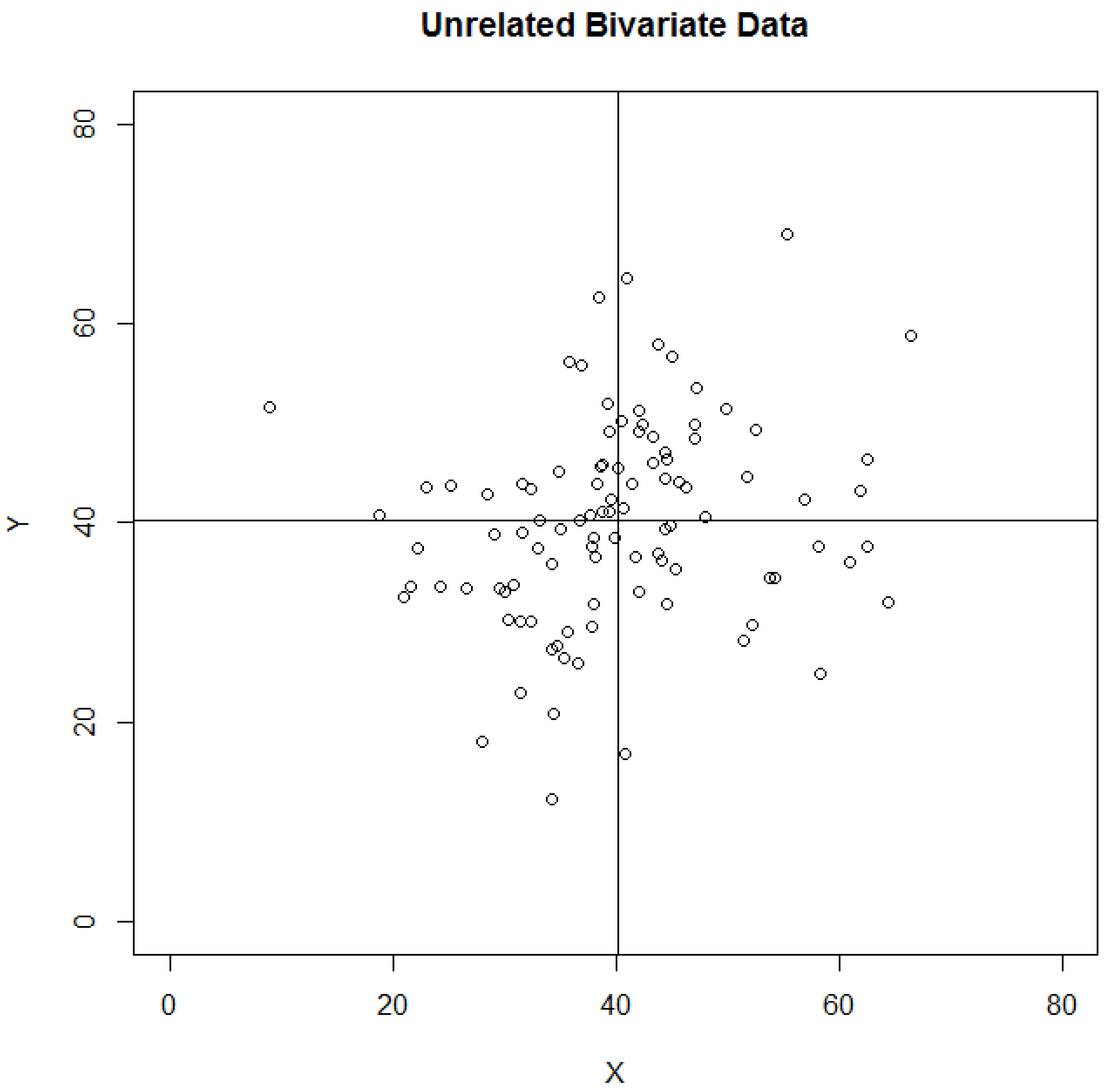
BIM 105
Probability and Statistics for Biomedical Engineers

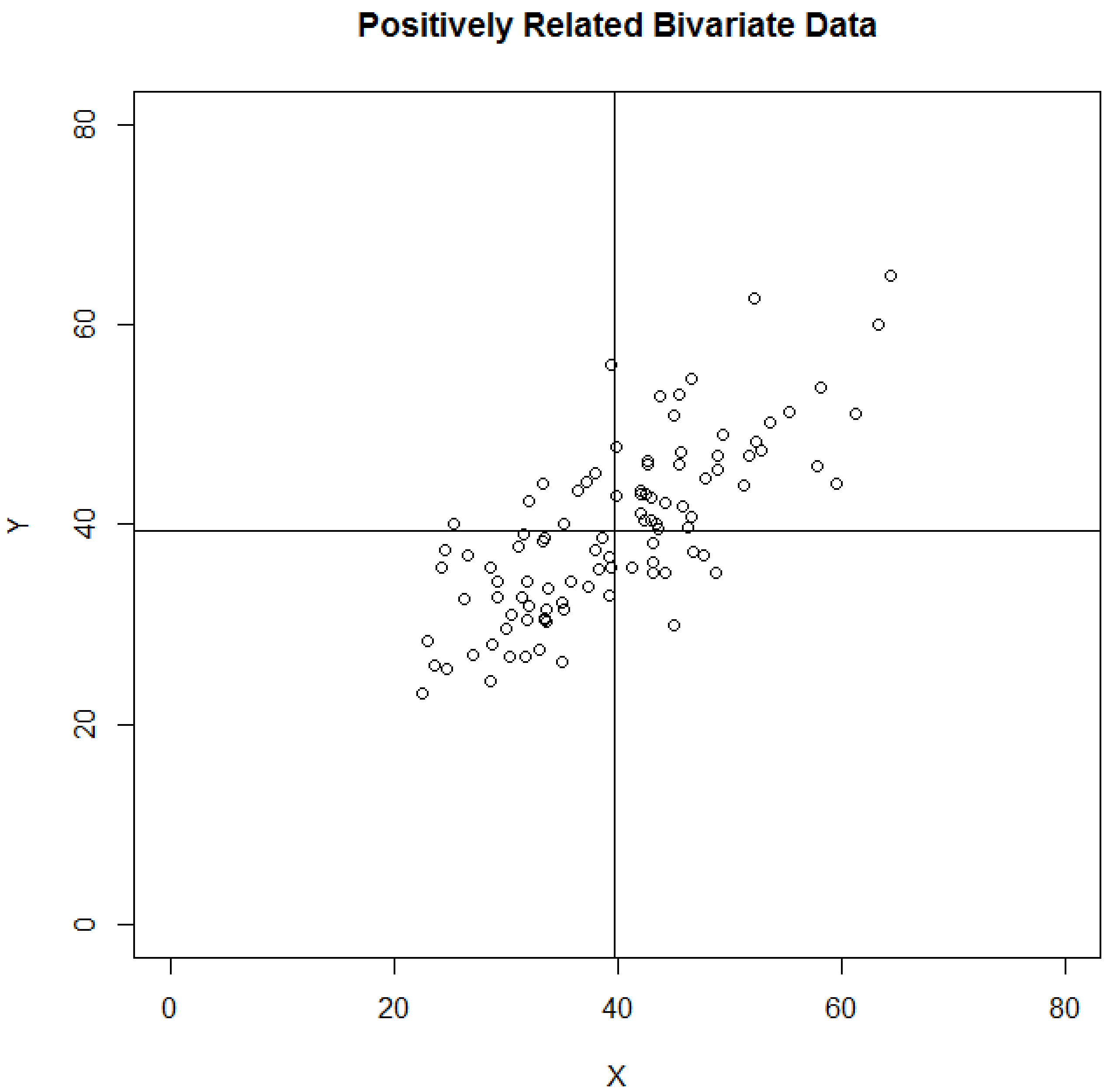
David M. Rocke

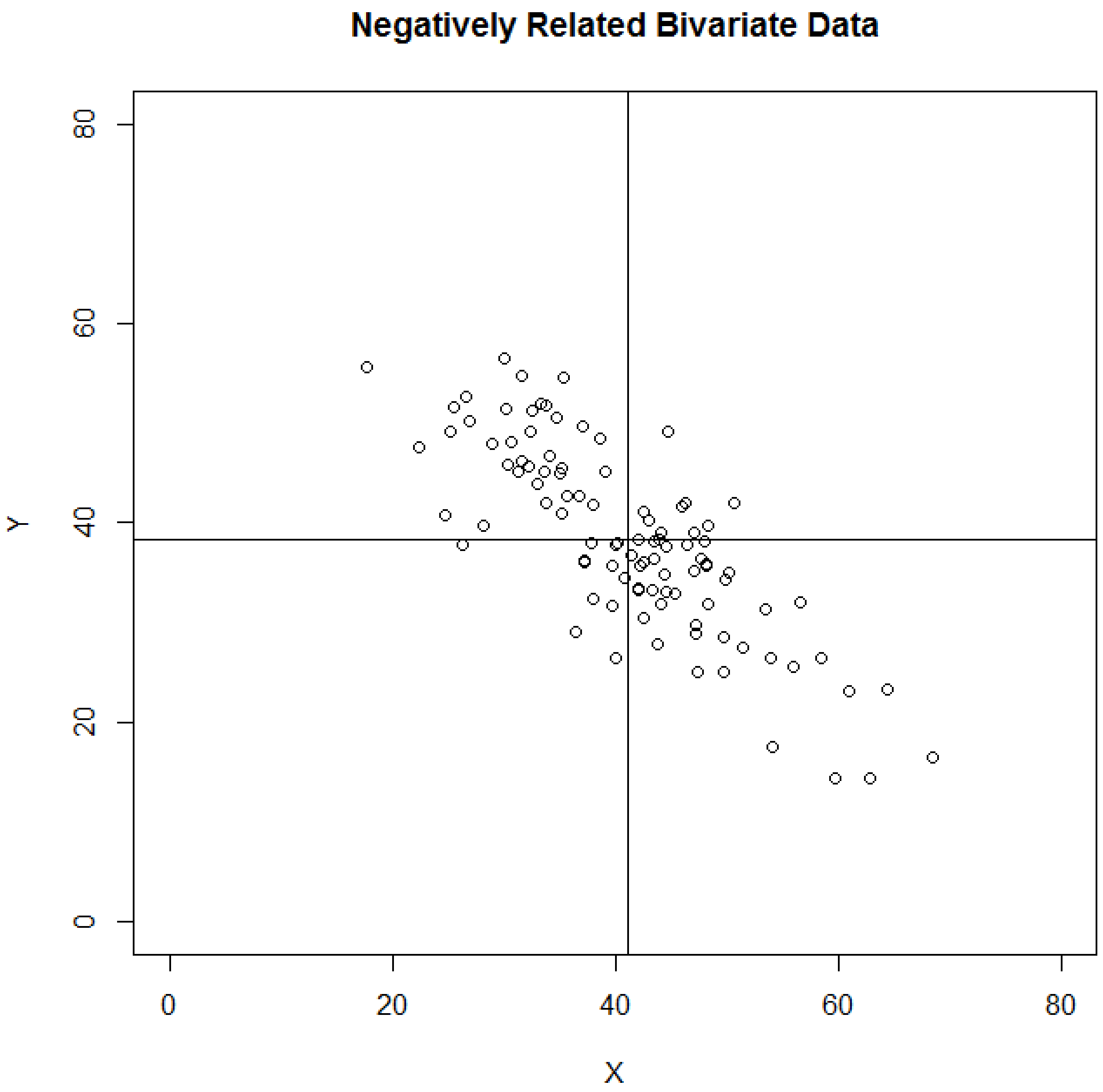
Department of Biomedical Engineering

Summaries for Bivariate Data

- If we have two measurements on each unit in a sample, we call that *bivariate* data.
- For example, we have 17 subjects with measurements
  - X = peak air flow by the standard Wright meter
  - Y = peak air flow by the mini Wright meter
- We have summaries of location and spread for each variable
  - The mean
  - The variance/standard deviation
- Are X and Y "related"?


Unrelated Bivariate Data
Y
0
20
40
60
80
X
0
20
40
60
80


Positively Related Bivariate Data
0
20
40
60
80
Y
0
20
40
60
80
X


Negatively Related Bivariate Data
0
20
40
60
80
0
20
40
60
80
X
Y

Measuring Relatedness

Variance of X

$$S_X^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Variance of Y

$$S_Y^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

Covariance of X and Y

$$S_{XY} = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Product is + when x and y are both above the mean

Product is + when x and y are both below the mean

Product is $-$ when x and y are on opposite sides of the mean

Scaling

Correlation of X and Y

$$\rho_{XY} = (n-1)^{-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{S_X}\right)\left(\frac{y_i-\overline{y}}{S_Y}\right)$$

$$= \frac{(n-1)^{-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{S_X S_Y}$$

$$= \frac{S_{XY}}{S_X S_Y}$$

Scaling

$\left(\dfrac{x_i - \overline{x}}{S_X}\right)$ always has mean 0 and standard deviation 1

It is often said to be *standardized*.

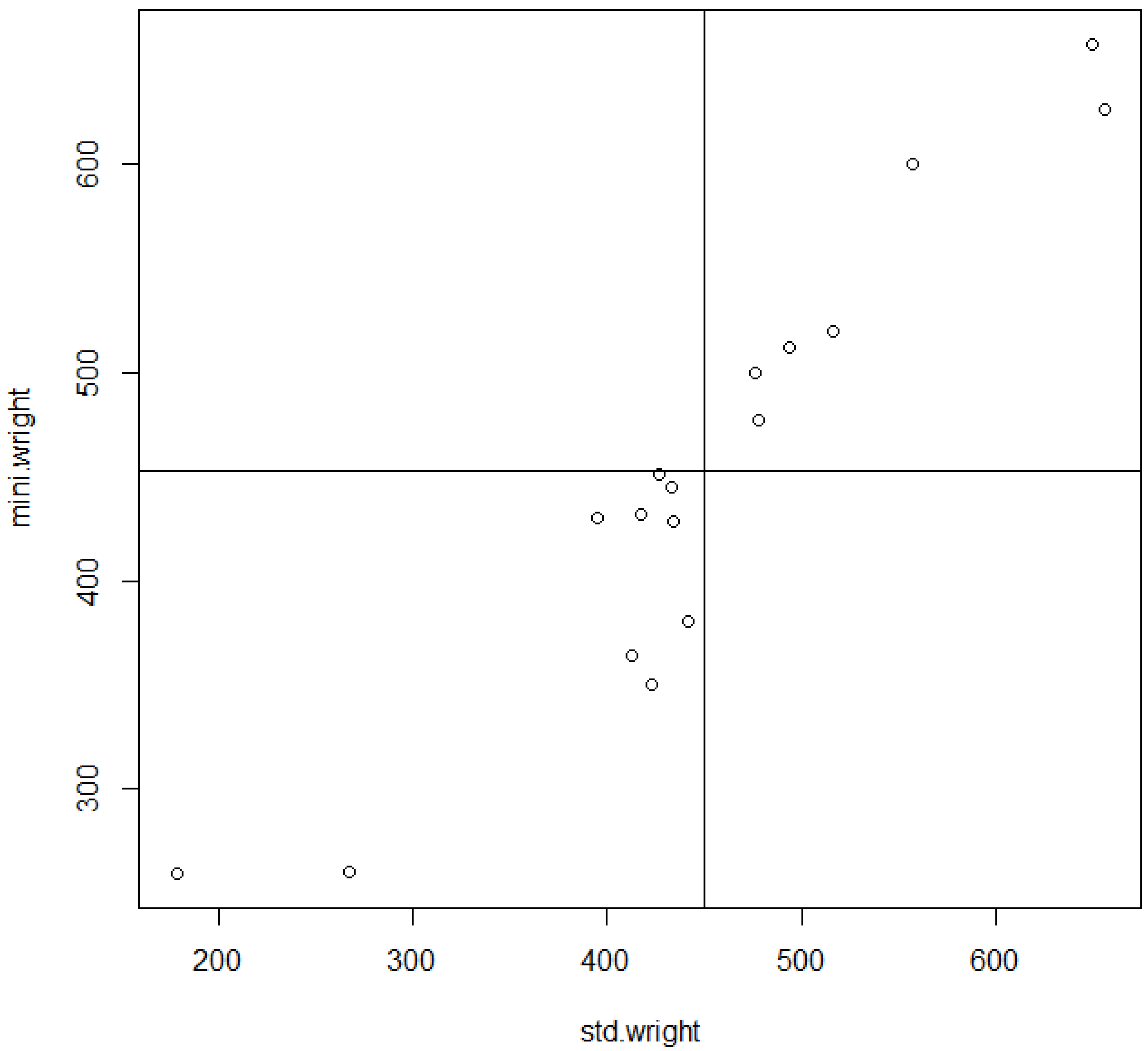
ρ_{XY} is always between -1 and 1

1 if the points all lie on a line with positive slope

–1 if the points all lie on a line with negative slope

0 if the points lie in a circular shape with no elliptical tilt

Correlation is the covariance of standardized variables



mini.wright
600
500
400
300
200
300
400
500
600
std.wright

Correlation for the Wright Meter Data

```
Means
 std.wright mini.wright
   450.3529    452.4706
Variances
 std.wright mini.wright
   13,528.62    12,795.01
Standard Deviations
 std.wright mini.wright
   116.3126    113.1151
Covariance and Correlation
   12410.45
   0.94327945
```

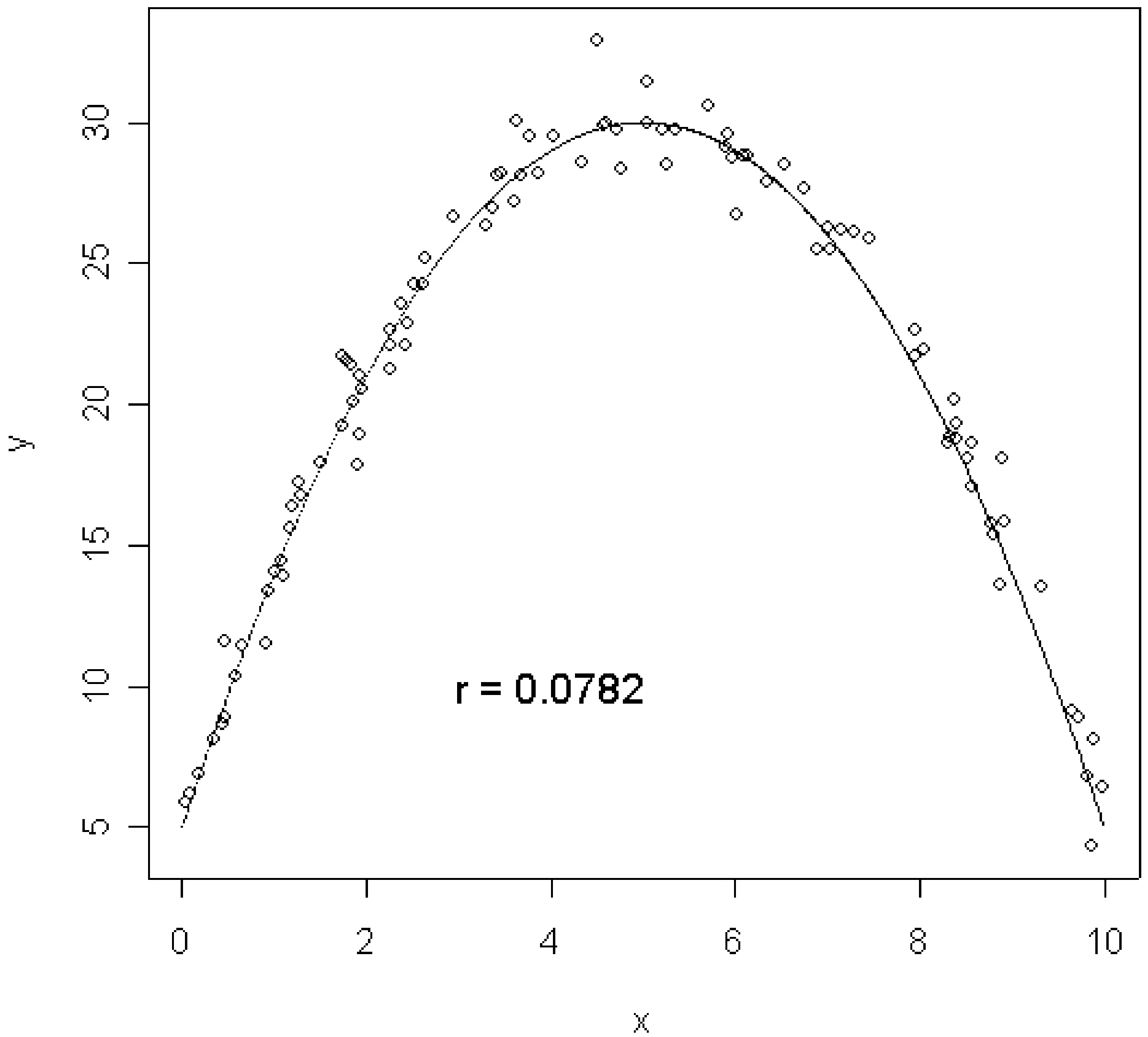
Correlation in MATLAB

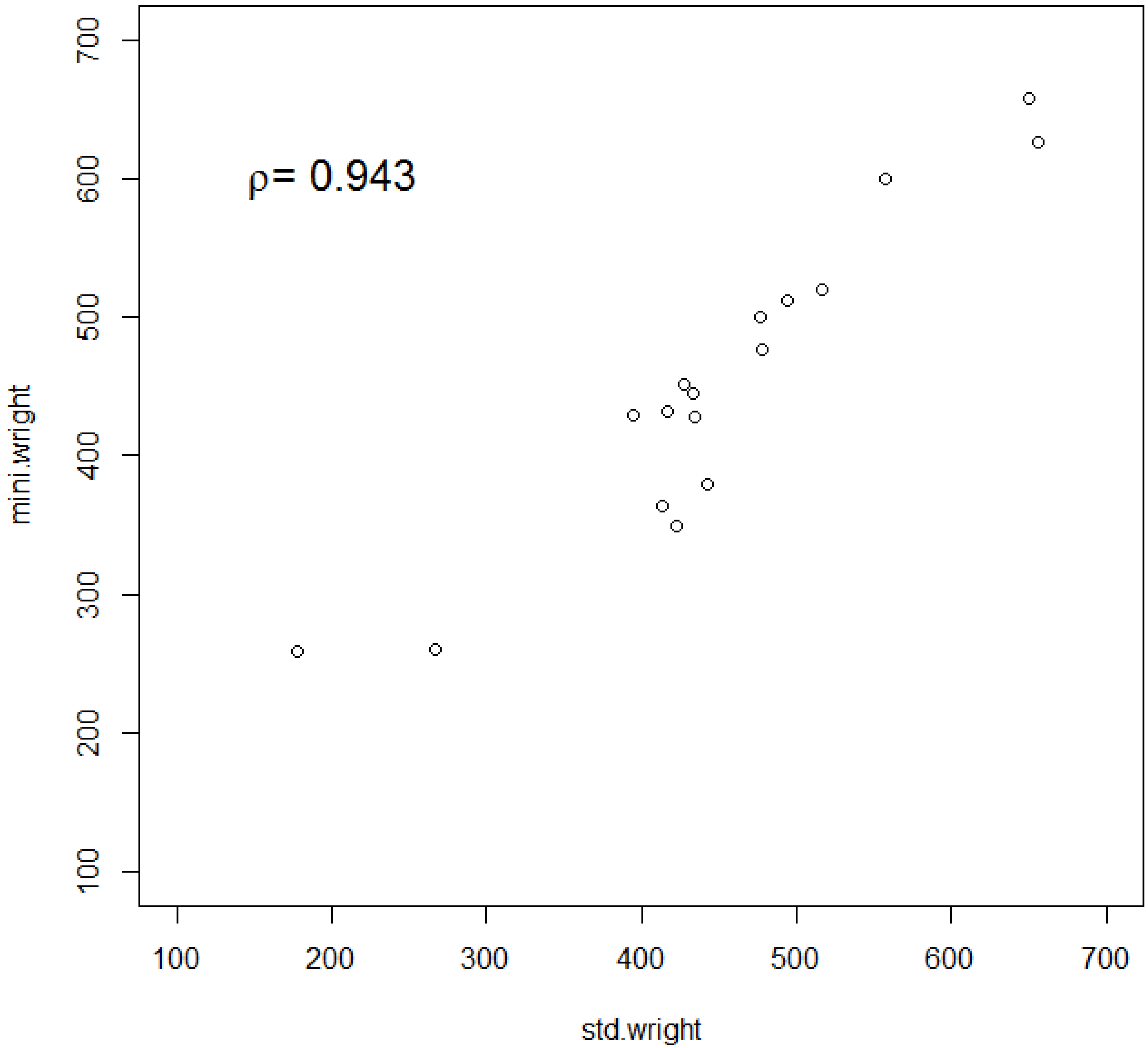
```
>> corrcoef(stdwright,miniwright)

ans =

    1.0000    0.9433
    0.9433    1.0000

>> cov(stdwright,miniwright)

ans =

   1.0e+04 *

    1.3529    1.2410
    1.2410    1.2795
```

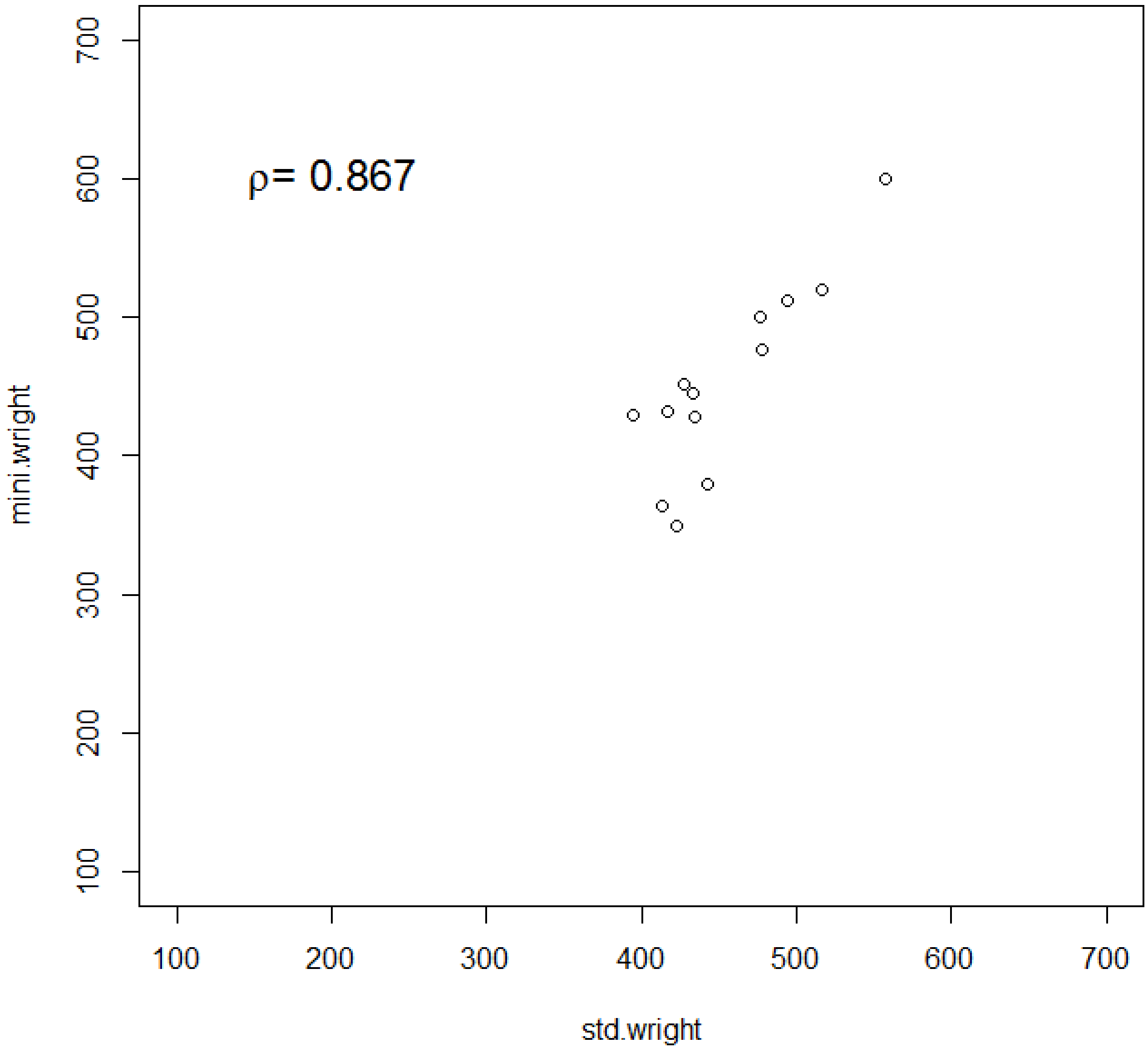
Cautions about Correlation

- The coefficient of correlation measures linear association. If the relationship is non-linear, a more sophisticated measure is needed.
- Correlation depends not only on how close the values in X and Y are, but also on the range of X
- Correlation coefficients can be distorted by outliers
- Correlation does not imply causation (storks do not bring babies)

A strong nonlinear relationship with low correlation

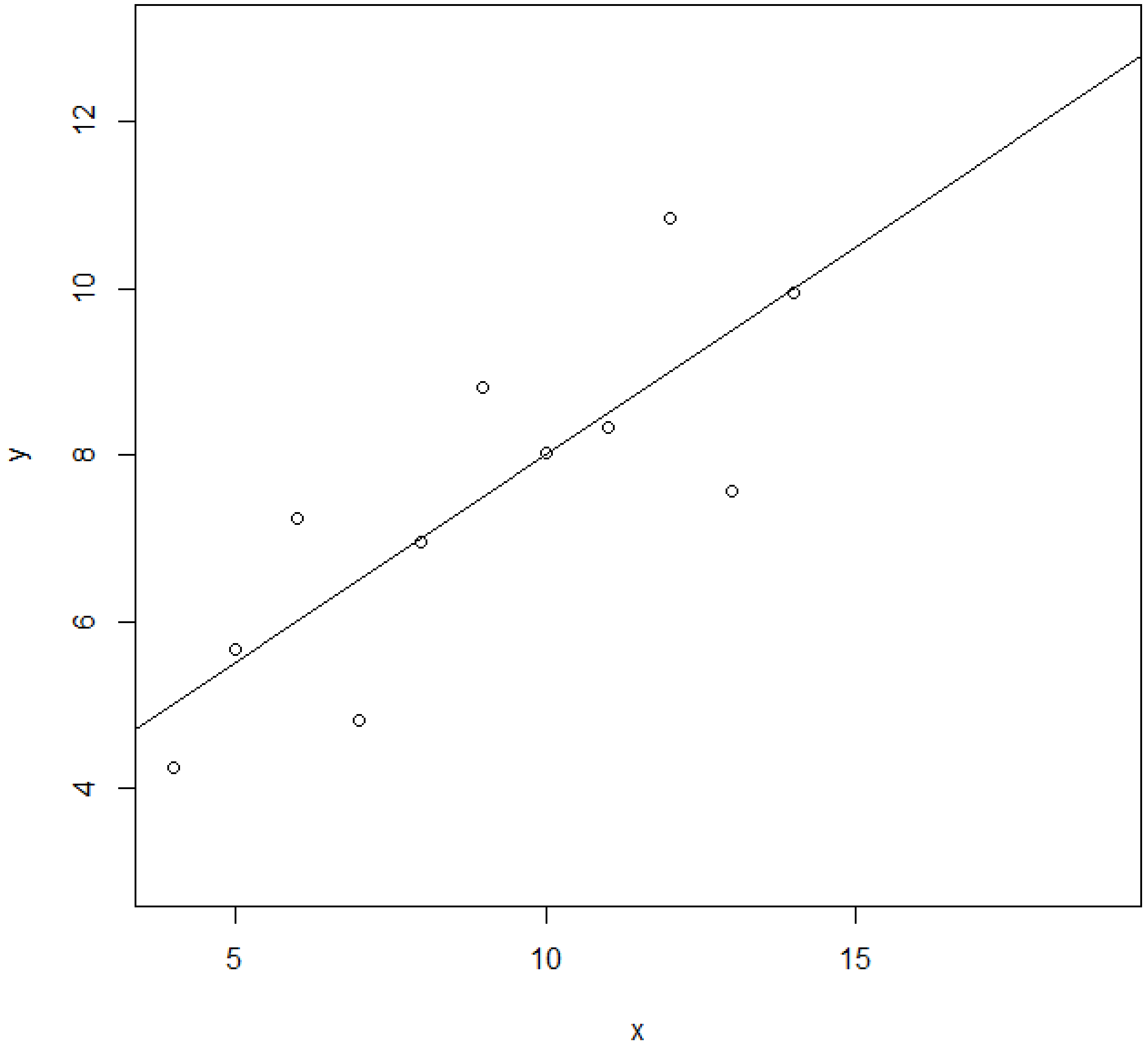



ρ= 0.943
mini.wright
std.wright
100
200
300
400
500
600
700
100
200
300
400
500
600
700

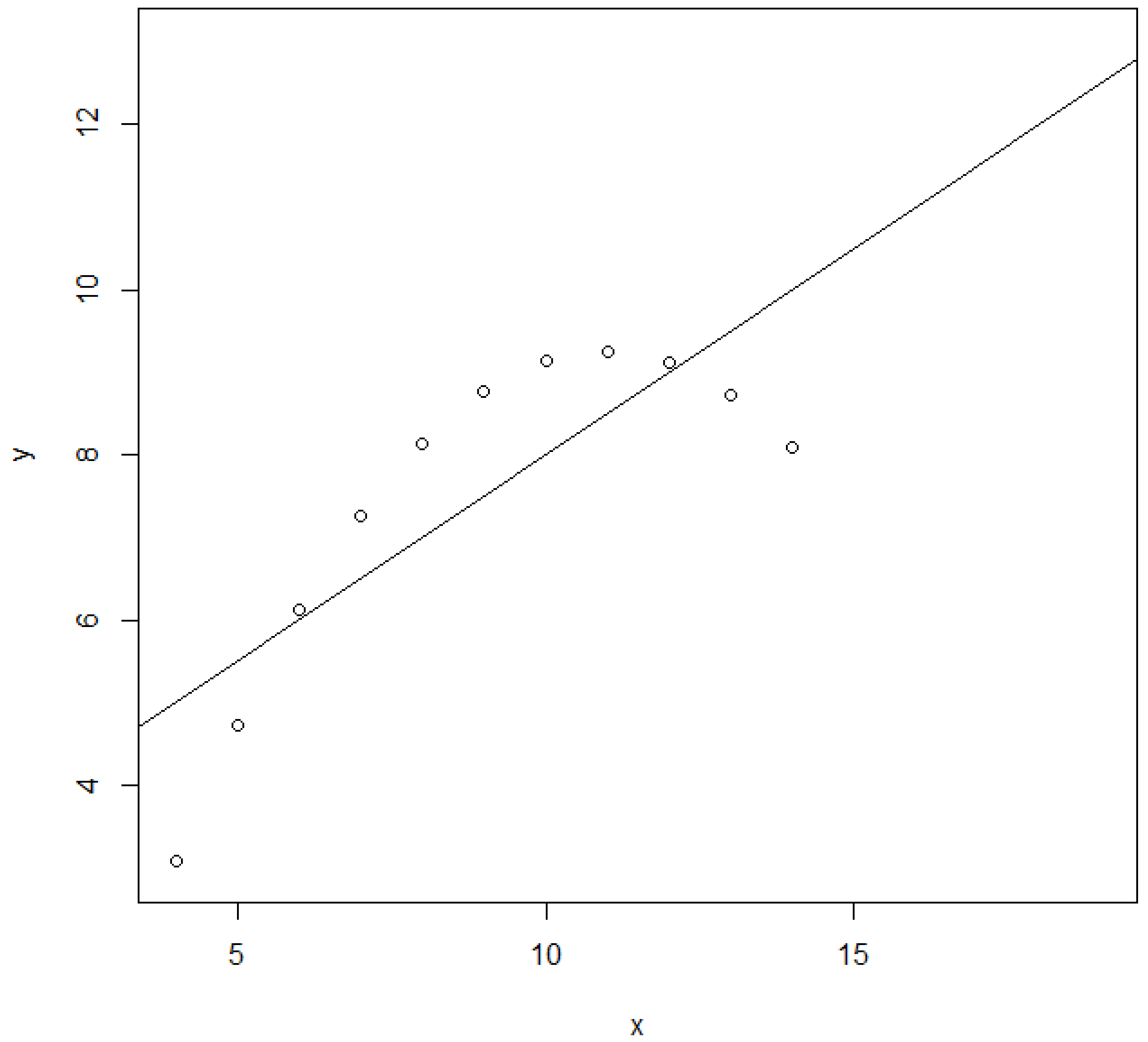

ρ= 0.867
mini.wright
std.wright
100
200
300
400
500
600
700

Summaries vs. Plots

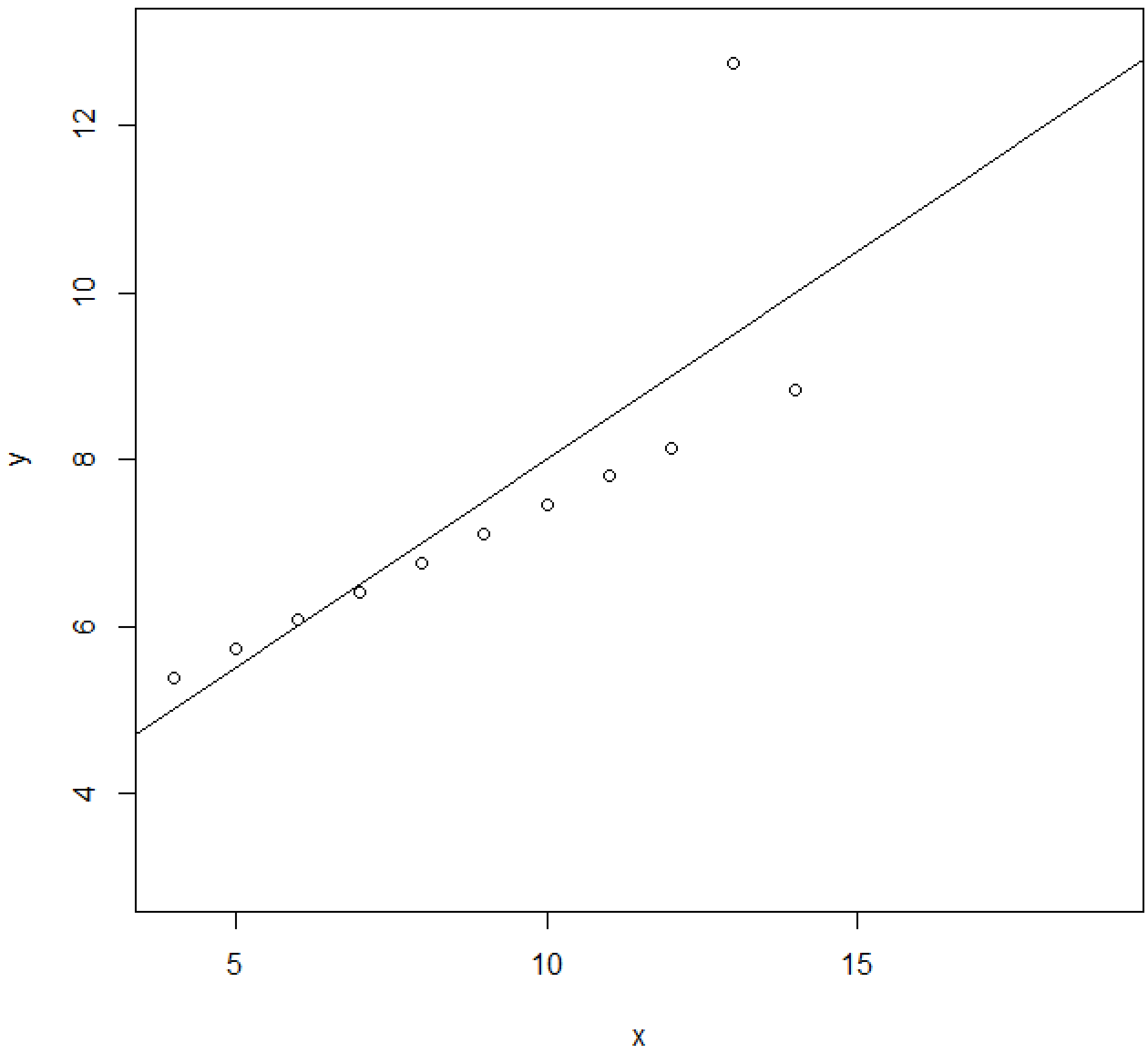
- Four data sets of x/y pairs.
- In each case the mean of x is 9, with variance 11.
- The mean of y is 2.031 with variance 4.13.
- The correlation between x and y is 0.816
- So the summaries are all the same.
- But the appearance and interpretation is very different.
- This example is due to Anscombe.

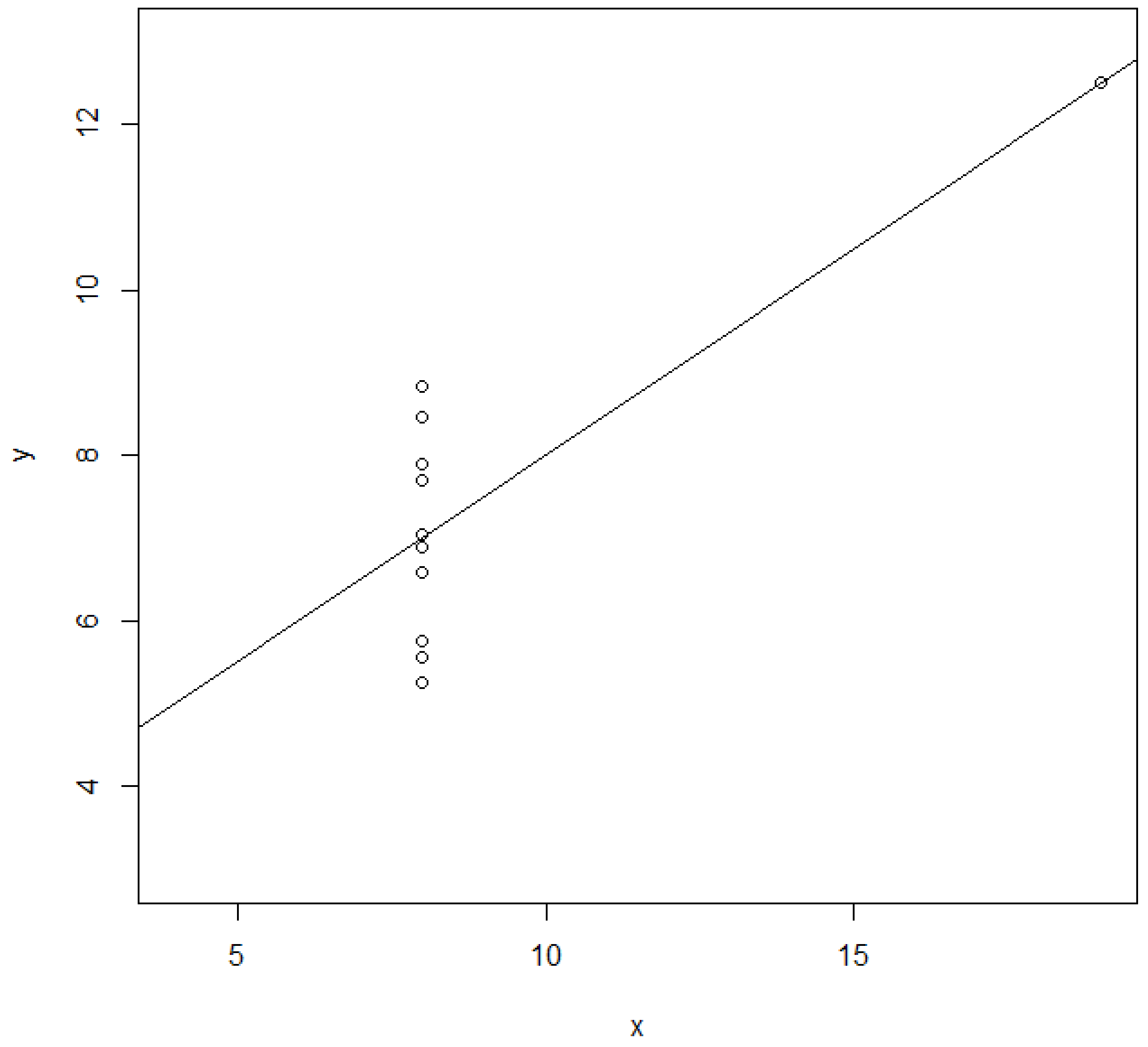


12
10
8
6
4
y
5
10
15
x



12
10
8
6
4
y
5
10
15
x


y
4
6
8
10
12
5
10
15
x



4
6
8
10
12
y
5
10
15
x