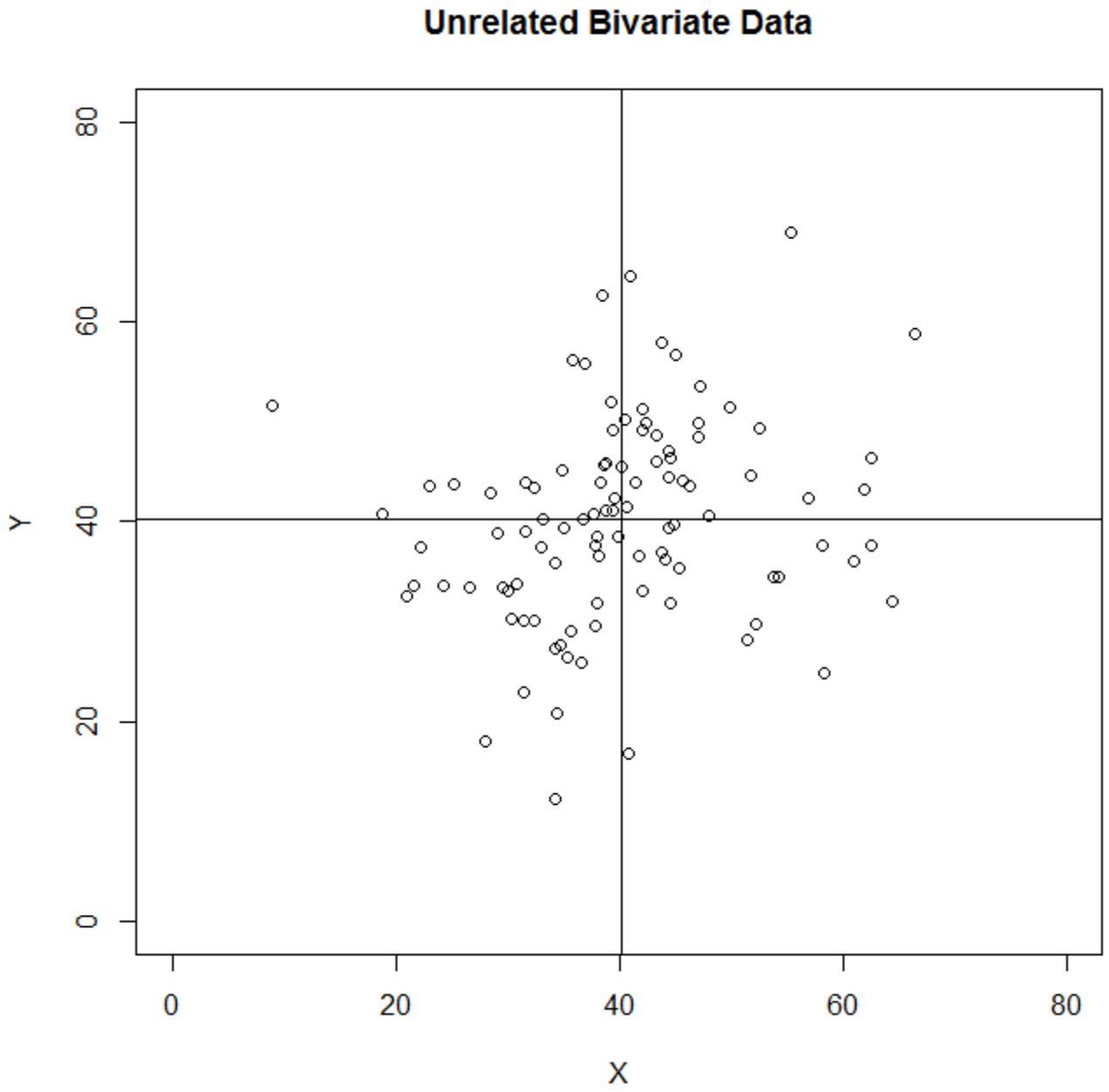
BIM 105
Probability and Statistics for Biomedical Engineers

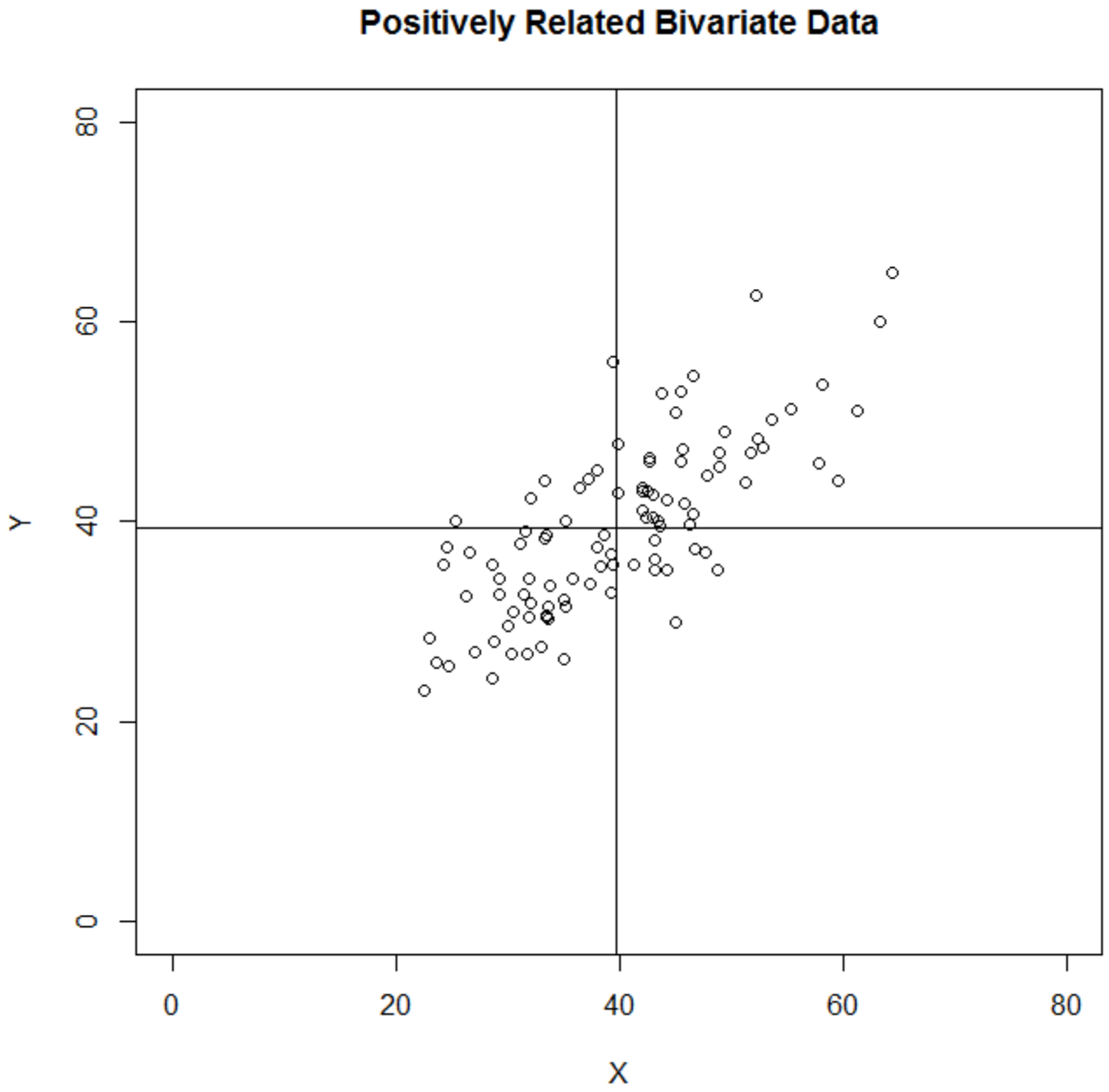
David M. Rocke

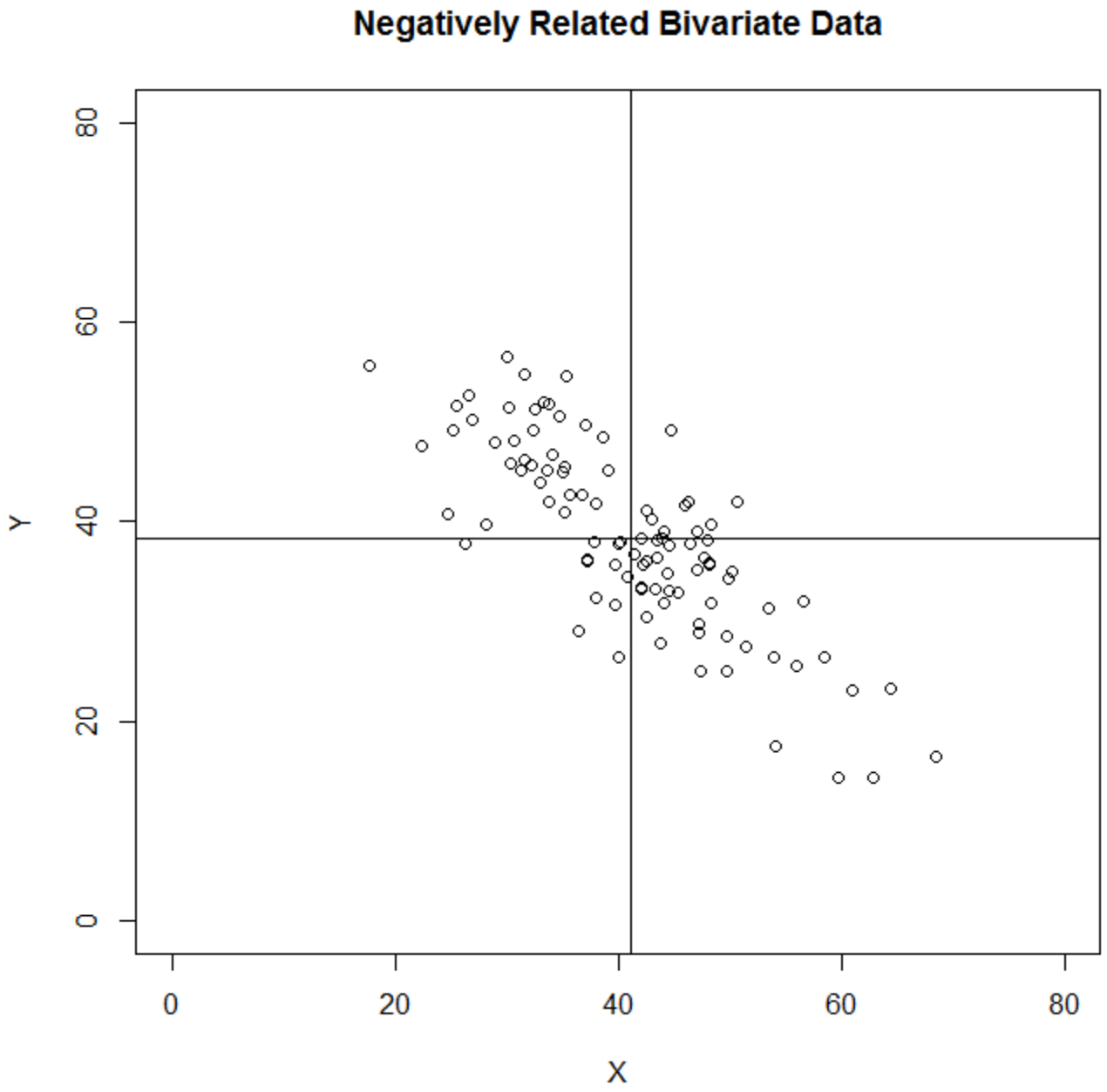
Department of Biomedical Engineering

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
X
0
20
40
60
80


Positively Related Bivariate Data
Y
X
0
20
40
60
80


Negatively Related Bivariate Data
Y
X
0
20
40
60
80

Measuring Relatedness

Variance of X

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

Variance of Y

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

Covariance of X and Y

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

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

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

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

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

Scaling

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

It is often said to be *standardized*.

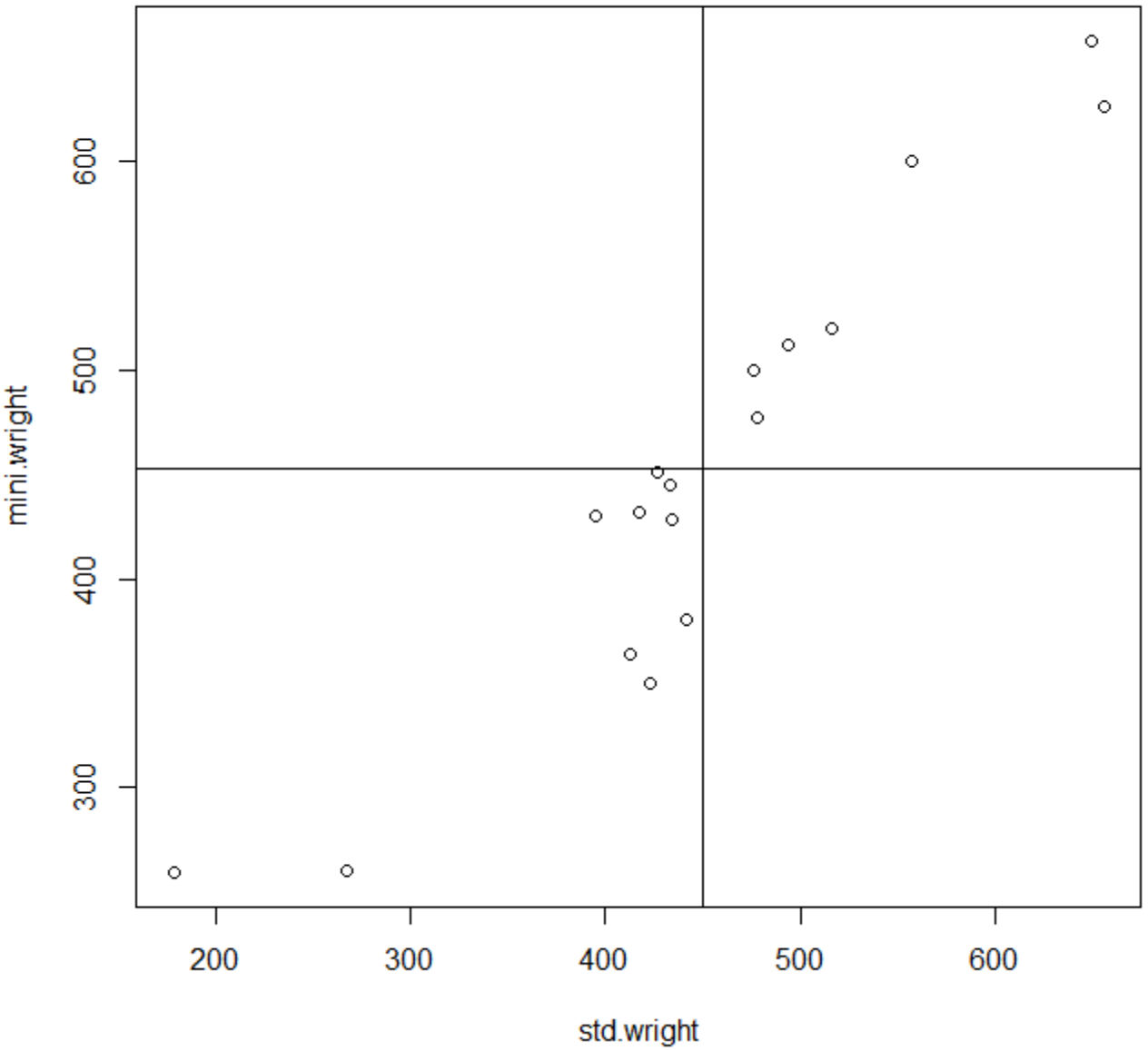
ρ_{XY} is always between -1 and 1

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

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
```

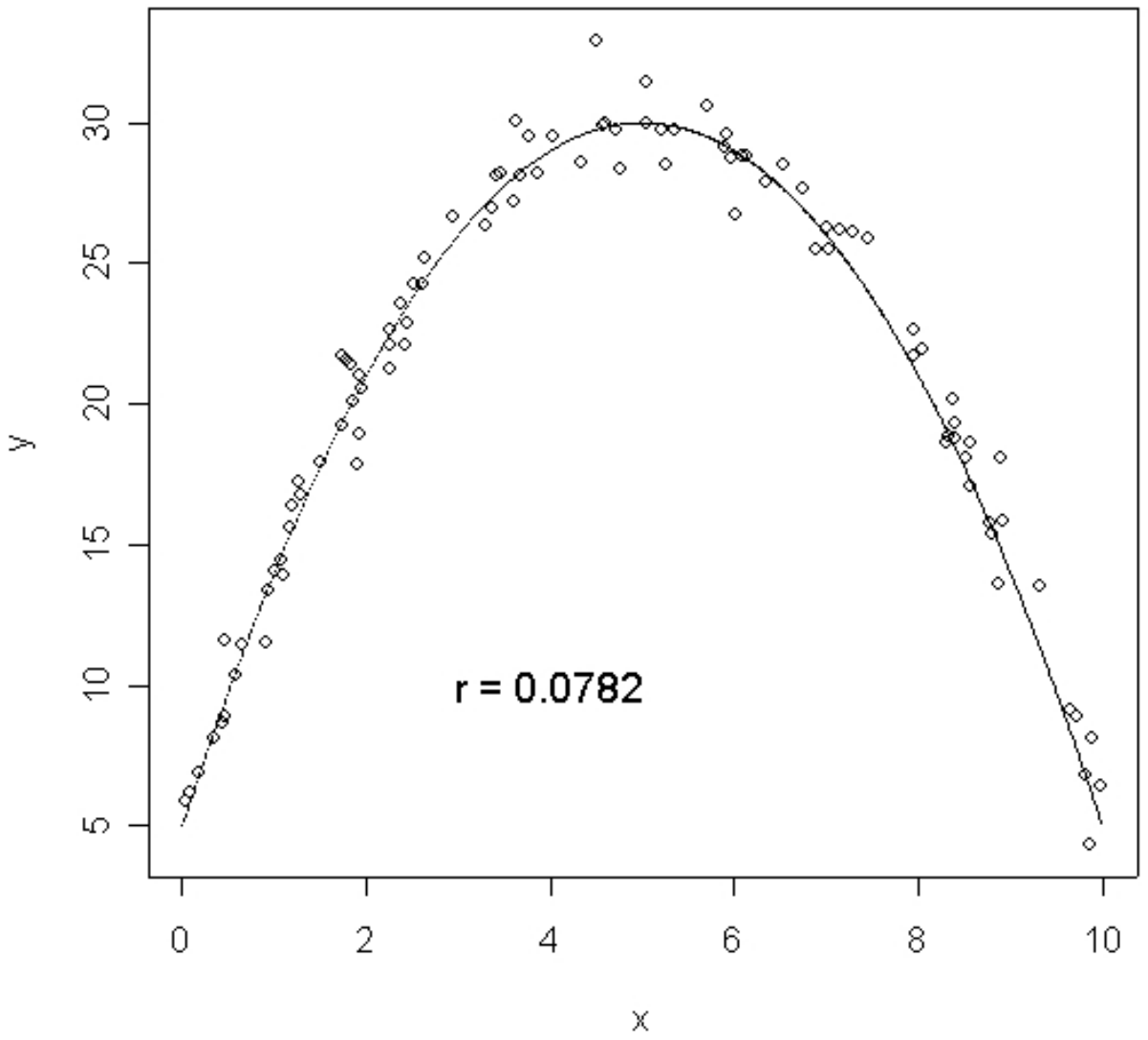
Correlation in MATLAB

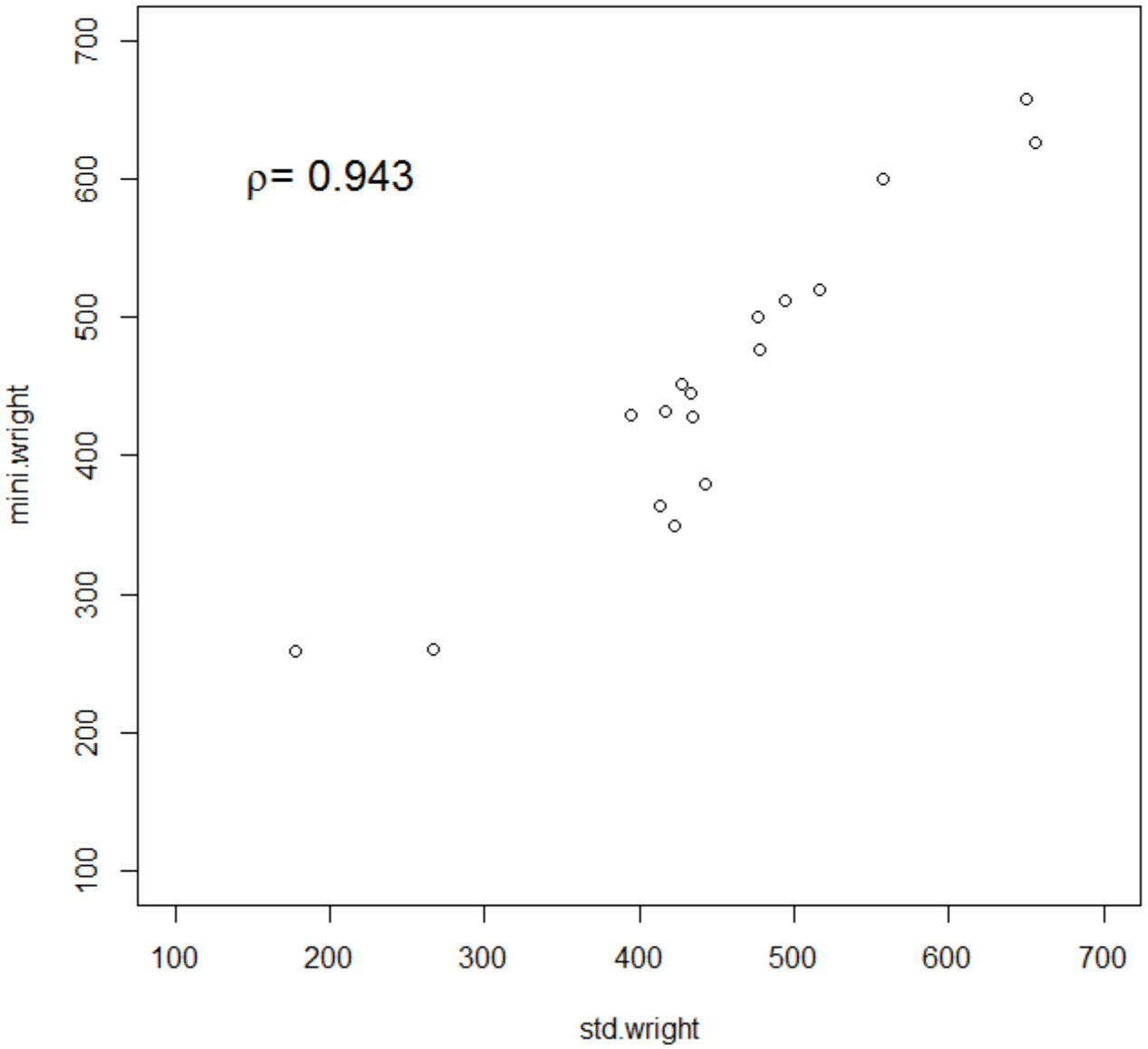
```
>> corrcoef(stdwright,miniwright)

ans =

    1.0000    0.9433
    0.9433    1.0000

>> cov(stdwright,miniwright)

ans =

   1.0e+04 *

    1.3529    1.2410
    1.2410    1.2795
```

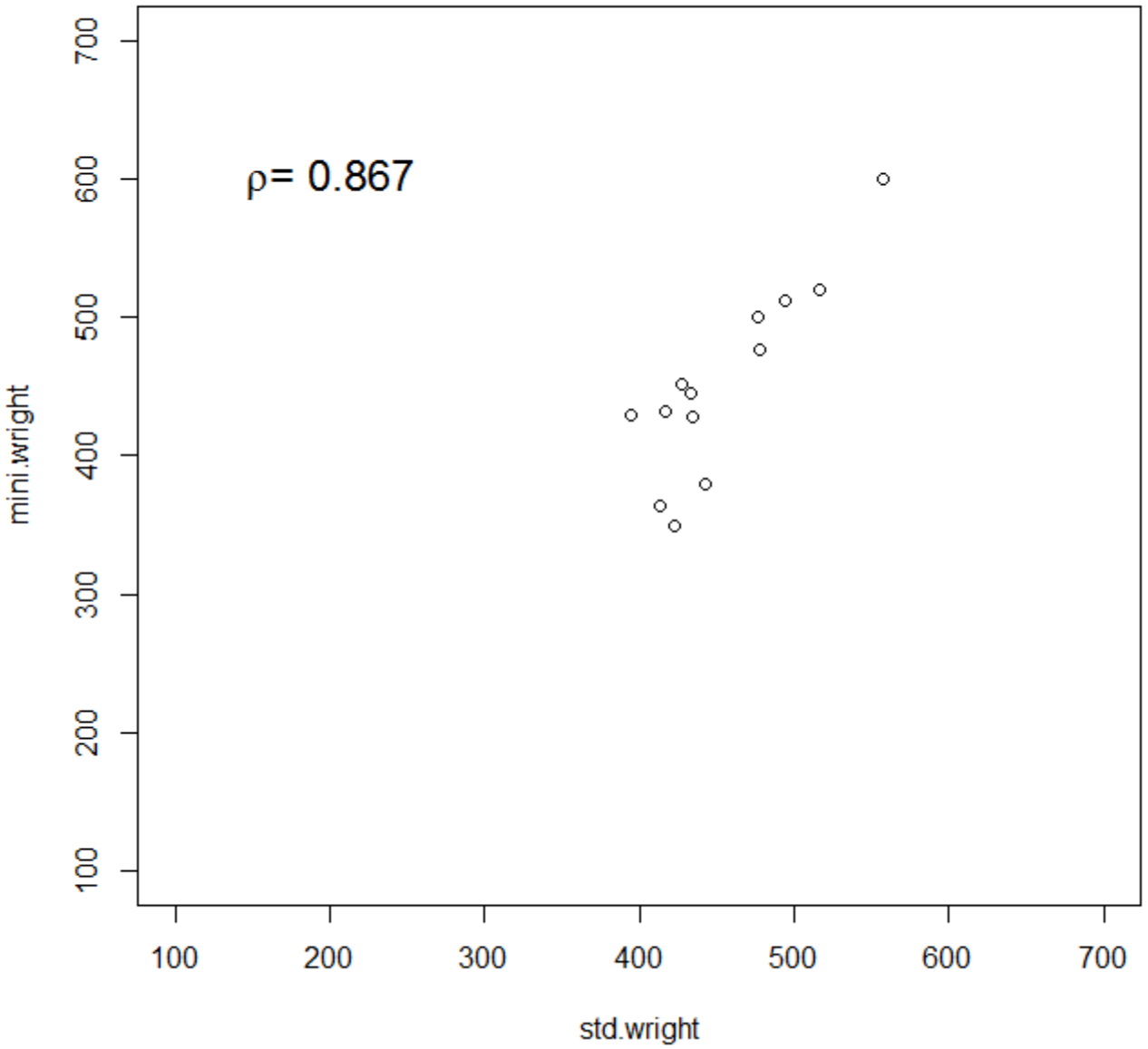
Cautions about Correlation

- 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

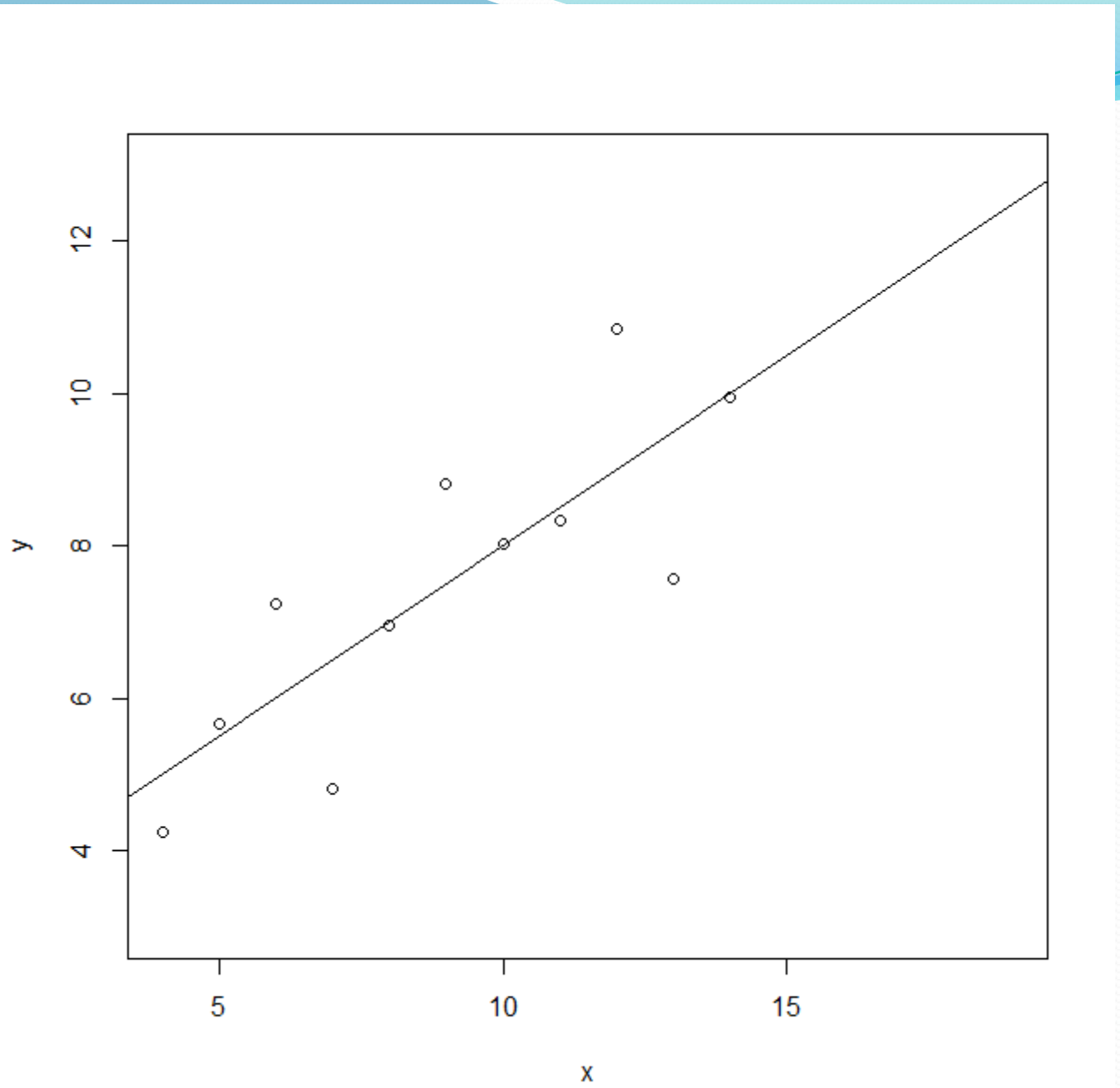


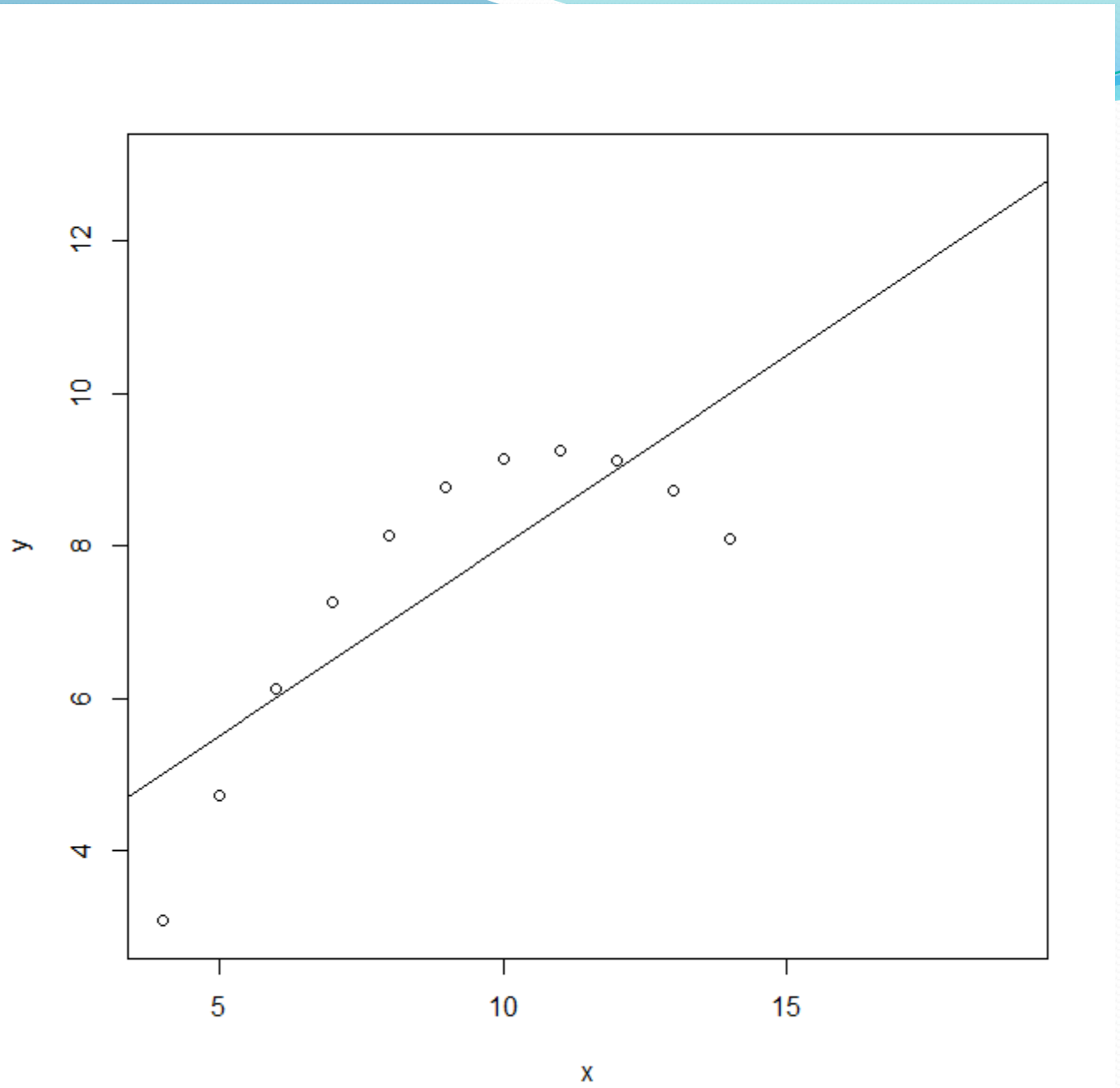

ρ= 0.943
mini.wright
std.wright
100
200
300
400
500
600
700

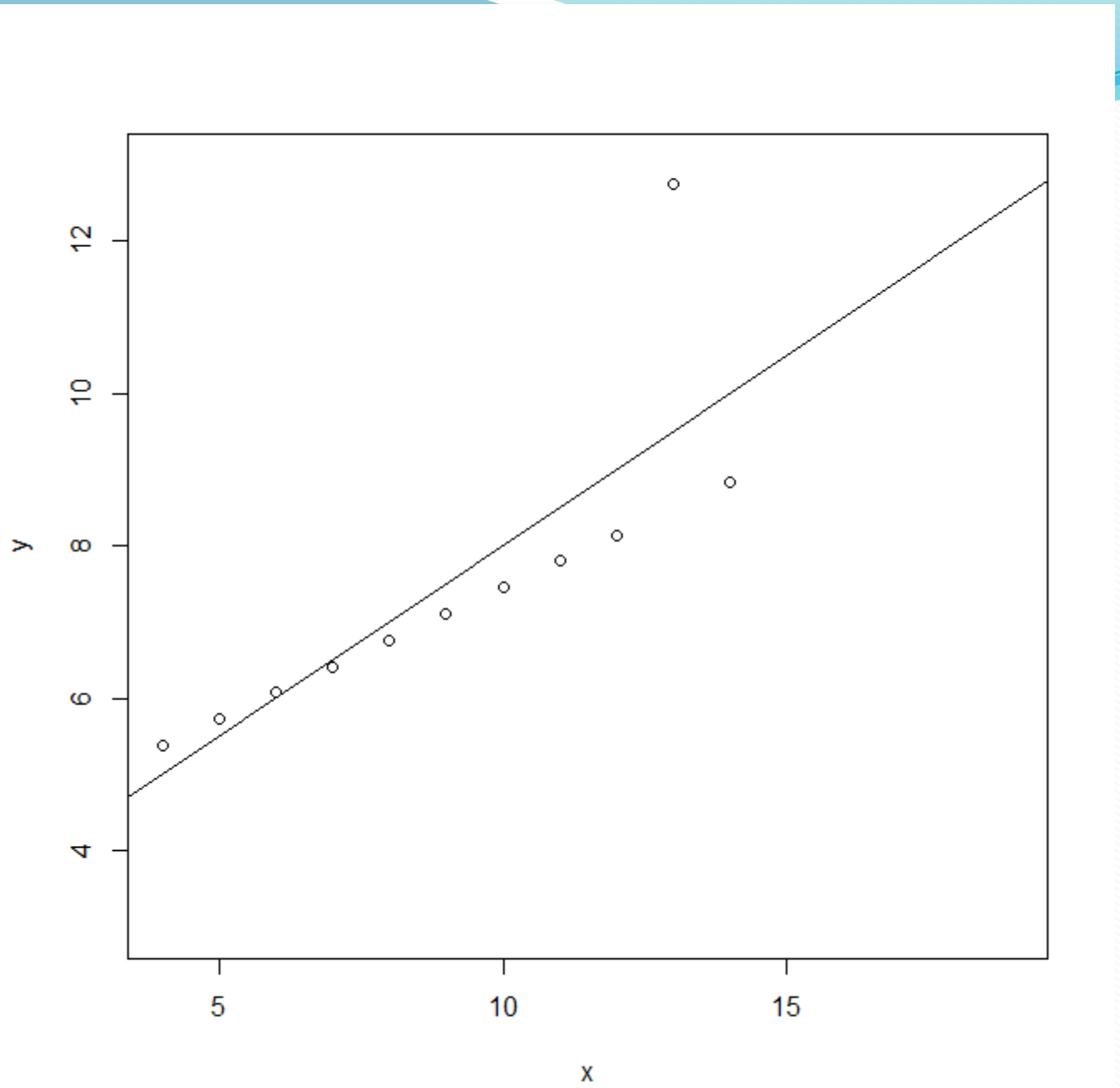

ρ= 0.867
mini.wright
std.wright
100
200
300
400
500
600
700

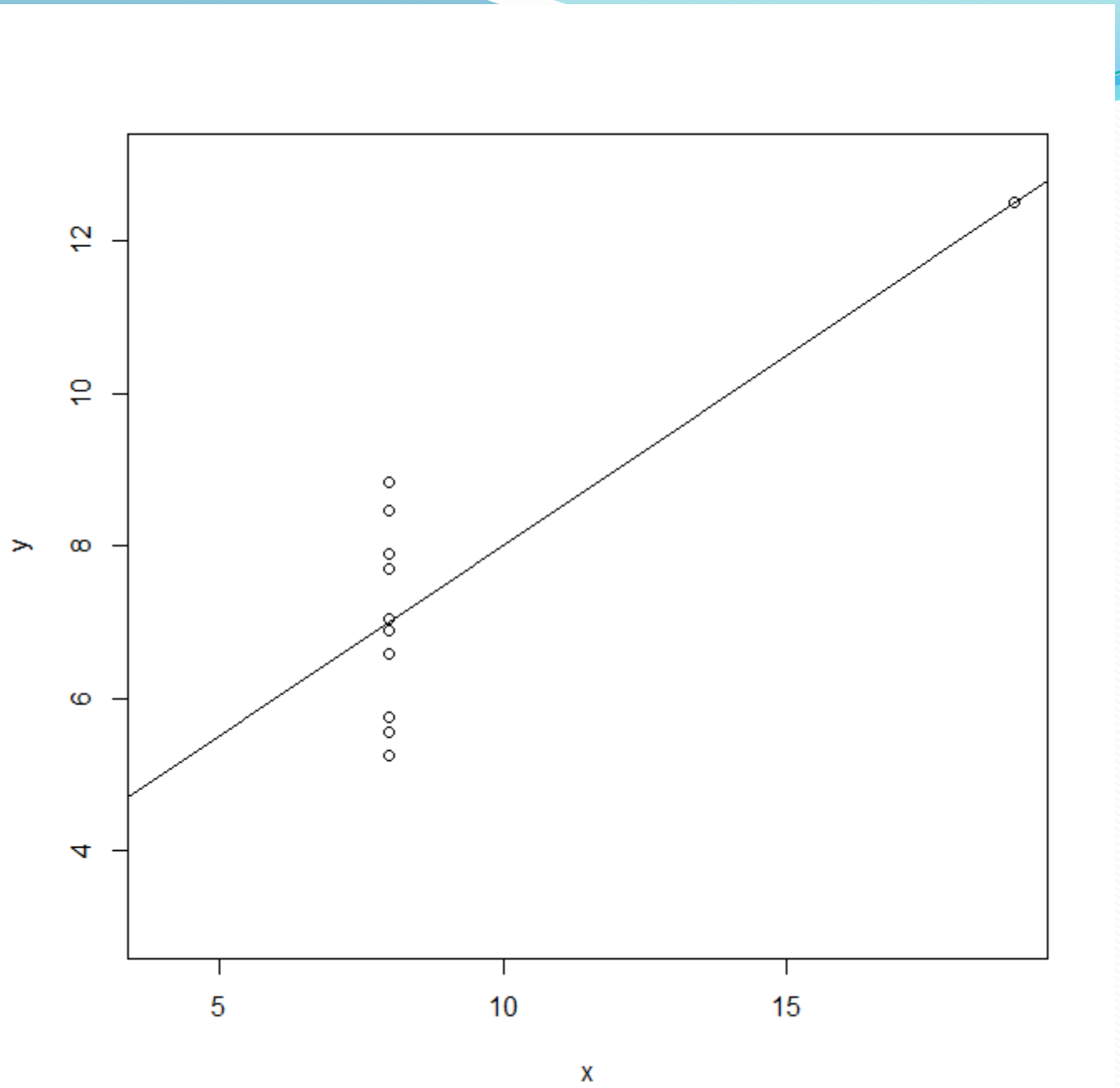
Summaries vs. Plots

- 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.


y
12
10
8
6
4
5
10
15
x

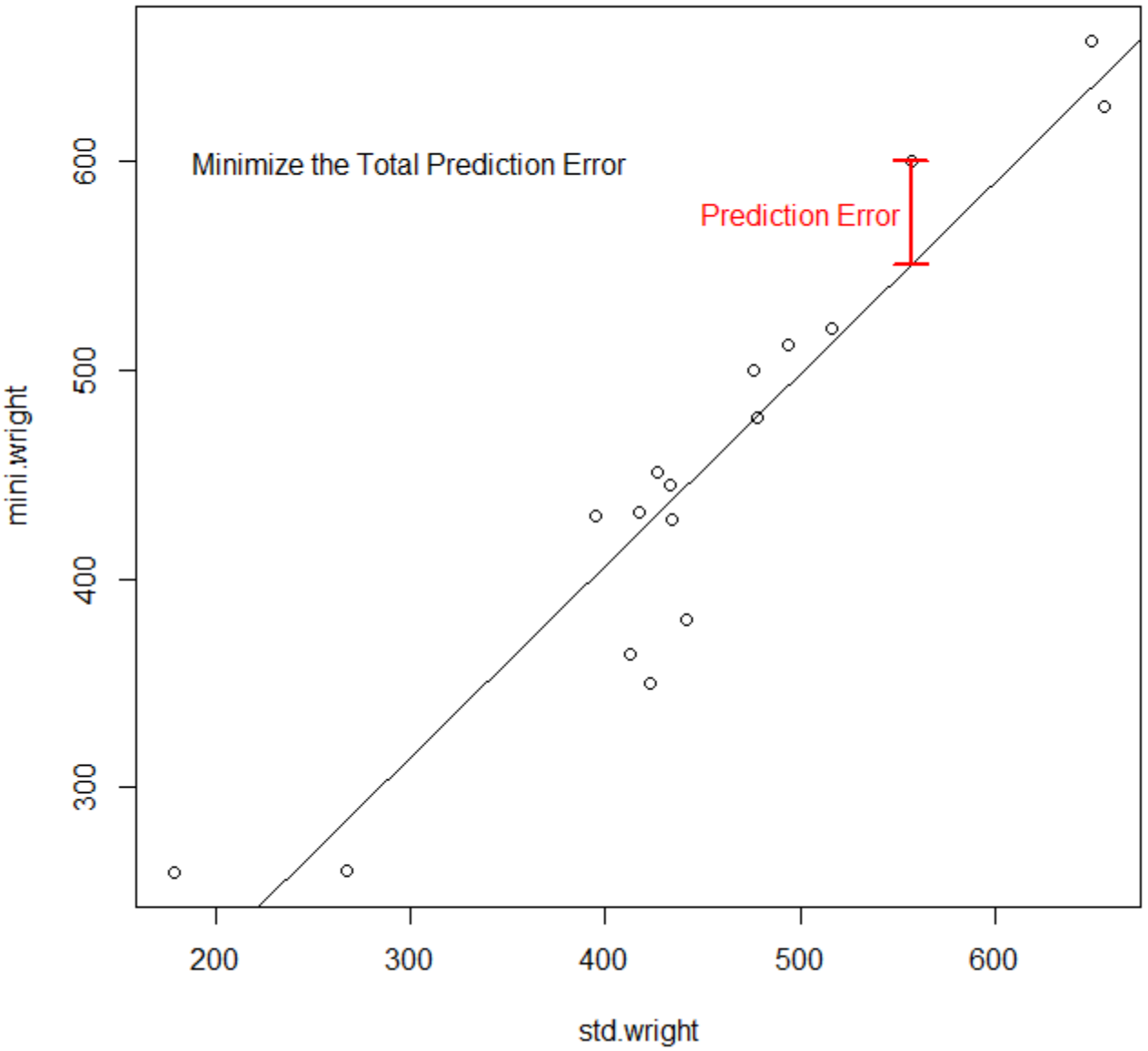

y
12
10
8
6
4
5
10
15
x


y
12
10
8
6
4
5
10
15
x


y
12
10
8
6
4
5
10
15
x

The Least-Squares Line

- Correlation is symmetric in X and Y
- If we want to predict Y from X, we can construct a prediction function $y = f(x)$ which given x will make the best prediction it can about the value of y.
- Often, we use a straight-line prediction function $y = \beta_0 + \beta_1 x$
- We need to find the intercept and slope so that the line fits as well as it can.
- One solution is the least-squares line.


Minimize the Total Prediction Error
Prediction Error
mini.wright
std.wright
300
400
500
600
200
300
400
500
600

Minimize Total Prediction Error

A possible prediction line is

$$y = f(x) = \beta_0 + \beta_1 x$$

For one of the data points (x_i, y_i), the prediction error is

$$y_i - f(x_i) = y_i - \beta_0 - \beta_1 x_i$$

We want to minimize the total magnitude of the errors,

so we don't want positive and negative errors to cancel.

We need to use something like the absolute value or the square of the error.

The total square error is

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - f(x_i) \right)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

The least-squares line is the choice of intercept and slope that

minimizes the sum of the squares of the errors

Finding the Least-Squares Line

Once we have the data points, the SSE depends only on

The choice of intercept and slope

$$\mathrm{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

The minimum of the SSE is at the point where the two partial derivatives are 0

$$\frac{1}{2}\frac{\partial}{\partial \beta_0}\mathrm{SSE}(\beta_0, \beta_1) = 0$$

$$\frac{1}{2}\frac{\partial}{\partial \beta_1}\mathrm{SSE}(\beta_0, \beta_1) = 0$$

We multiply by 1/2 to eliminate the factor of 2 that we will get from differentiation

Finding the Least-Squares Line

$$\frac{1}{2}\frac{\partial}{\partial\beta_0}\mathrm{SSE}(\beta_0,\beta_1)=0=\frac{1}{2}\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2=\frac{1}{2}\sum_{i=1}^{n}2\left(y_i-\beta_0-\beta_1x_i\right)(-1)$$

$$=\sum_{i=1}^{n}\left(-y_i+\beta_0+\beta_1x_i\right)$$

$$=-\sum_{i=1}^{n}y_i+\sum_{i=1}^{n}\beta_0+\beta_1\sum_{i=1}^{n}x_i$$

$$0=-n\bar{y}+n\beta_0+n\beta_1\bar{x}$$

$$\bar{y}=\beta_0+\beta_1\bar{x}$$

The point $(\bar{x},\bar{y})$ is on the least squares line

Finding the Least-Squares Line

$$\frac{1}{2}\frac{\partial}{\partial \beta_1}\mathrm{SSE}(\beta_0,\beta_1)=\frac{1}{2}\frac{\partial}{\partial \beta_1}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2=\frac{1}{2}\sum_{i=1}^{n}2\left(y_i-\beta_0-\beta_1 x_i\right)(-x_i)$$

$$=\sum_{i=1}^{n}\left(-x_i y_i+\beta_0 x_i+\beta_1 x_i^2\right)$$

$$=-\sum_{i=1}^{n}x_i y_i+\sum_{i=1}^{n}\beta_0 x_i+\beta_1\sum_{i=1}^{n}x_i^{\,2}$$

$$=-SS_{XY}+n\beta_0\bar{x}+\beta_1 SS_{XX}$$

$$0 = \frac{1}{2}\frac{\partial}{\partial \beta_0}\text{SSE}(\beta_0, \beta_1) = -n\bar{y} + n\beta_0 + n\beta_1\bar{x}$$

$$0 = \frac{1}{2}\frac{\partial}{\partial \beta_1}\text{SSE}(\beta_0, \beta_1) = -SS_{XY} + n\beta_0\bar{x} + \beta_1 SS_{XX}$$

$$\beta_0 + \beta_1\bar{x} = \bar{y}$$

$$n\beta_0\bar{x} + \beta_1 SS_{XX} = SS_{XY}$$

Two equations in two unknowns, β_0 and β_1

The *normal equations*

$$\beta_0 + \beta_1 \bar{x} = \bar{y}$$
$$n\beta_0 \bar{x} + \beta_1 SS_{XX} = SS_{XY}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$
$$n(\bar{y} - \beta_1 \bar{x})\bar{x} + \beta_1 SS_{XX} = SS_{XY}$$
$$n\overline{xy} - n\beta_1 \bar{x}^2 + \beta_1 SS_{XX} = SS_{XY}$$
$$\beta_1 = \frac{SS_{XY} - n\overline{xy}}{SS_{XX} - n\bar{x}^2}$$

Equivalently

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_X^2}$$

The slope is the ratio of the covariance to the variance of X

Finding the Least Squares Line

- The computational formulas work for relatively small samples.
- Most of the time, the least squares line is computed in a computer program (MATLAB, Excel, R, …)
- In MATLAB there are several commands that can compute the least squares line.

```
>> fitlm(stdwright,miniwright)
```

Fits the least squares line

```
>> lsline
```

Adds the least squares line to a plot (the last one plotted)

Finding the Least Squares Line

```
mdl = fitlm(tbl)
mdl = fitlm(tbl,modelspec)
mdl = fitlm(X,y)
mdl = fitlm(X,y,modelspec)
```

`tbl` is a table object, using the last column as `y`

`X` is a matrix of predictors

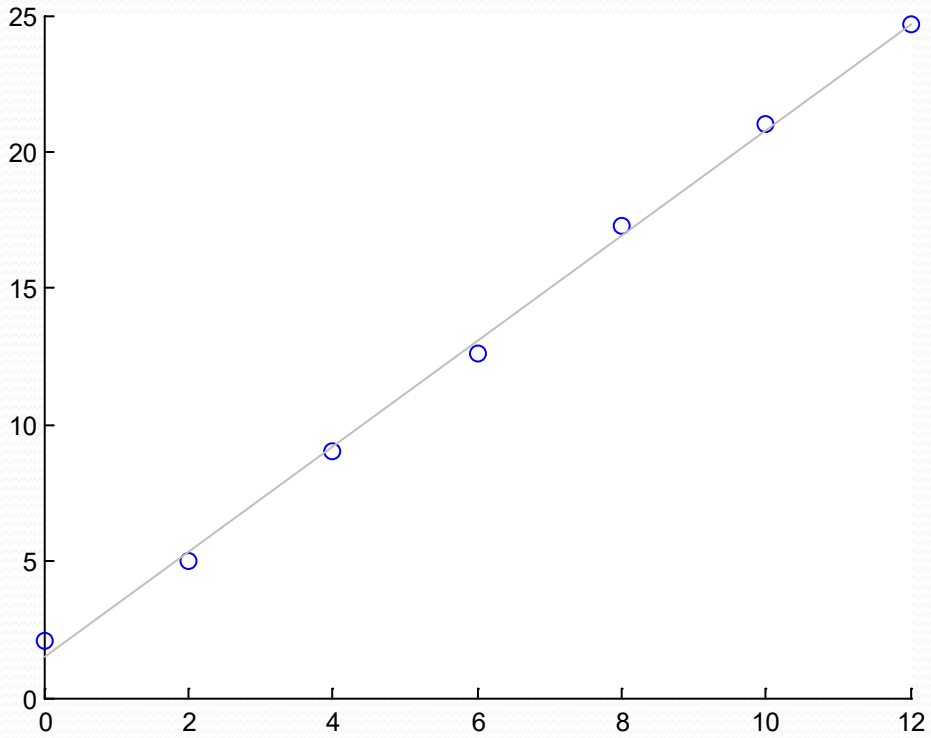
`modelspec` says which variables to use and how

Calibration

- Standard aqueous solutions of fluorescein (in pg/ml) are examined in a fluorescence spectrometer and the intensity (arbitrary units) is recorded
- What is the relationship of intensity to concentration
- Use later to infer concentration of labeled analyte

Concentration (pg/ml)	0	2	4	6	8	10	12
Intensity	2.1	5.0	9.0	12.6	17.3	21.0	24.7

```
>> concentration = [0 2 4 6 8 10 12]
>> intensity = [2.1 5.0 9.0 12.6 17.3 21.0 24.7]
>> scatter(concentration,intensity)
>> lsline
>> fitlm(concentration,intensity)

Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    ________    ______    __________

    (Intercept)    1.5179      0.29494    5.1464    0.0036258
    x1             1.9304       0.0409    47.197    8.066e-08

Number of observations: 7, Error degrees of freedom: 5
Root Mean Squared Error: 0.433
R-squared: 0.998,  Adjusted R-Squared 0.997
F-statistic vs. constant model: 2.23e+03, p-value = 8.07e-08
```


25
20
15
10
5
0
0
2
4
6
8
10
12

Use of the calibration curve

$\hat{y} = 1.52 + 1.93x$

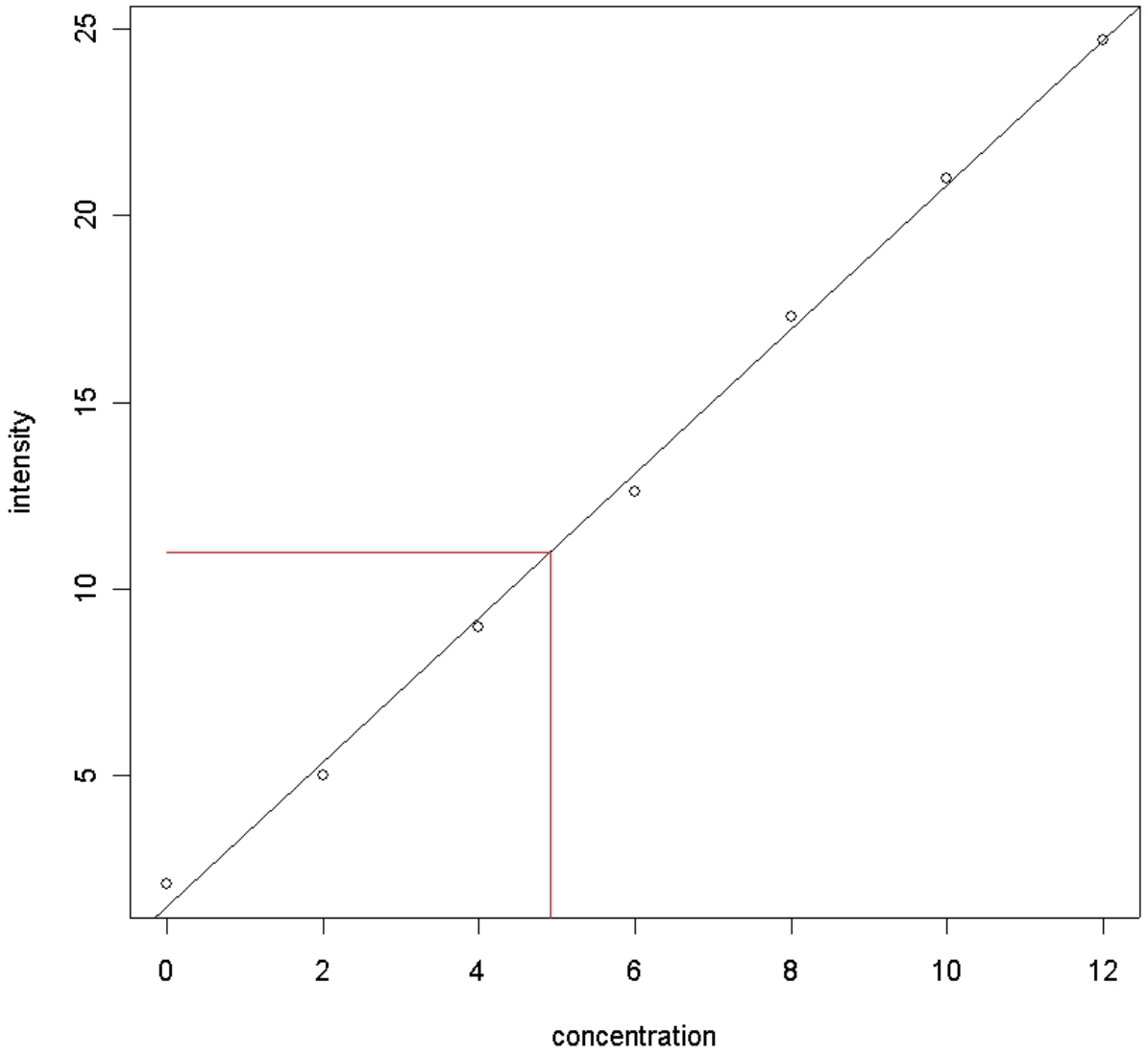
$\hat{y}$ is the predicted average intensity

x is the true concentration

$\hat{x} = \dfrac{y - 1.52}{1.93}$

y is the observed intensity

$\hat{x}$ is the estimated concentration


intensity
25
20
15
10
5
0
2
4
6
8
10
12
concentration

Measurement and Calibration

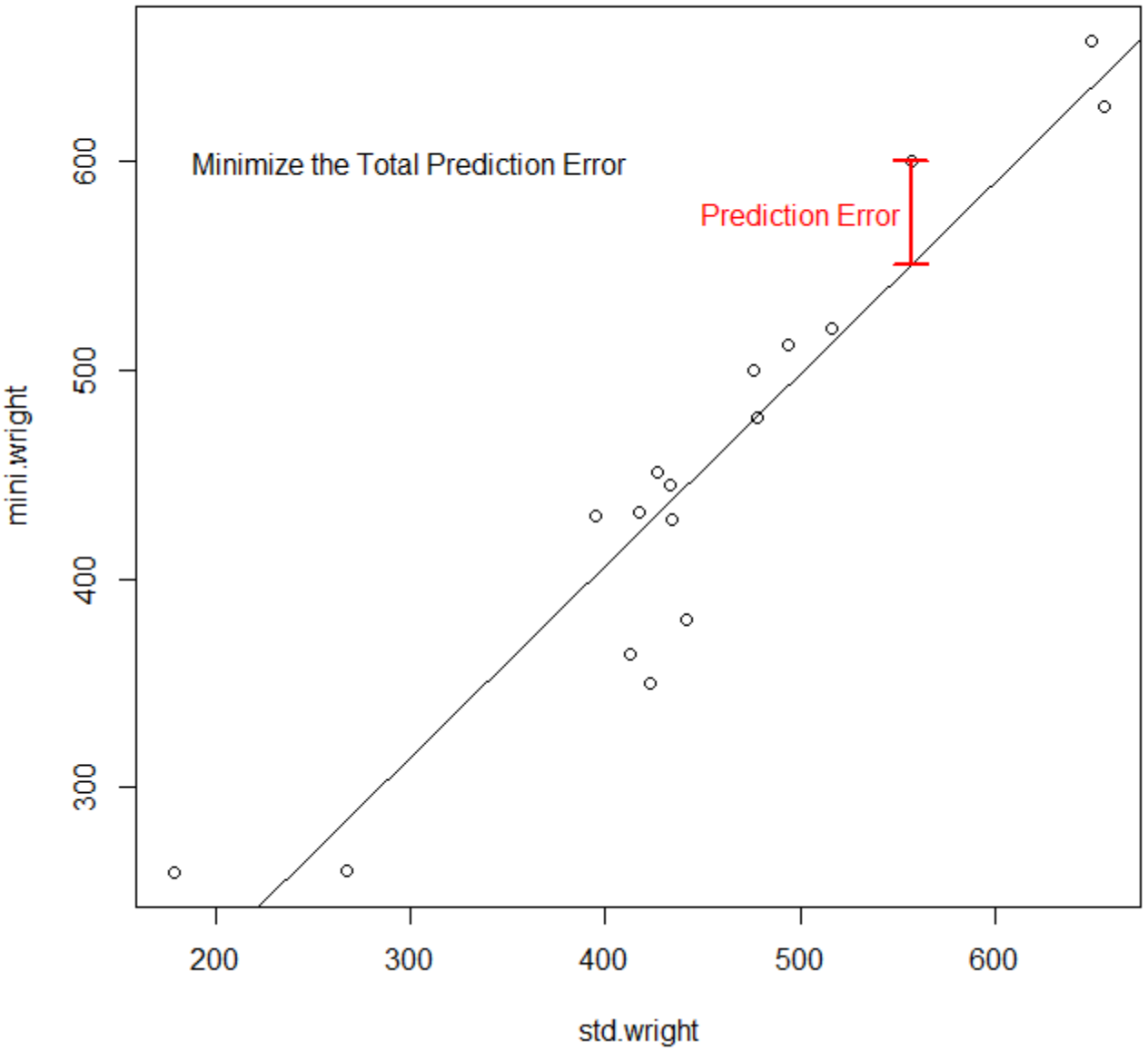
- Essentially all things we measure are indirect
- The thing we wish to measure produces an observed transduced value that is related to the quantity of interest but is not itself directly the quantity of interest
- Calibration takes known quantities, observes the transduced values, and uses the inferred relationship to quantitate unknowns

Measurement Examples

- Weight is observed via deflection of a spring (calibrated)
- Concentration of an analyte in mass spec is observed through the electrical current integrated over a peak (possibly calibrated)
- Gene expression can observed via fluorescence of a spot to which the analyte has bound (usually not calibrated)
- Or via a relative count of DNA fragments that map to a known gene sequence

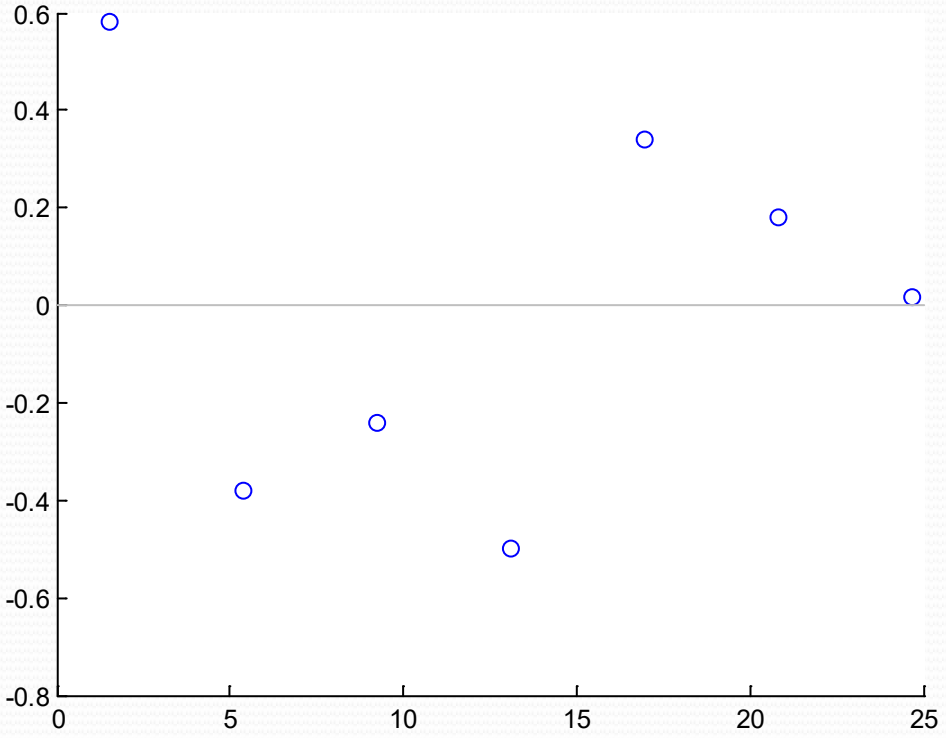
Fitted Values and Residuals

- We have a set of data (x_i, y_i) and a least-squares line $y = a + bx$
- Each value x_i has a fitted value which is $a + bx_i$. All the fitted values lie on the line
- The residual is the difference between the fitted value and the actual value of y.

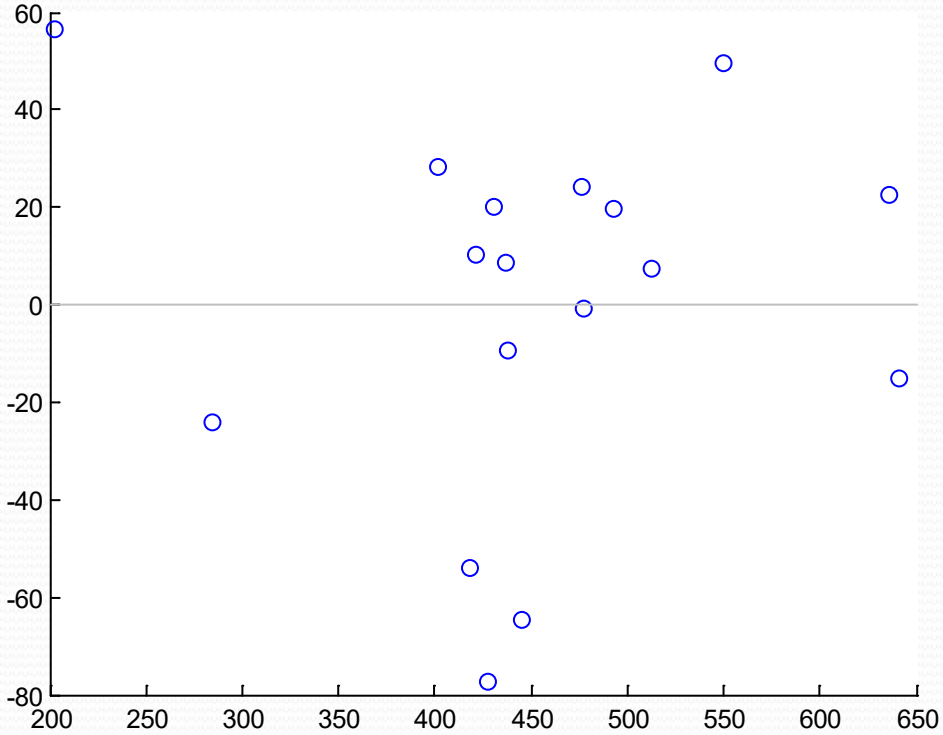

Minimize the Total Prediction Error
Prediction Error
mini.wright
std.wright
300
400
500
600
200
300
400
500
600

```
>> model1 = fitlm(concentration,intensity)
>> fits = model1.Fitted
fits =

    1.5179
    5.3786
    9.2393
   13.1000
   16.9607
   20.8214
   24.6821

>> resids = model1.Residuals.Raw
resids =

    0.5821
   -0.3786
   -0.2393
   -0.5000
    0.3393
    0.1786
    0.0179
```

```
>> scatter(fits,resids)
>> lsline
```



```
>> model2 = fitlm(stdwright,miniwright)

Estimated Coefficients:
                    Estimate       SE       tStat       pValue
                    ________    ________    _______    __________

    (Intercept)       39.34       38.704    1.0164        0.32554
    x1              0.91735     0.083365    11.004    1.3995e-08
>> fits2 = model2.Fitted
>> resids2 = model2.Residuals.Raw
>> scatter(fits2,resids2)
>> lsline
```


60
40
20
0
-20
-40
-60
-80
200
250
300
350
400
450
500
550
600
650

Matlab Objects

fitlm() creates a linear model object.

Parts of the object can be addressed using suffixes such as

```
>> model1 = fitlm(concentration,intensity)
>> fits = model1.Fitted
>> resids = model1.Residuals.Raw
```

For a complete list of components see

www.mathworks.com/help/stats/linearmodel-class.html

Other Issues

- A least squares (straight) line is only useful if the relationship between x and y is roughly linear. Check with a scatter plot and plot of residuals vs. fitted values.
- Outlying values can badly distort the computed line. Check with a scatter plot and plot of residuals vs. fitted values.