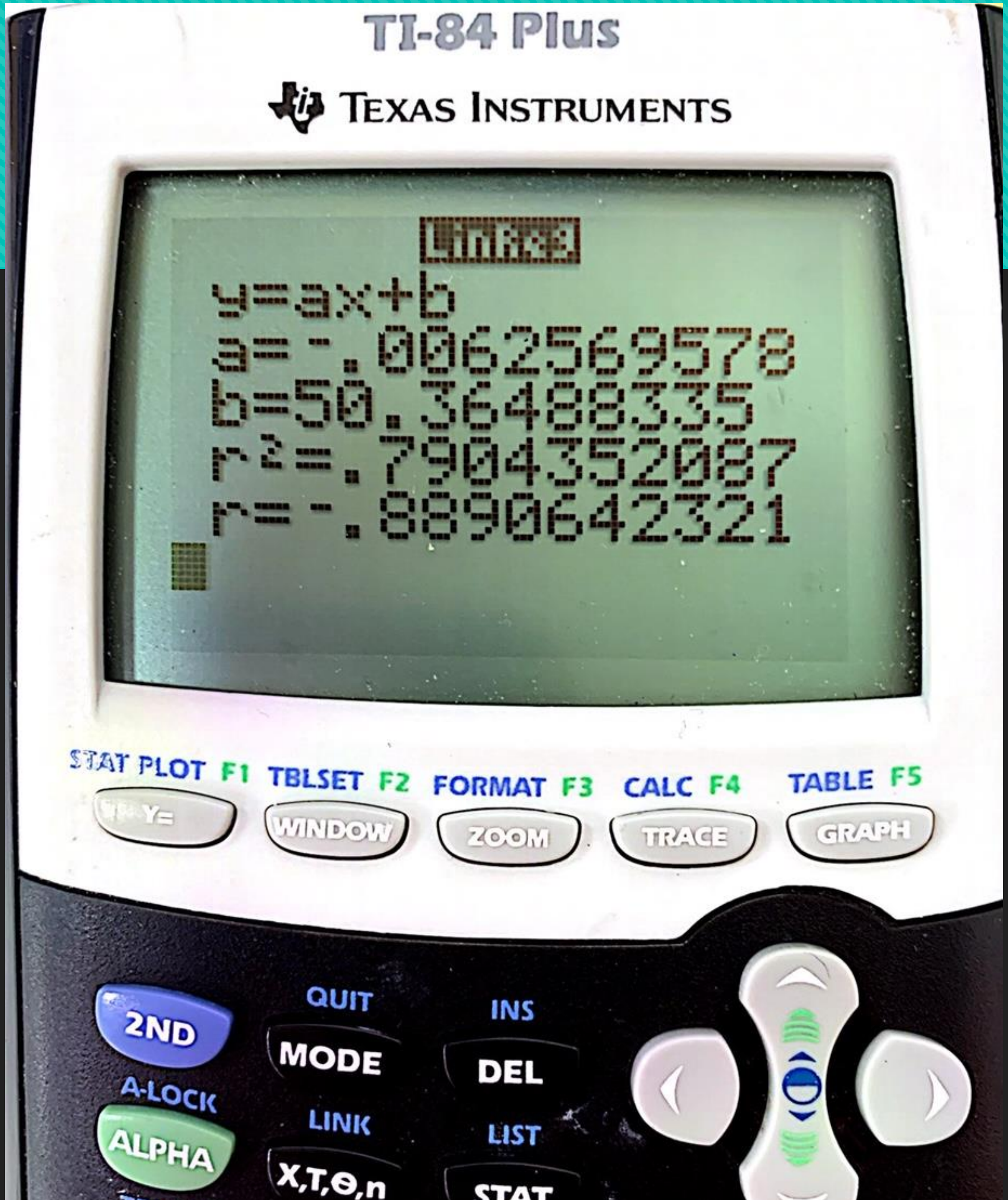
scatterplots



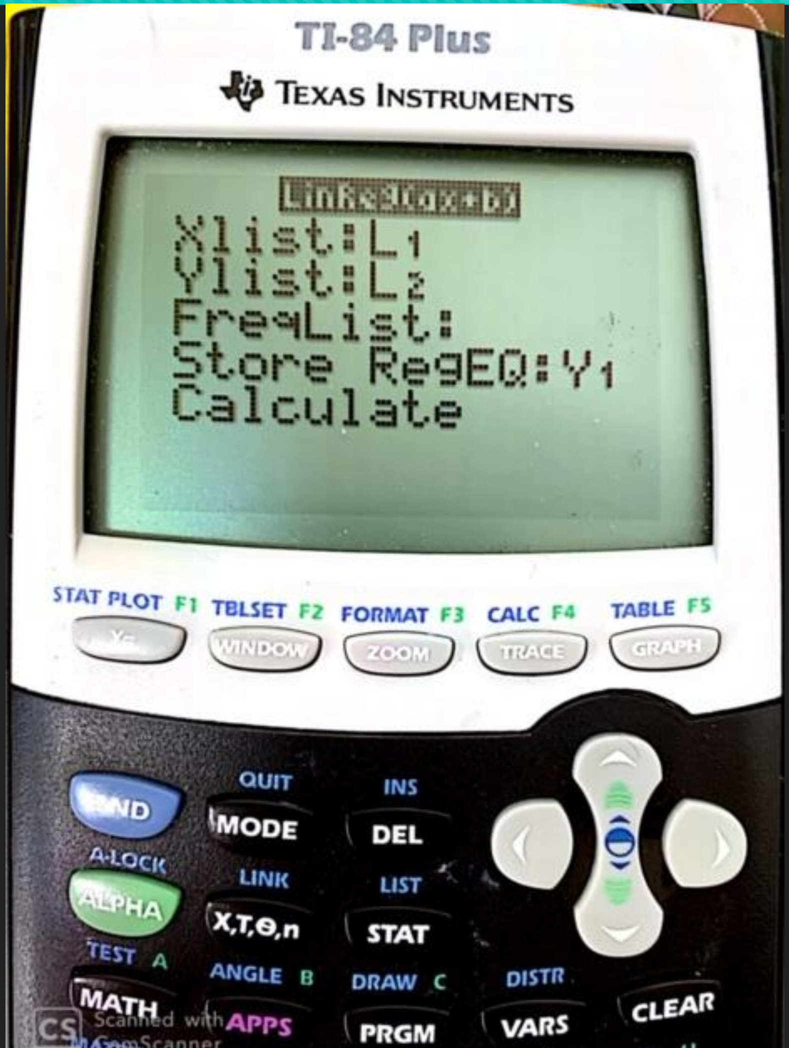


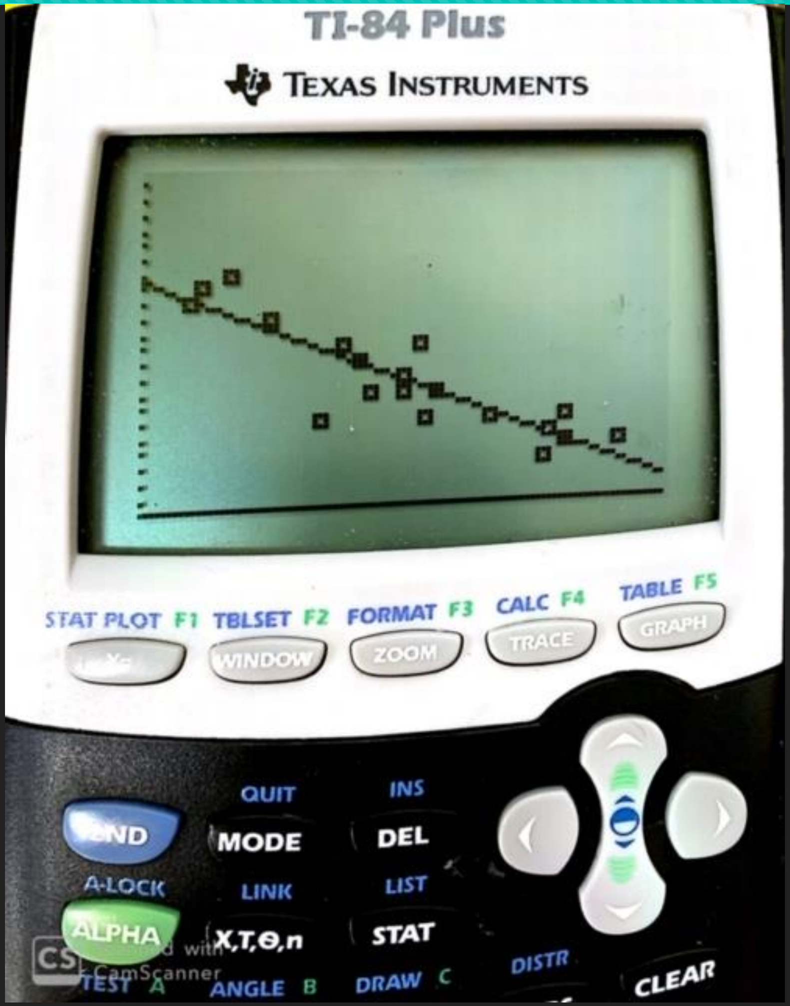
Linear regression

- Use three significant digits
- $\hat{y} = 50.4 - 0.006x$



Storing the regression equation





Making predictions

$$\hat{y} = 50.4 - 0.006x$$

- Find the expected Highway Fuel efficiency for a car weighing 5000 lbs.
- Find the fuel efficiency of the Hummer H2 weighing 6,400 lbs
- Find the fuel efficiency of the Smart Car weighing 1,500 lbs
- How much should a car weight to get 60 MPG??

What if there is no linear correlation!?

- Then the value of $\hat{y}$ is assumed to be $\bar{y}$ for any predictor value of x
- Why?
  - Because $\bar{y}$ is the expected value of $\hat{y}$

Homeowrk

- P529 #13-16, 24, 29