

Correlation and Linear Regression

You are responsible for the following:

1. Be able to enter the paired data (x,y) into your calculator and know how to retrieve the following:

r the sample correlation coefficient

r^2 the coefficient of determination

n the sample size of the paired data

\bar{y} the mean for random variable y (know when to use y for prediction)

\hat{y} the equation for the sample regression line $\hat{y} = a + bx$

a the y-intercept (know when to use \hat{y} for prediction)

b the slope

\bar{x} the mean for random variable x

s_e the standard error of the estimate (s_e)

$\Sigma x, \Sigma x^2, \Sigma y, \Sigma y^2, \Sigma xy$ summation for $x, y, x^2, y^2,$ and xy

2. Test for significant linear correlation using the P-Value method or Table A-5. Make a conclusion that states one of the three following conclusions:

No Linear Correlation

Significant Positive Linear Correlation

Significant Negative Linear Correlation

3. Draw a scatter-plot of the paired data (include the regression line if correlation statistically significant). You may draw the regression line by approximation (try to make it look like the line of best fit). Use the Stat Plot option on the TI-83 to graph to help you with your plot.

4. Provide a one sentence statement using the coefficient of determination to express the percentage of variation in y expressed by the variation in x .

5. Be able to construct a prediction interval. The formula can be programmed into your calculator (see program "PREDINT").

$$\hat{y} - E < Y < \hat{y} + E$$

$$E = t_{\frac{\alpha}{2}} S_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}} \quad S_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

6. Be able to test for significant correlation between ranked data or data that may not be a bivariate normal distribution. Use Spearman's rank correlation test. The procedure can be programmed into your calculator (see program "RANKCORR").

EXAMPLE: Is murder rate related to the number of registered automatic weapons? Test at the 0.05 level of significance. The data below represents the number of automatic weapons in thousands and the murder rate is the rate per one hundred. What is the best predicted murder rate for a state with 10,000 registered automatic weapons? What is the 95% prediction interval for murder rate based on 10,000 registered automatic weapons? What is the coefficient of determination? Explain.

Automatic weapons	11.6	8.3	3.6	0.6	6.9	2.5	2.4	2.6
Murder rate	13.1	10.6	10.1	4.4	11.5	6.6	3.6	5.3

(1) Enter Automatic weapon registration as explanatory variable (x) and Murder rate as response variable (y).

$$r \approx 0.8849841667$$

$$r^2 \approx 0.7831969753$$

$$n = 8$$

$$\bar{y} = 8.15$$

$$\hat{y} \approx 4.047250782 + 0.852519318x$$

$$\bar{x} = 4.8125$$

$$\sum x = 38.5, \quad \sum x^2 = 283.15, \quad \sum y = 65.2, \quad \sum y^2 = 622.2, \quad \sum xy = 397.21$$

(2) Test for significant correlation at

$$H_0: \rho = 0$$

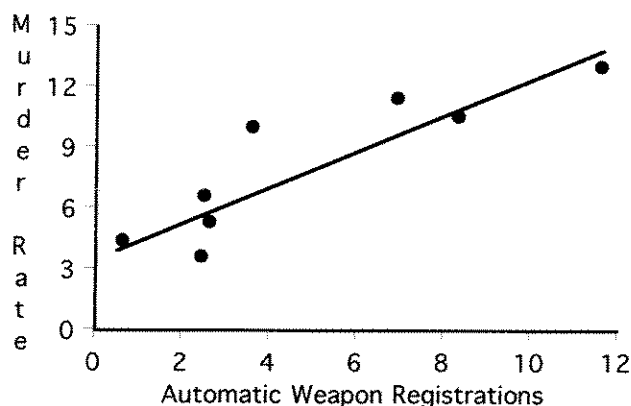
$$H_1: \rho \neq 0$$

Test Statistic (from TI-83) $r \approx 0.885$ (Use Table A-5 or P-Value)

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \quad t = \frac{0.88984}{\sqrt{\frac{1-0.783197}{6}}} \approx 4.655629994 \quad \text{P-Value} \approx 0.0034831867$$

Conclusion: Reject Null Hypothesis
Significant Positive Linear Correlation

(3) Draw a scatter-plot of the paired data.



(4) Coefficient of determination

$r^2 \approx 0.783$ 78.3% of the variation in the murder rate can be explained by the variation in automatic weapon registrations.

(5) Construct a prediction interval (choose $x_0 = 10$ (10,000 registered auto. weapons)).

Best Point Estimate: $\hat{y} \approx 12.57244396$

$$E = 5.244931687$$

Prediction interval (95% with 6 degrees of freedom)

$$\text{CV: } t = 2.447 \quad x_0 = 10$$

$$\hat{y} - E < Y < \hat{y} + E$$

$$12.57244396 - 5.244931687 < Y < 12.57244396 + 5.244931687$$

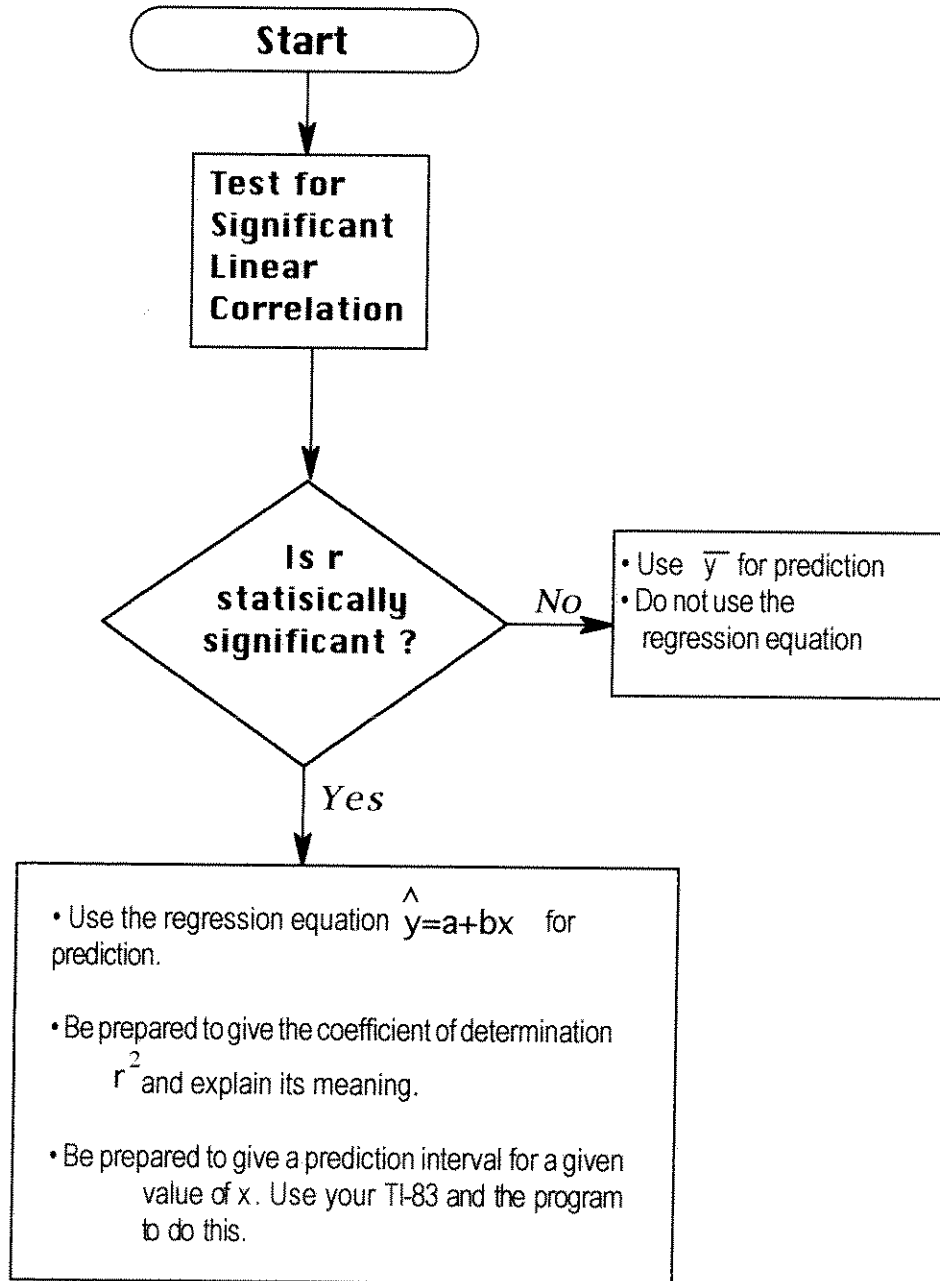
$$\boxed{7.328 < Y < 17.817}$$

(6) Using the Murder Rate data for Spearman's Rank Correlation:

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \text{ from the TI-83 program RANKCORR } r_s = 0.9285714286$$

from Table A-6 with $\alpha = 0.05$ the CV = 0.738

Conclusion: Reject the Null Hypothesis, Significant Positive Linear Correlation



Homework - 8 Problems

For the following seven problems do these steps:

1. Enter the data into your TI-83.
2. Test for significant linear correlation and make a conclusion. Use Rank Correlation if appropriate.
3. Draw a scatter plot of the data set and include the regression line if statistically significant.
4. Give the coefficient of determination and a one sentence statement using it as a percentage.
5. Provide a predicted value for the response variable given a value for the explanatory variable.

(1) Randomly selected girls are given the Wide Range Achievement Test. Their ages are listed along with their scores on the reading part of that test. (Based on data from The National Health Survey) Give a prediction interval for an 8 year old girl (if appropriate).

AGE	6.1	7.2	5.9	6.3	10.5	11.0
SCORE	17.8	47.4	25.8	24.3	66.6	91.4

(2) At one point during a recent season of the National Basketball Association, USA Today reported the current statistics. Give below are the total minutes played and the total points scored by 9 randomly selected NBA players. Give a prediction interval for 500 minutes played (if appropriate).

Minutes	1364	53	457	717	384	1432	365	1626	840
Points	652	20	163	210	175	821	143	1098	459

(3) The following data lists numbers of patio tiles and the costs in dollars of having them cut manually to fit. Give a prediction interval for the cost of 4 tiles (if appropriate).

Tiles :	1	2	3	5	6
Cost \$:	5	8	11	17	20

(4) The following data represents the living space in square footage (hundreds of square feet) and the selling price (in thousands of \$) of houses in Shasta County. Give a prediction interval for an 1,800 square foot house (if appropriate).

Sq. Footage :	15	38	23	16	16	13	20	24
Selling Price :	145	228	150	130	160	114	142	265

(5) A study was conducted to investigate any relationship between age (in years) and BAC (blood alcohol concentration) measured when convicted DWI jail inmates were first arrested. Give a prediction interval for a 25 year old.

Age (in yrs):	17.2	43.5	30.7	53.1	37.2	21.0	27.6	46.3
BAC :	0.19	0.20	0.26	0.16	0.24	0.20	0.18	0.23

continue

(6) The accompanying table lists weights (in hundreds of pounds) and highway fuel usage rate (in miles per gallon) for a sample of domestic new cars. Based on the results, can you expect to pay more for gas if you buy a heavier car? Give a prediction interval for 2,700 lbs. (if appropriate).

Weight :	29	35	28	44	25	34	30	33	28	24
Fuel :	31	27	29	25	31	29	28	28	28	33

(7) A *Raiders of the Lost Ark* pinball machine is used to measure learning that results from repeating manual functions. Subjects were selected so that they are similar in important characteristics of age, gender, intelligence, education and so on. We expect that there should be an association between the number of games played and the pinball score. Is there sufficient evidence to support the claim that there is such an association?

Number of games played :	9	13	21	5	6	25	7	33	11	104
Score :	22	62	70	2	10	82	26	78	58	86

(8) *Blood Pressure Measurements* Fourteen different second-year medical students took blood pressure measurements of the same patient and the results are listed below (data provided by Marc Triola, MD).

Systolic	138	130	135	140	120	125	120	130	130	144	143	140	130	150
Diastolic	82	91	100	100	80	90	80	80	80	98	105	85	70	100

Find the: (a) explained variation (b) unexplained variation (c) total variation
 (d) coefficient of determination (r^2) (e) standard error of estimate (s_e).
 (Assume that there is significant linear correlation.)

(1) Randomly selected girls are given the Wide Range Achievement Test. Their ages are listed along with their scores on the reading part of that test. (Based on data from The National Health Survey)

AGE	6.1	7.2	5.9	6.3	10.5	11.0
SCORE	17.8	47.4	25.8	24.3	66.6	91.4

1. Enter the girls age as the explanatory variable (x) and the score on the test as the response variable (y).

$$r \approx 0.958$$

TS: $t \approx 6.713$ Table A-3 CV: $t \approx \pm 2.776$

P-Value ≈ 0.0026

Table A-6 $r \approx \pm 0.811$

2. Conclusion: **Significant Positive Linear Correlation**

3.



$$\bar{y} \approx 45.55$$

$$\hat{y} \approx -48.3 + 12.0x$$

4. Coefficient of Determination $r^2 \approx 0.918$

91.8% of the variation in the test scores can be explained by the variation in the ages of the girls.

5. Construct a 95% prediction interval for an 8 year old girls.

CV: $t = 2.776$ $X_0 = 8$ $\hat{y} \approx 47.546$ $E \approx 27.62$ **$19.9 < Y < 75.2$**
Best point estimate

(2) At one point during a recent season of the National Basketball Association, *USA Today* reported the current statistics. Give below are the total minutes played and the total points scored by 9 randomly selected NBA players.

Minutes	1364	53	457	717	384	1432	365	1626	840
Points	652	20	163	210	175	821	143	1098	459

1. Enter minutes as the explanatory variable and points as the response variable.

$$r \approx 0.967$$

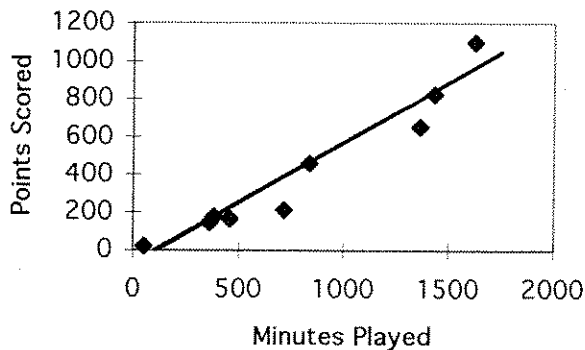
TS: $t \approx 9.97$ Table A-3 CV: $t \approx \pm 2.365$

P-Value ≈ 0.00002

Table A-6 $r \approx \pm 0.666$

2. Conclusion: Significant Positive Linear Correlation

3.



$$\bar{y} = 415.7$$

$$\hat{y} = -101 + 0.643x$$

4. Coefficient of Determination $r^2 \approx 0.934$

93.4% of the variation in points scored can be explained by the variation in minutes played.

5. Construct a 95% prediction interval for 500 minutes played.

CV: 2.365 $X_0 = 500$ $y \approx 220.1$ $E \approx 255.7$

$$\text{-35.7} < Y < \text{475.8}$$

(3) The following data lists numbers of patio tiles and the costs in dollars of having them cut manually to fit.

Tiles :	1	2	3	5	6
Cost \$:	5	8	11	17	20

1. Enter tiles as the explanatory variable and cost as the response variable.

$$r = 1$$

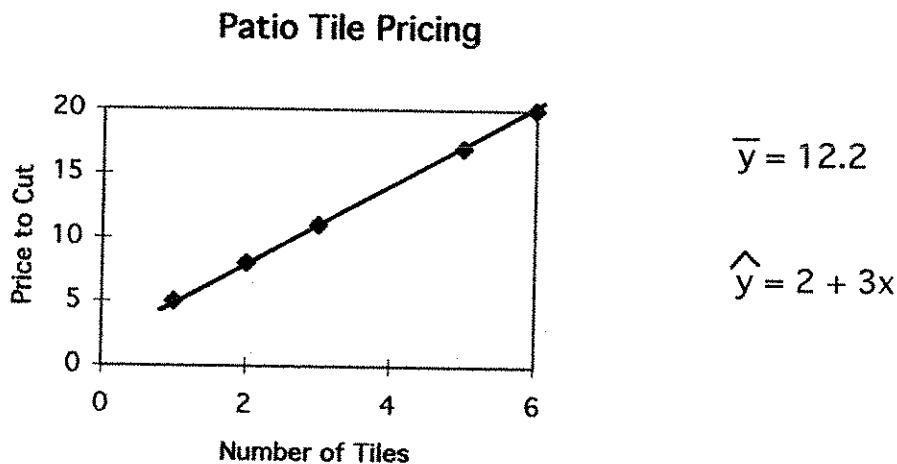
Test Statistic $t = 1E99$ (think about this) $CV_{0.05}: t = \pm 3.182$

P-Value = 0 (think about this)

Table A-6 CV: $r = \pm 0.878$

2. Conclusion: Significant Positive Linear Correlation

3.



4. Coefficient of Determination $r^2 = 1$

100% of the variation in the price of tiles can be explained by the number of tiles cut.

5. Construct a 95% prediction interval for 4 tiles cut.

CV: 3.182 $X_0 = 4$ $y \approx 14$ $E \approx 0$ <-- (think about this)

$$14 < Y < 14$$

(4) The following data represents the living space in square footage (hundreds of square feet) and the selling price (in thousands of \$) of houses in Shasta County.

Sq.Footage : 15 38 23 16 16 13 20 24
 Selling Price : 145 228 150 130 160 114 142 265

1. Enter square footage as the explanatory variable and selling price as the response variable.

$$r \approx 0.717$$

Test Statistic $t \approx 2.517$ $CV_{0.05}: t = \pm 2.447$

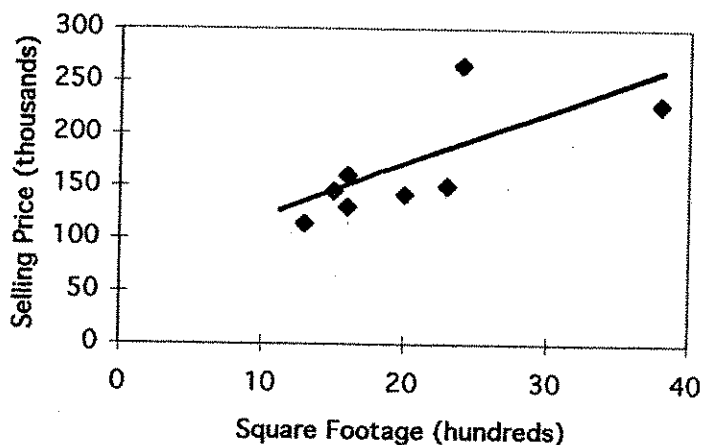
P-Value = 0.045

Table A-6 CV: $r = \pm 0.707$

2. Conclusion: **Significant Positive Linear Correlation**

3.

Shasta County Home Prices



$$\bar{y} = 166.75$$

$$\hat{y} = 71.0 + 4.6x$$

4. Coefficient of Determination $r^2 \approx 0.514$

51.4% of the variation in the selling price of a house can be explained by the variation in the square footage.

5. Construct a 95% prediction interval for a house with 1,800 square feet.

CV: 2.447 $X_0 = 18$ $y \approx 154.57$ $E \approx 102.41$

$$52.16 < Y < 256.97$$

(5) A study was conducted to investigate any relationship between age (in years) and BAC (blood alcohol concentration) measured when convicted DWI jail inmates were first arrested.

Age (in yrs) : 17.2 43.5 30.7 53.1 37.2 21.0 27.6 46.3
 BAC : 0.19 0.20 0.26 0.16 0.24 0.20 0.18 0.23

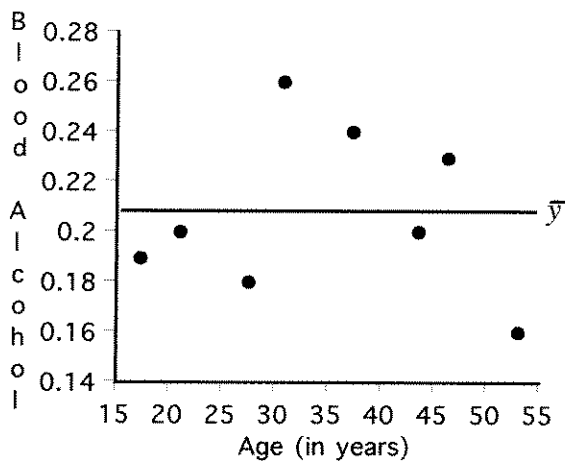
1. Enter age as the explanatory variable and blood alcohol concentration (BAC) as the response variable.

$$r \approx -0.069$$

Test Statistic $t \approx -0.170$ $CV_{.05}: t = \pm 2.447$
 P-Value = 0.871

2. Conclusion: No Linear Correlation

3.



$$\bar{y} = 0.2075$$

$$\hat{y} = 0.21 - 1.822x$$

4. Coefficient of Determination $r^2 \approx 0.0048$

0.5% of the variation in the blood alcohol concentration can be explained by the variation in the age.

5. Construct a 95% prediction interval for a house with 1,800 square feet.

Don't bother to construct a prediction interval. There is no linear correlation. If we want a predicted value use \bar{y} .

(6) The accompanying table lists weights (in hundreds of pounds) and highway fuel usage rate (in miles per gallon) for a sample of domestic new cars. Based on the results, can you expect to pay more for gas if you buy a heavier car?

Weight :	29	35	28	44	25	34	30	33	28	24
Fuel :	31	27	29	25	31	29	28	28	28	33

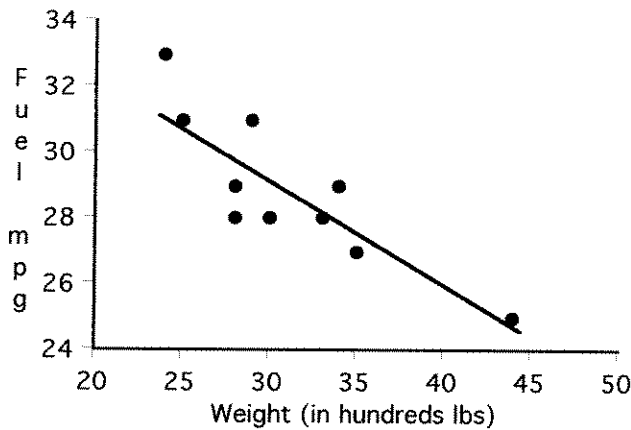
1. Enter the weight as the explanatory variable and fuel consumption as the response variable.

$$r \approx -0.851$$

Test Statistic $t \approx -4.592$ $CV_{.05}: t = \pm 2.306$
 P-Value = 0.0018

2. Conclusion: Significant Negative Linear Correlation

3. Graph



$$\bar{y} = 28.9$$

$$\hat{y} = 39.23 - 0.33x$$

4. Coefficient of Determination $r^2 \approx 0.725$

72.5% of the variation in the fuel consumption of a car can be explained by the variation in the weight of the car.

5. Construct a 95% prediction interval for a car weighing 2,700 lbs.

CV: $t = 2.306$ $x_0 = 27$ $\hat{y} \approx 30.23$ $E \approx 3.14$

$$27.09 < Y < 33.38$$

(7) A *Raiders of the Lost Ark* pinball machine is used to measure learning that results from repeating manual functions. Subjects were selected so that they are similar in important characteristics of age, gender, intelligence, education and so on. We expect that there should be an association between the number of games played and the pinball score. Is there sufficient evidence to support the claim that there is such an association?

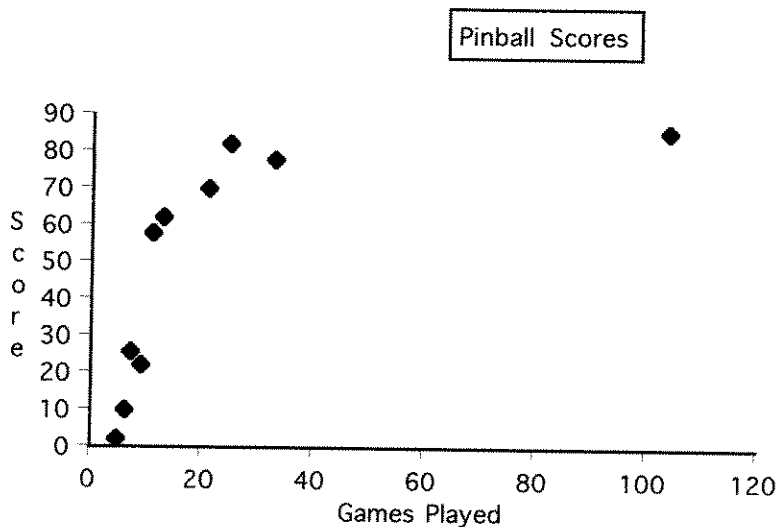
Number of games played:	9	13	21	5	6	25	7	33	11	104
Score :	22	62	70	2	10	82	26	78	58	86

1. Enter the number of games played as the explanatory variable (x) and the score and the response variable (y).

$$r \approx 0.629 \quad P\text{-Value} \approx 0.0512$$

at a significance level of $\alpha = 0.05$ *Fail to Reject the Null*
No Linear Correlation

2. Looking at the graph we can see a positive association but not a linear one. This problem might better be tackled using Rank Correlation. (Look at the graph.)



3. Using Rank Correlation techniques

$$r_s \approx 0.976 \quad CV_{r_s} = \pm 0.648 \text{ (Table A-6)}$$

Reject the Null Hypothesis

There is Significant Correlation

Higher numbers of games played appear to be associated with higher scores.

(8) *Blood Pressure Measurements* Fourteen different second-year medical students took blood pressure measurements of the same patient and the results are listed below (data provided by Marc Triola, MD).

Systolic	138	130	135	140	120	125	120	130	130	144	143	140	130	150
Diastolic	82	91	100	100	80	90	80	80	80	98	105	85	70	100

Find the: (a) explained variation (b) unexplained variation (c) total variation
 (d) coefficient of determination (r^2) (e) standard error of estimate (s_e).
 (Assume that there is significant linear correlation.)

Enter your data into L1 and L2
 then do E:LinRegTTest...

L1	L2	L3	Z
138	82		
130	91		
135	100		
140	100		
120	80		
125	90		
120	80		
130	80		
130	80		
144	98		
143	105		
140	85		
130	70		
150	100		

Calculate the total variation $\sum (y - \bar{y})^2$

Cursor on L3

L1	L2	L3	Z
138	82		
130	91		
135	100		
140	100		
120	80		
125	90		
120	80		
130	80		
130	80		
144	98		
143	105		
140	85		
130	70		
150	100		

L1	L2	L3	Z
138	82		
130	91		
135	100		
140	100		
120	80		
125	90		
120	80		
130	80		
130	80		
144	98		
143	105		
140	85		
130	70		
150	100		

LIST	MATH	5:Sum(
SUM(L3)		
1453.214286		
Ans→T		

Total Variation ≈ 1453.214286

Store in T

Calculate the Explained Variation:

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \text{Explained Variation} = r^2 \cdot \text{Total Variation}$$

r ² *T	628.9602705
Ans→E	

Store in E

Calculate the Unexplained Variation:

$$\text{Unexplained Variation} = \text{Total Variation} - \text{Explained Variation}$$

T-E	824.2540152
Ans→U	

Store in U

ANSWERS

- (a) Explained variation is stored in E
- (b) Unexplained variation is stored in U
- (c) Total variation is stored in T
- (d) Coefficient of determination
- (e) Standard error of estimate

$\text{Explained Variation} \approx 628.96027$

$\text{Unexplained Variation} \approx 824.25402$

$\text{Total Variation} \approx 1453.2143$

$r^2 \approx 0.43280628$

$s_e \approx 8.287812413$

VAR	5:Statistics	EQ	8:r ²
VAR	5:Statistics	TEST	0:s