Introductory Statistics Lectures

Linear correlation

Testing two variables for a linear relationship

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

(Compile date: Tue May 19 14:51:18 2009)

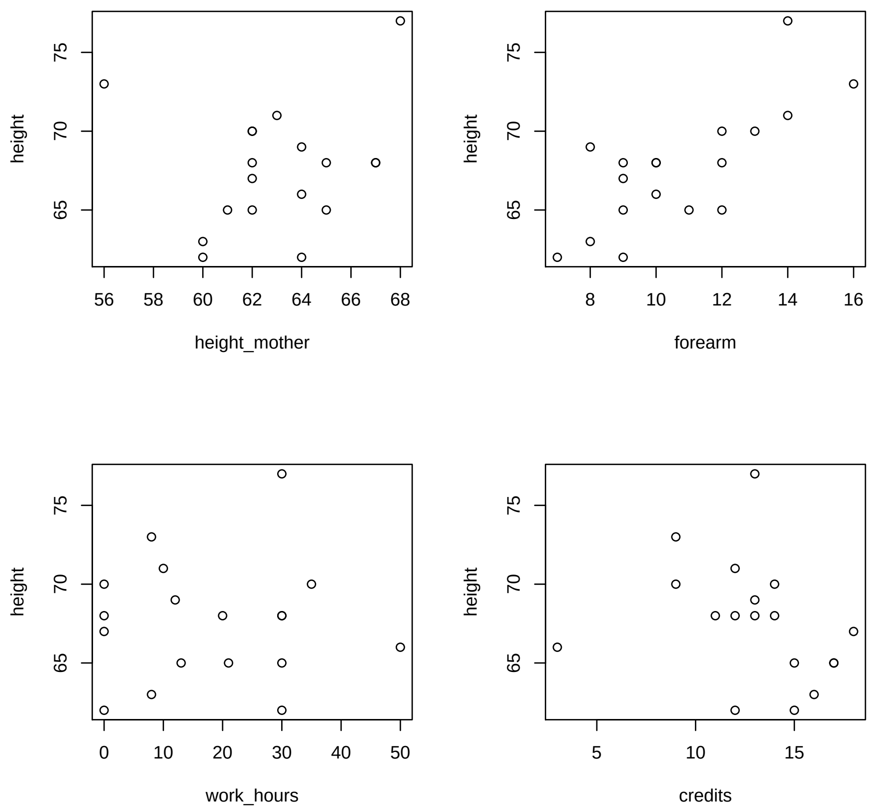
Contents

1 Linear correlation

1.1 Introduction

Motivation

Is there a relationship — correlation — between your height and ... (1) your mother's height? (2) your forearm height? (3) your work hours per week? (4) your commute distance?



Motivation

Example 1. How much of a individual's height is explained by their mother's height? Use our class data to determine if there is a linear relationship between a mother's height and their child's height (your height) and how much variation in the child's height can be explained by the mother's height.

```
R: height = class.data$height
R: height_mother = class.data$height_mother
```

The first few data points are:

Use a scatter plot to see if a relationship exists

```
R: plot(height_mother, height, main = "Height in inches")
```

	height	height_mother
1	65	65
2	68	67
3	71	63
4	66	64
5	68	65
6	65	62

Question 1. Does it look like there is a linear relation ship? Draw a best fit line in the data

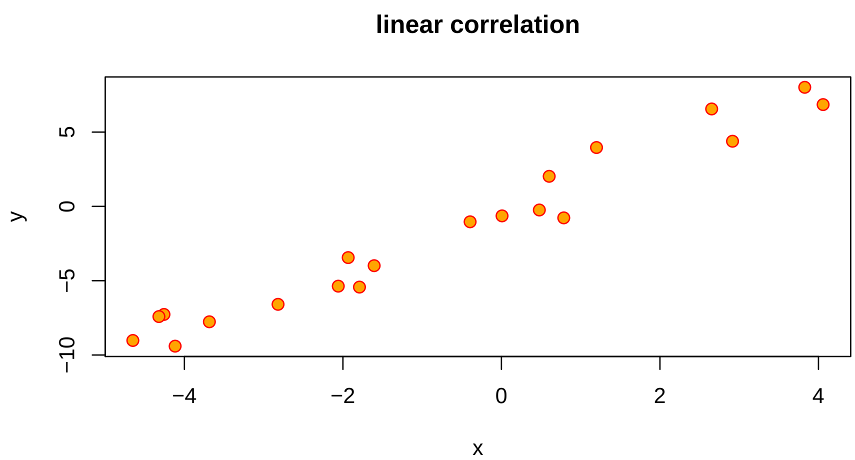
Paired data. Definition 1.1

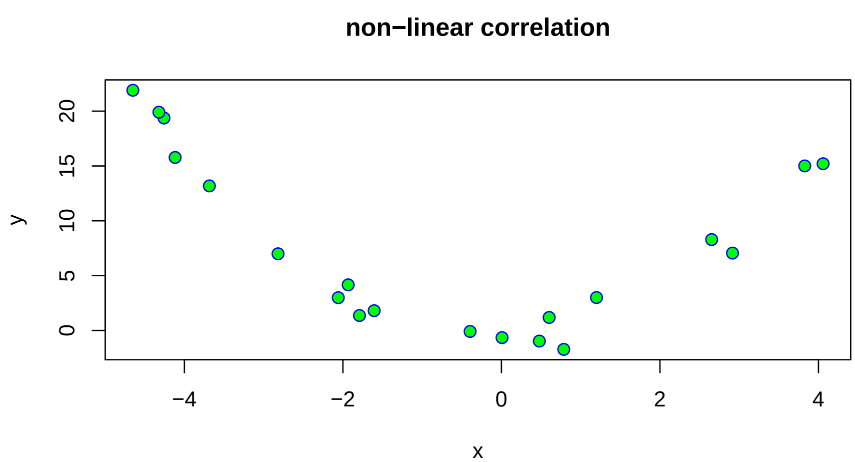
A set of (x_i, y_i) data where each pair is **related**. (Dependent samples.)
ex: mother height, child height.

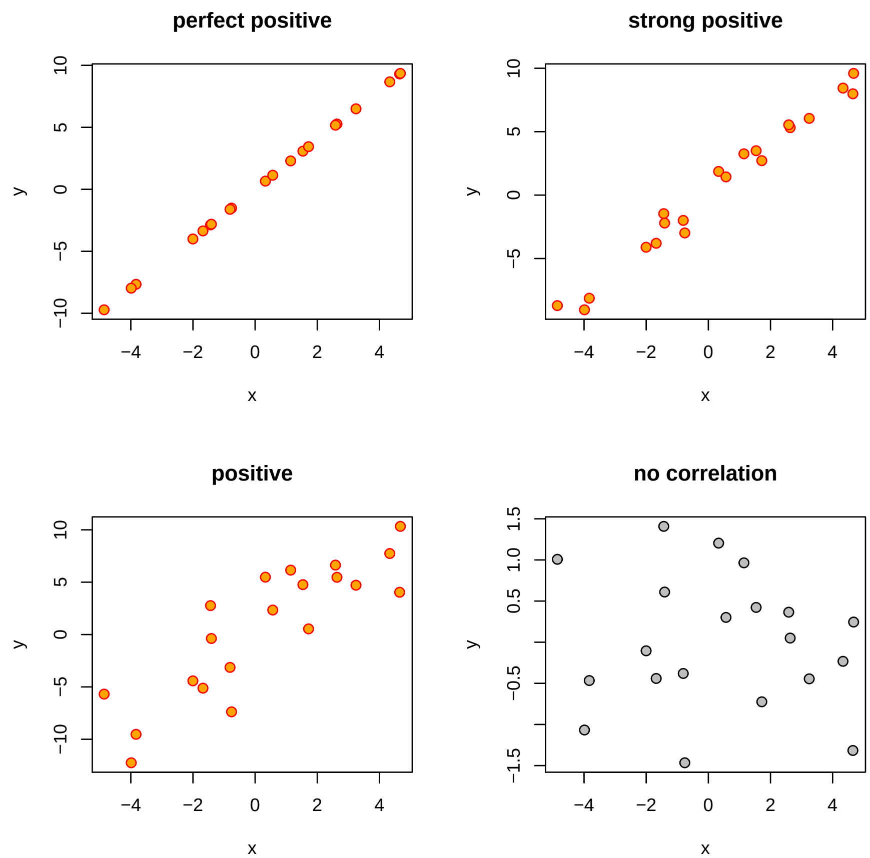
1.2 Linear Correlation

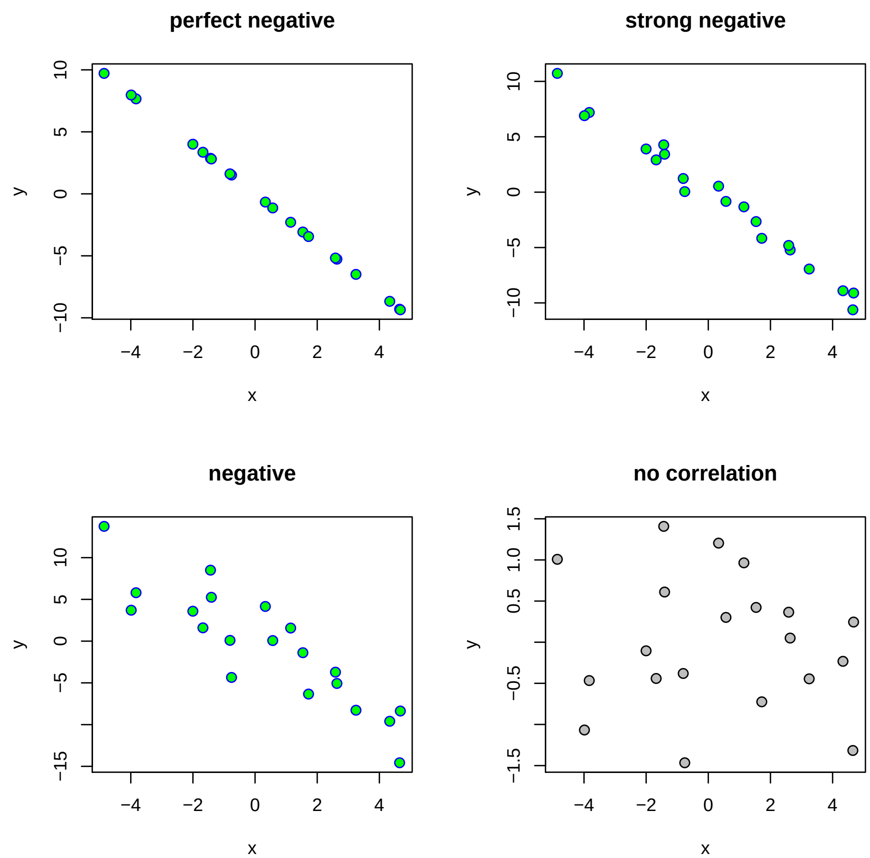
Correlation. Definition 1.2

exists when two variables have a relationship with one another.


linear correlation
y
x
5
0
−5
−10
−4
−2
0
2
4


non−linear correlation
y
x
20
15
10
5
0
−4
−2
0
2
4


perfect positive
strong positive
positive
no correlation
x
y



USE

Often used to help answer:

1. Is there a **linear** relationship between X and Y?
2. Can X be used to predict Y?
3. How much of the variation in X can be predicted with Y?

COMPUTATION

DEFINITION 1.3

LINEAR CORRELATION COEFFICIENT.

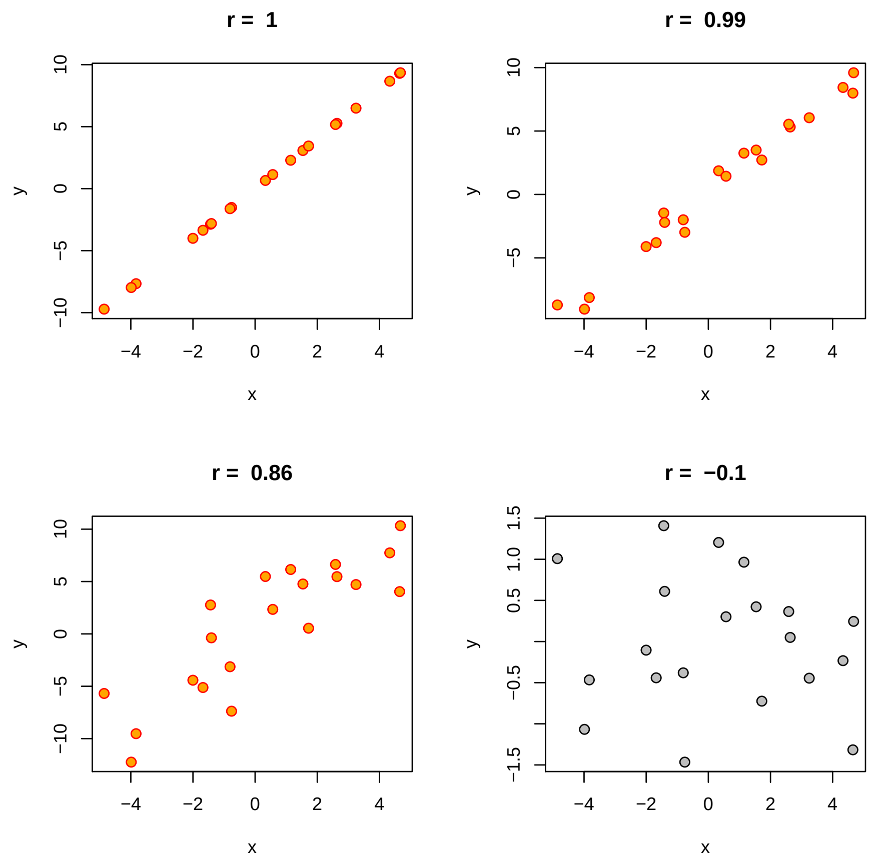
The linear correlation coefficient for a population is denoted with ρ. We can estimate ρ via a sample and calculate Pearson's linear correlation coefficient r:

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

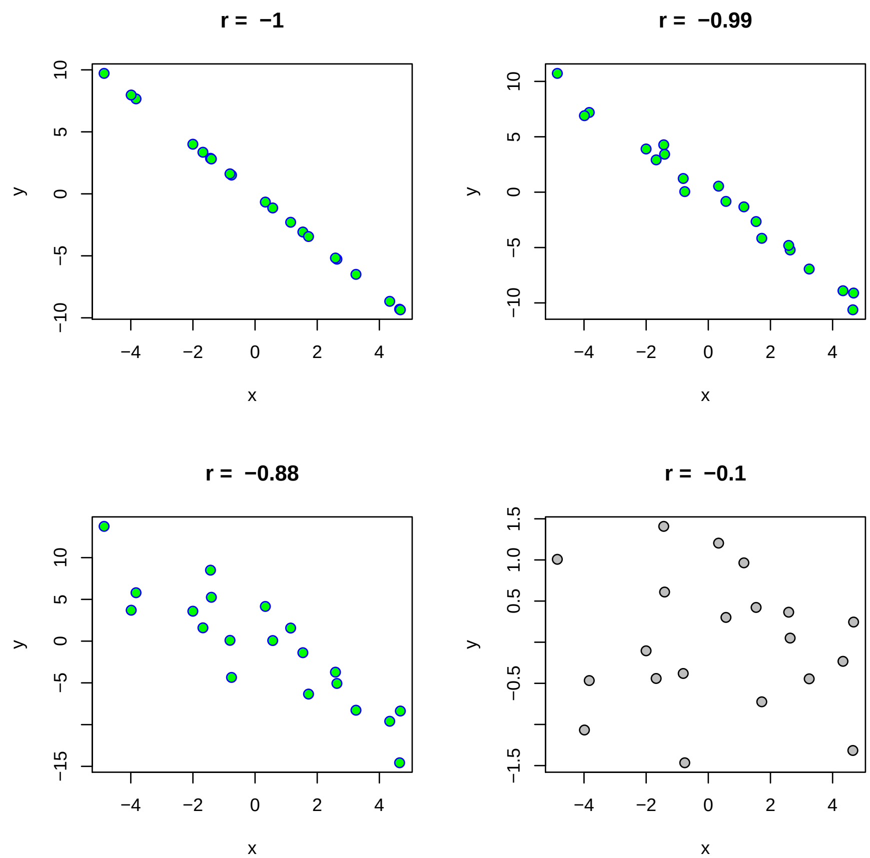
n is the number of **pairs** of data points (length of x or y).

- Measures the **strength** of the linear relationship between x and y.
- Larger values of $|r|$ indicate stronger linear relationship.[1]
- Positive r indicates positive slope, negative r indicates negative slope.

Examples of r



[1] Larger $|r|$ does not indicate a steeper slope. We will find the slope later using regression.



Properties of r (ρ for populations)

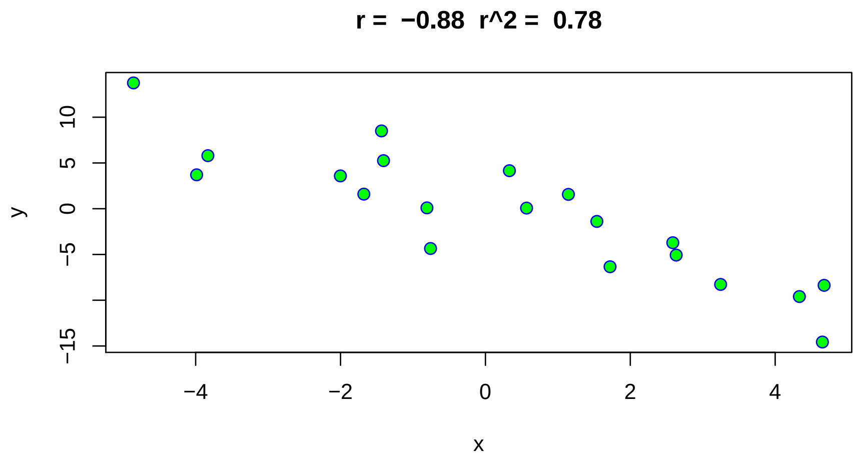
1. $-1 \leq r \leq +1$
2. r is scale invariant.
3. r is invariant if x and y are interchanged.
4. r **only** measures the **strength** of **linear** relationships.

DEFINITION 1.4 COEFFICIENT OF DETERMINATION (EXPLAINED VARIATION).

r^2 is the proportion of linear variation in y that is explained by x.

- $0 \leq r^2 \leq 1$
- The closer r^2 is to 1 the stronger the linear relationship and likewise the more variation in y that can be explained by x.

Example of r^2



HYPOTHESIS TEST FOR LINEAR CORRELATION. DEFINITION 1.5

requirements (1) simple paired (x, y) random samples, (2) Pairs of (x, y) have a bivariate normal distribution[2], (3) correlation is linear.

null hypothesis $\rho = 0$ (no linear correlation)

alternative hypothesis $\rho \neq 0$ (a linear correlation exists[3])

Always make a scatter plot first to see if the relationship is linear.

LINEAR CORRELATION COEFFICIENT r AND HYPOTHESIS TEST:

`cor.test(x, y)`

Calculates r from the sample and conducts the hypothesis test for $H_0 \rho = 0$.

x vector of **ordered** x data.

y vector of **ordered** y data.

R COMMAND

Test statistic for linear correlation coefficient

$$t = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}} \tag{2}$$

where $df = n - 2$.

Note: n is number of pairs.

Procedure for finding r

1. Define **two ordered vectors** (x and y) with the data.
2. Make a **scatterplot** to determine if a linear relationship exists.
3. If a linear relationship exists, run the **hypothesis test** `cor.test()` and do all 7 steps. It will give you a point estimate of ρ (which is r) and allow you to determine if it is significant (via the p-value).

[2] Effectively, a normal distribution for x and y.

[3] Alternative hypothesis involving $<$ or $>$ also possible and work just as before.

A COMPLETE EXAMPLE

Example of calculating r and testing significance

Back to our original example:

Example 2. How much of a individual's height is explained by their mother's height? Use our class data to determine if there is a linear relationship between a mother's height and their child's height (your height) and how much variation in the child's height can be explained by the mother's height.

```
R: height
 [1] 65 68 71 66 68 65 62 68 77 62 69 70 65 63 67 73 68 70
R: height_mother
 [1] 65 67 63 64 65 62 60 62 68 64 64 62 61 60 62 56 67 62
```

First check if relationship looks linear

```
R: plot(height_mother, height, main = "Height in inches")
```

Question 2. Does is look line a linear relationship? (Strong or weak?)

Question 3. What sign do we expect for r?

Question 4. What is our null hypothesis?

Run hypothesis test...

$H_0 : \rho = 0$, $H_a : \rho \neq 0$, $\alpha = 0.05$.

```
R: res = cor.test(height_mother, height)
R: res
        Pearson's product-moment correlation

data:  height_mother and height
t = 0.7725, df = 16, p-value = 0.4511
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.30417  0.60311
sample estimates:
    cor
0.18963
```

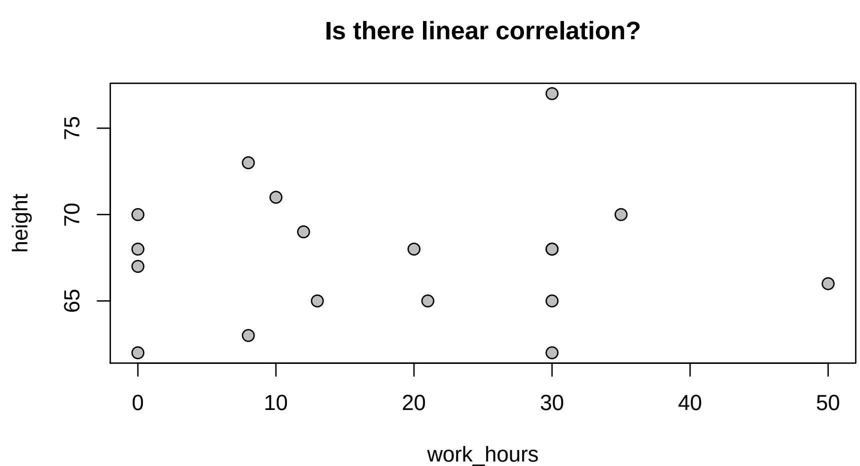
Question 5. What is the formal decision?

Question 6. What is r?

Question 7. What is the percent of variation of an individual's height is explained by their mother's height?

Question 8. What is the formal conclusion?

Importance of testing significance of r



The linear correlation coefficient for the above data is 0.0388.

Question 9. Why is it important to test the hypothesis $H_0 : \rho = 0$ and $H_a : \rho \neq 0$ since r is not zero?

```
R: cor.test(work_hours, height)
        Pearson's product-moment correlation

data:  work_hours and height
t = 0.1553, df = 16, p-value = 0.8785
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.43597  0.49667
sample estimates:
     cor
0.038799
```

Since we fail to reject $H_0 : \rho = 0$, r is not significant. Sampling error can account for its deviation from the null value of zero.

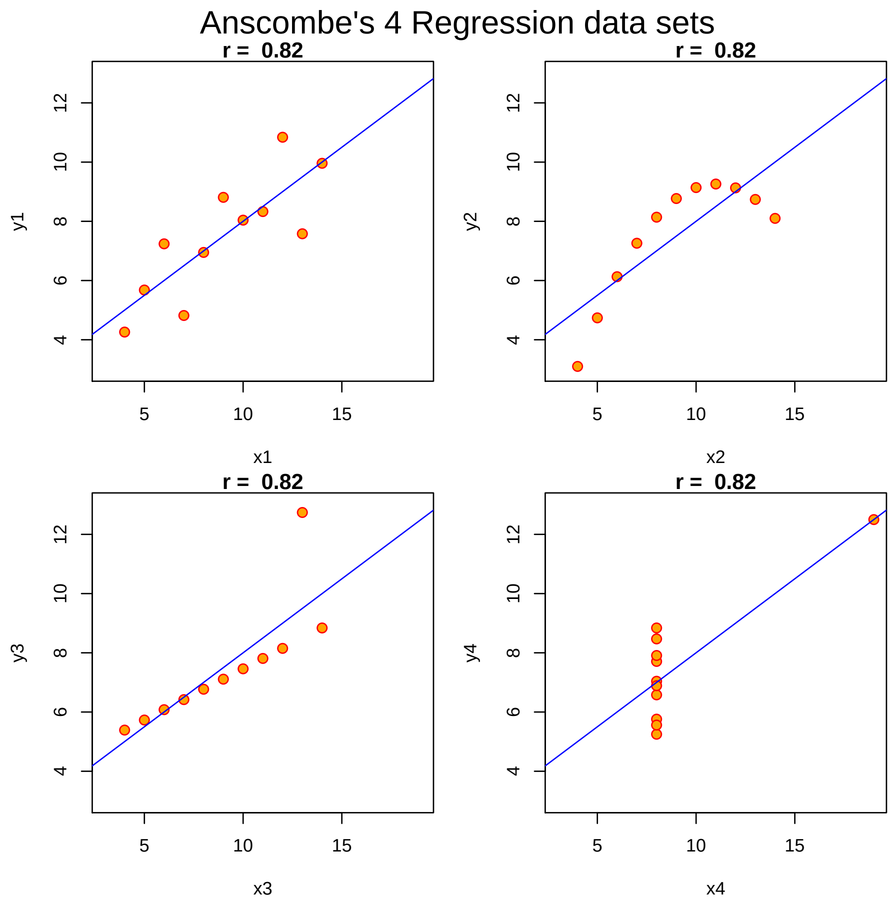
Conclusion: no significant liner correlation.

CAUTIONS

Cautions when calculating and interpreting correlation

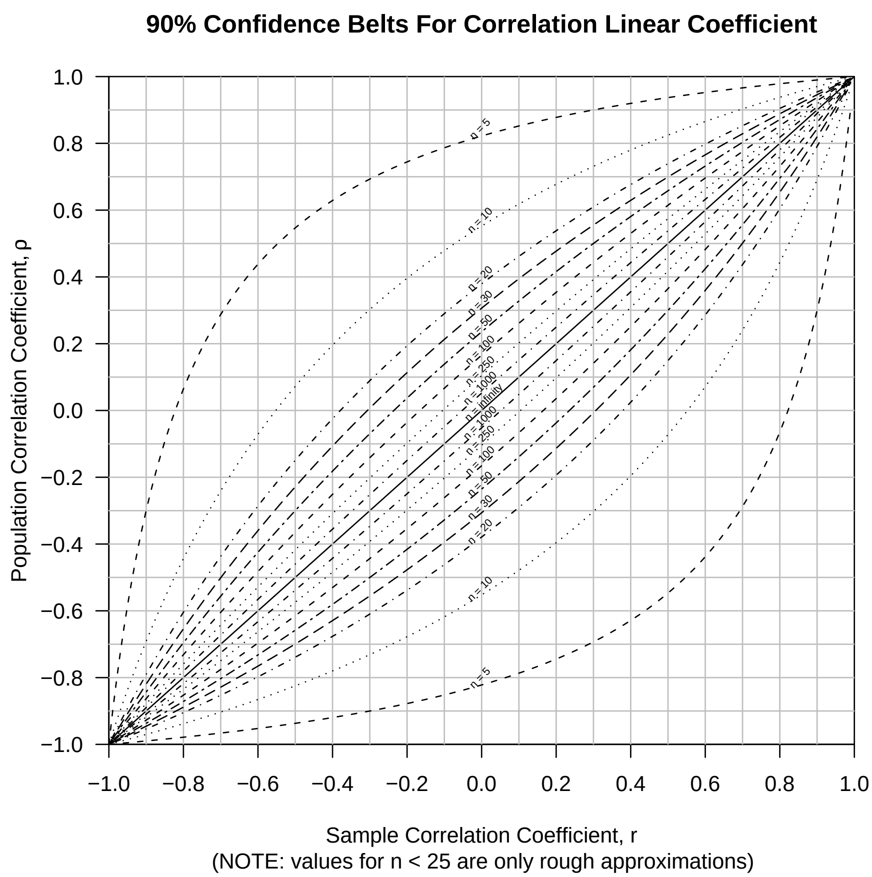
- If the relationship is nonlinear, r is not meaningful!
- Correlation does not imply causality!
- Calculating correlation based on averages may falsely inflate r.
- If no significant linear correlation exists, a relationship may still exist (namely, a nonlinear one).

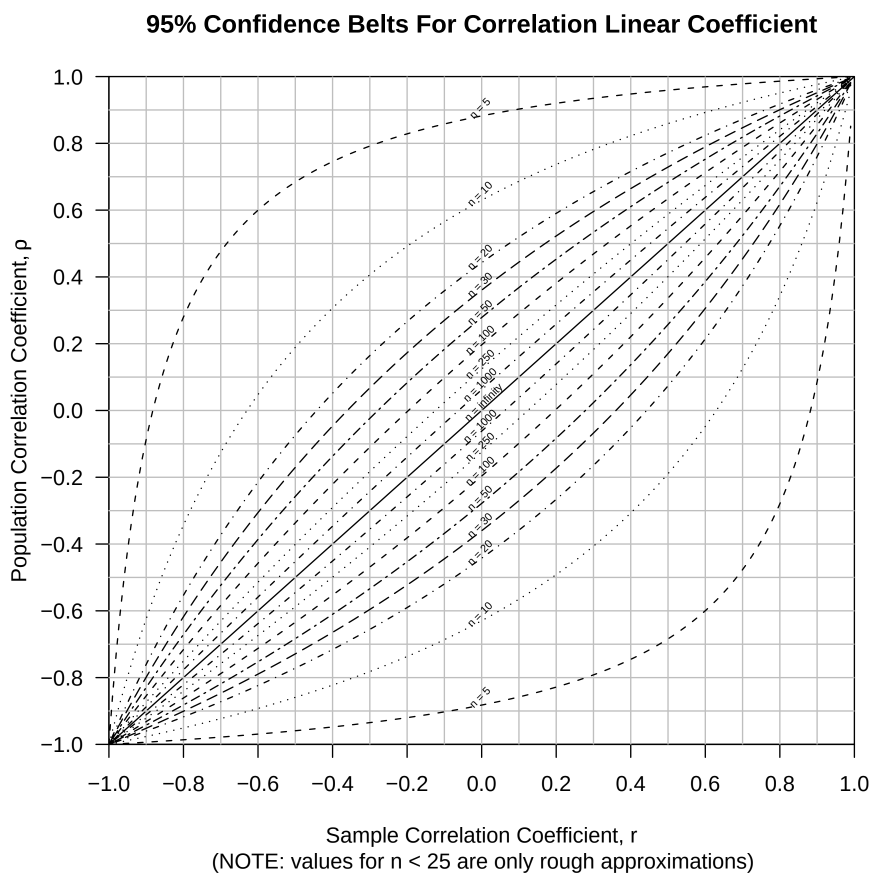
When relationship is not linear

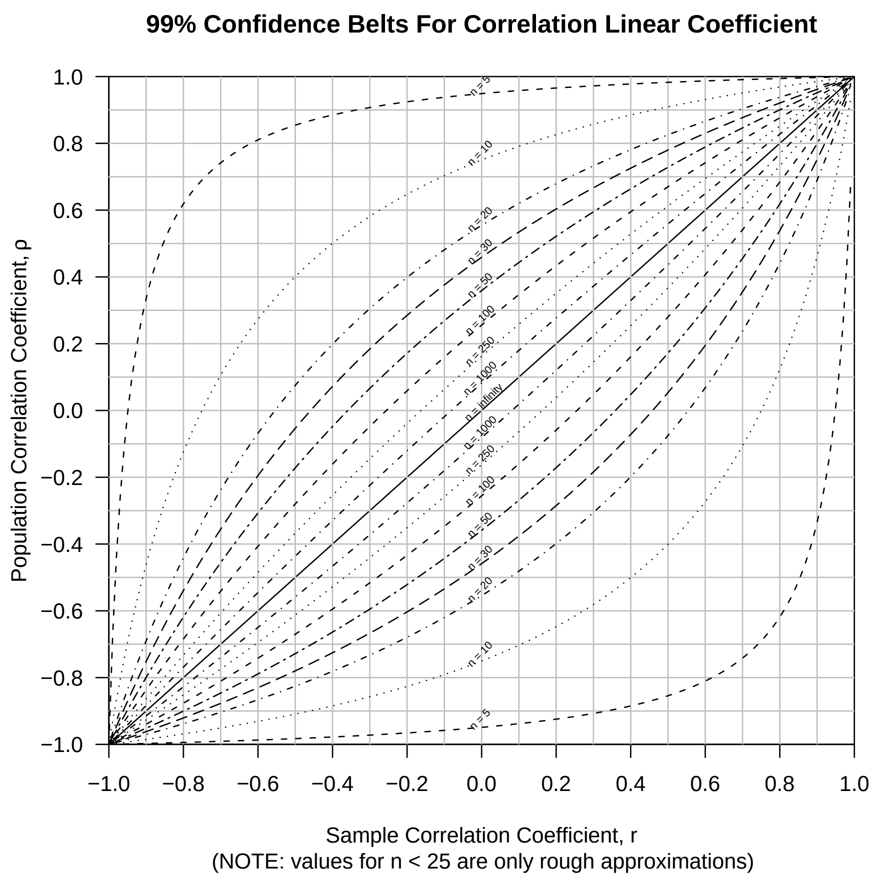


Note that r is the same for all plots, but not appropriate for all!

CONFIDENCE INTERVAL BELT GRAPHS


90% Confidence Belts For Correlation Linear Coefficient
Population Correlation Coefficient, ρ
Sample Correlation Coefficient, r
(NOTE: values for n < 25 are only rough approximations)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
n = 5
n = 10
n = 20
n = 30
n = 50
n = 100
n = 250
n = 1000
n = infinity


95% Confidence Belts For Correlation Linear Coefficient
Population Correlation Coefficient, ρ
Sample Correlation Coefficient, r
(NOTE: values for n < 25 are only rough approximations)
n = 5
n = 10
n = 20
n = 30
n = 50
n = 100
n = 250
n = 1000
n = infinity
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0


99% Confidence Belts For Correlation Linear Coefficient
Population Correlation Coefficient, ρ
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
n = 5
n = 10
n = 20
n = 30
n = 50
n = 100
n = 250
n = 1000
n = infinity
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Sample Correlation Coefficient, r
(NOTE: values for n < 25 are only rough approximations)

1.3 Summary

Linear correlation coefficient r

r measures the strength of linear correlation.

1. $-1 \le r \le +1$
2. r is scale invariant.
3. r is invariant if x and y are interchanged.
4. r **only** measures the **strength** of **linear** relationships.

r^2 proportion of linear variation in y that is explained by x.

Finding r check for linear relationship, then find r and make sure it is significant.

Hypothesis test to find r and test significance:

requirements (1) simple paired (x, y) random samples, (2) Pairs of (x, y) have a bivariate normal distribution, (3) correlation is linear.

null hypothesis $\rho = 0$ (no linear correlation)

alternative hypothesis $\rho \neq 0$ (a linear correlation exists.

test in R : `cor.test(x, y)`

x **vector** of **ordered** x data.

y **vector** of **ordered** y data.

You **must** make a scatter plot first!

1.4 Further examples

Now test to see if there is a significant linear correlation between a person's height and their forearm length. (Use the class data set, hint: `height=class.data$height` `forearm=class.data$forearm`)

Question 10. Does the relationship look linear?

Question 11. What is r (Check: $r = 0.753$)

Question 12. Is r significant? (Check: p-value= 0.000311)

Question 13. What is the 95% confidence interval for r (Check:(0.441, 0.902))

Question 14. What percent of an individual's height is explained by their forearm length?

Question 15. Which is better for predicting an individual's height: their forearm length or mother's height?