Correlation and linear regression

Petr Ocelík

ESS401 Social Science Methodology

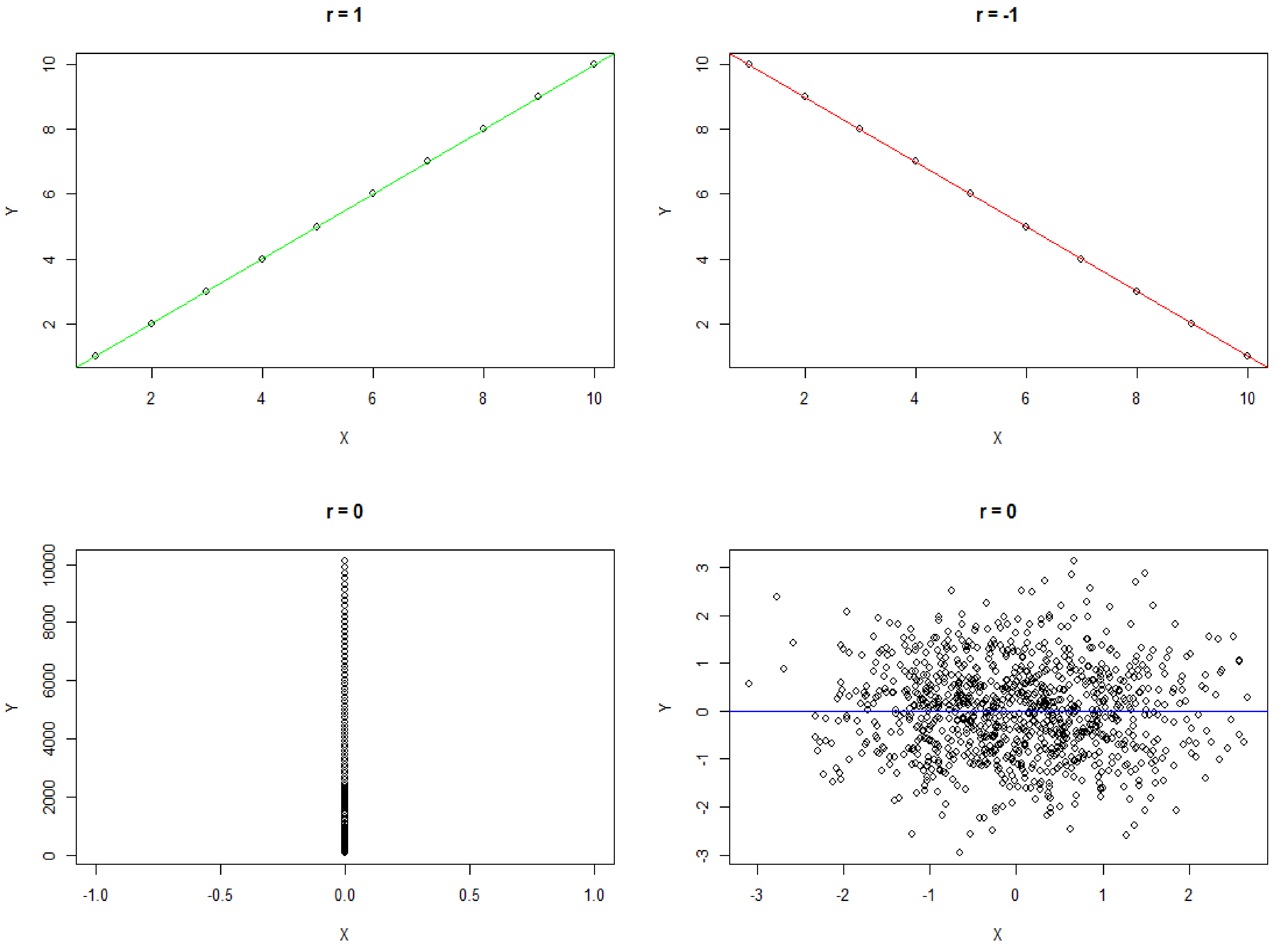
20th October 2015

Outline

- Correlation
- Simple linear regression
- Correlation and linear regression in R

Correlation

- Pearson's product-moment correlation coefficient (**r**).
- Correlation measures the **strength of the relationship between two variables**.
- Ranges between -1 (perfect negative corr) and 1 (perfect positive corr).
- 0 indicates no systematic **linear** relationship between variables.
- Value does not depend on variables' units.
- It is a **sample statistic.**

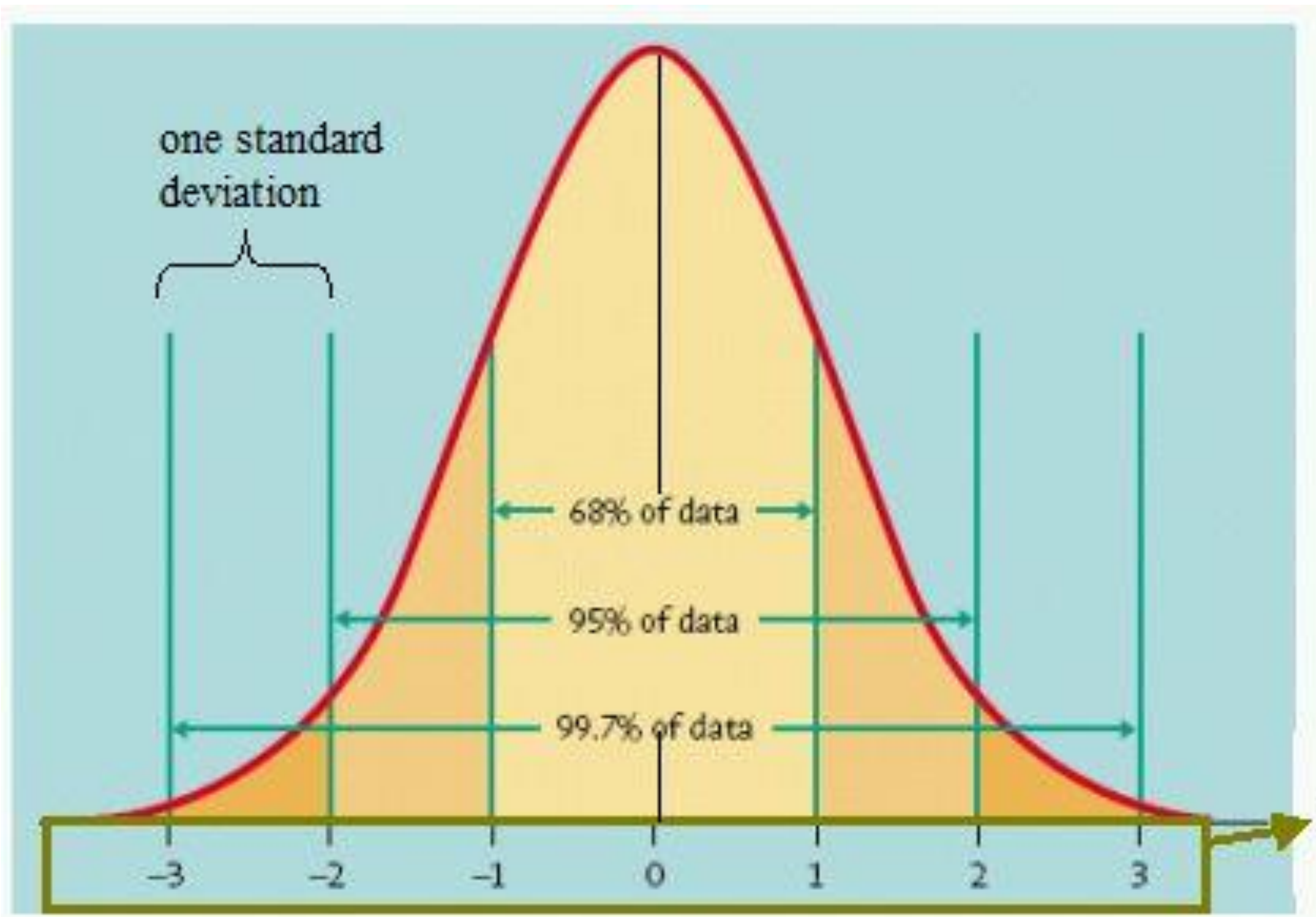


r = 1
r = -1
r = 0
r = 0
X
Y

Correlation

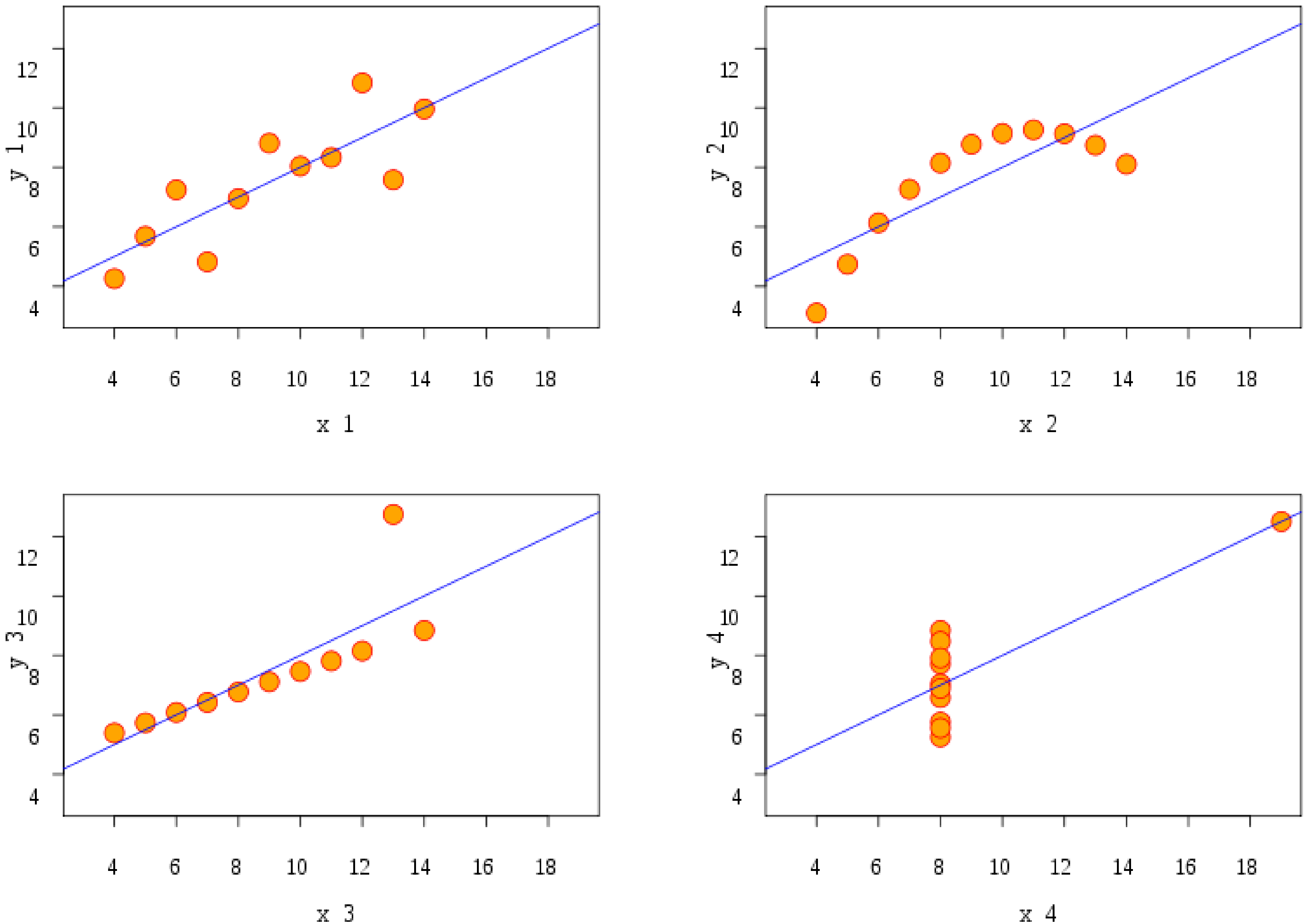
- Assumptions and limitations:
  - Normal distribution of X and Y
  - Linear relationship between X and Y
  - Homoscedasticity
  - Sensitive to outliers

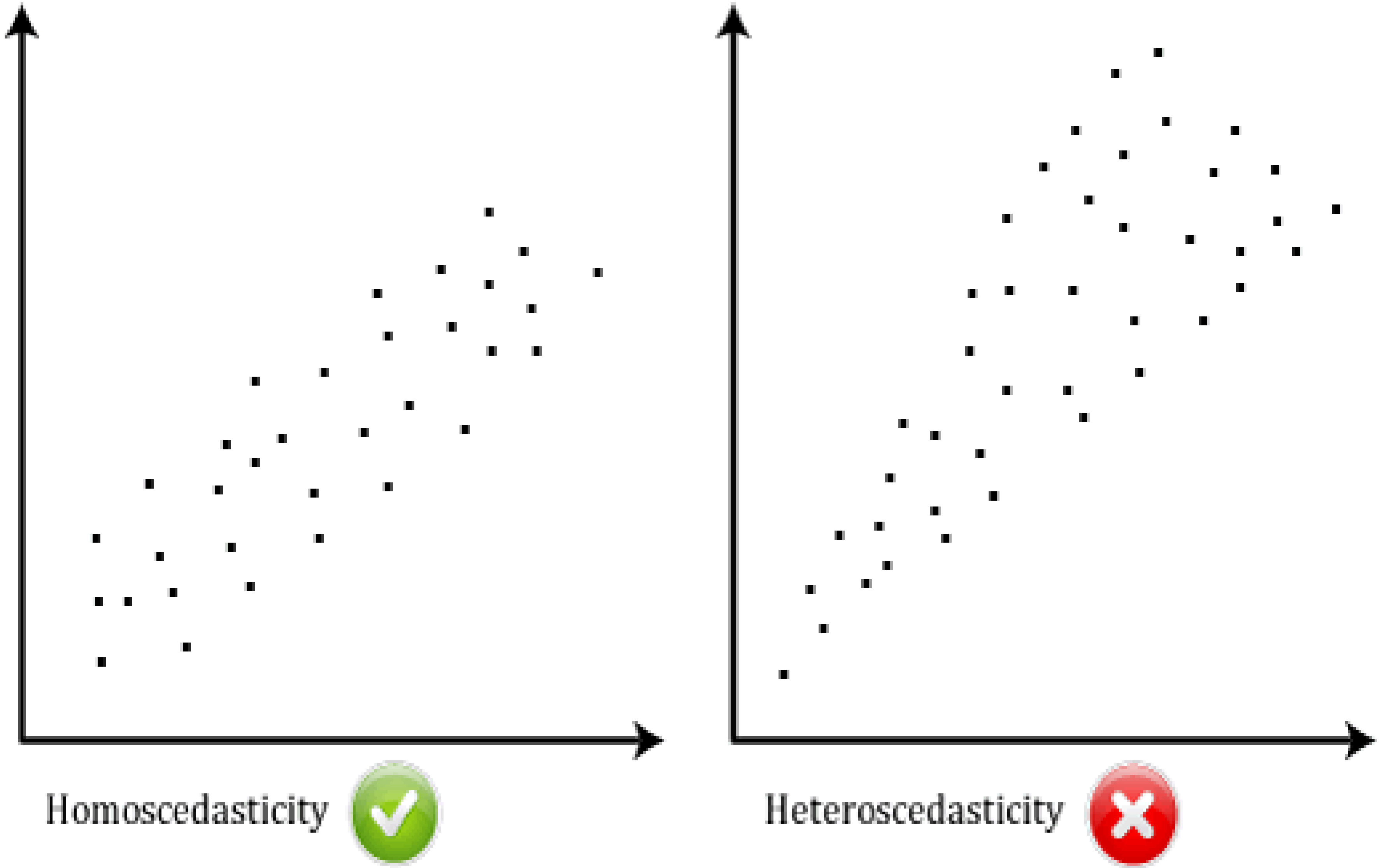
The standard normal distribution

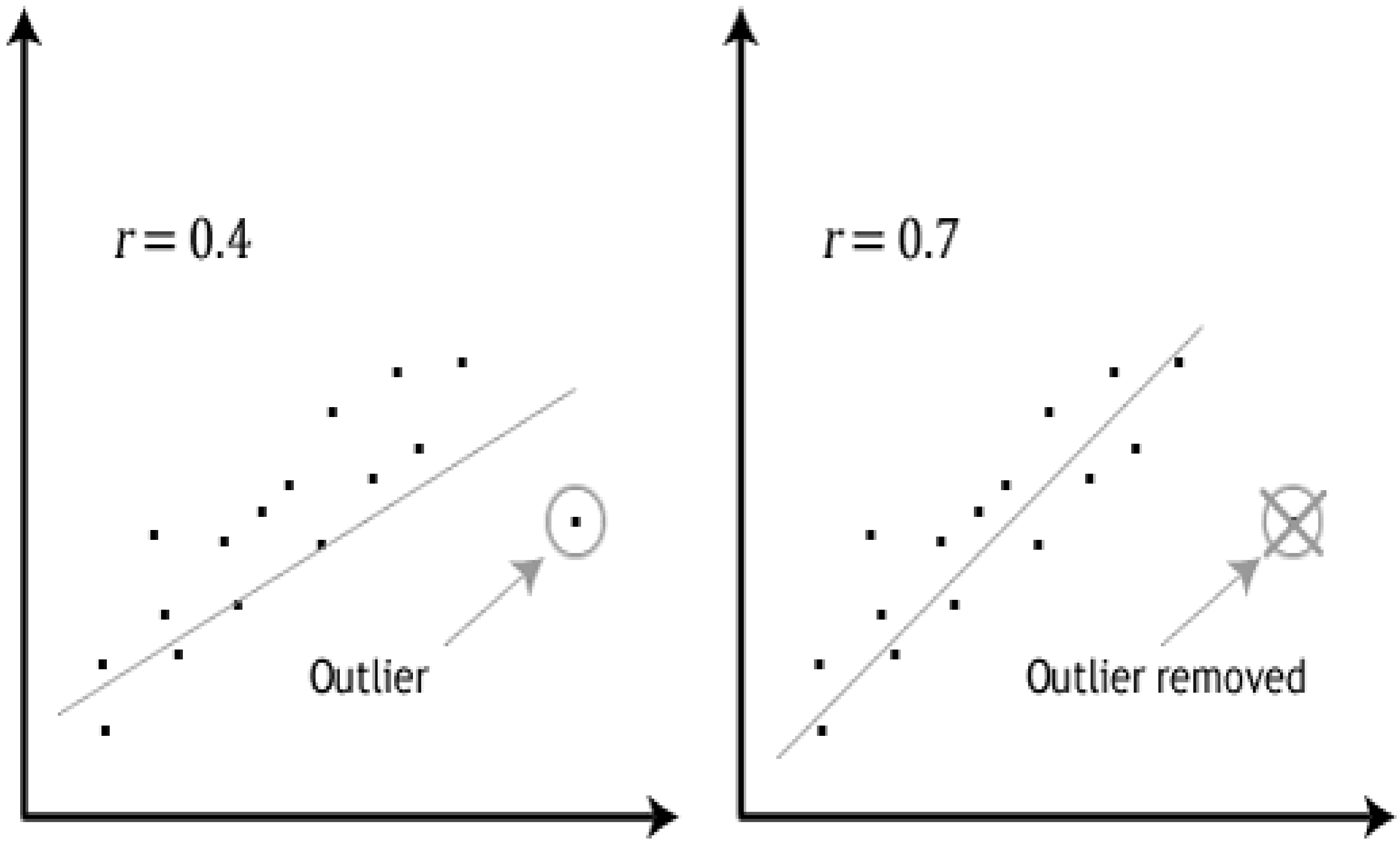


$$z = \frac{x - \mu}{\sigma}$$

Anscombe's quartet




Homoscedasticity
Heteroscedasticity



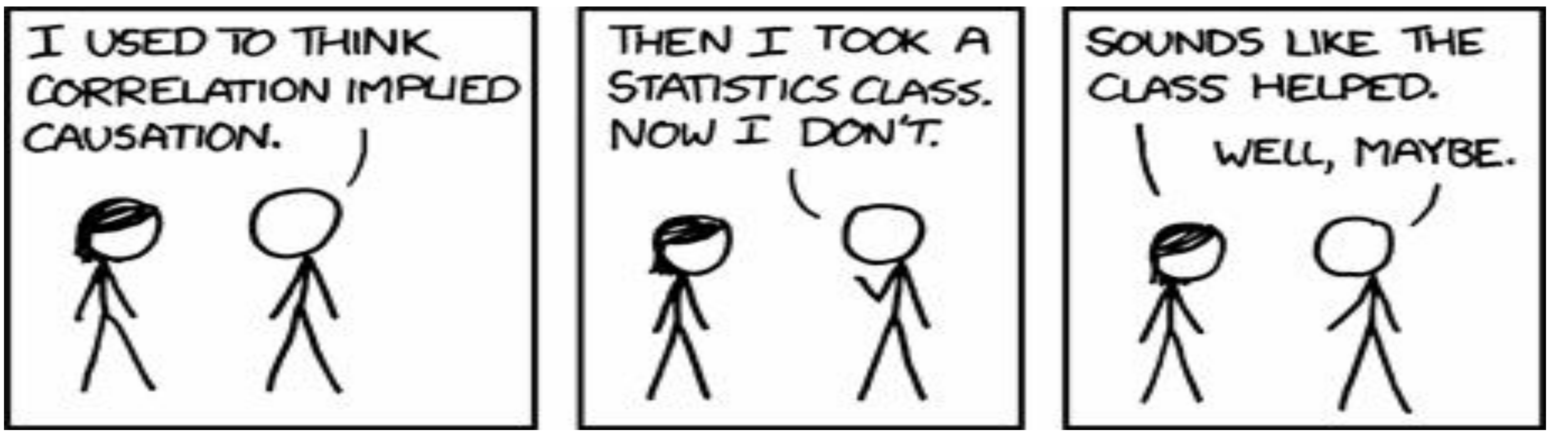
r = 0.4
Outlier
r = 0.7
Outlier removed

Correlation

- Normal distribution of X and Y
  - Histograms and descriptive statistics
- Linear relationship between X and Y
  - Scatterplot
  - Histogram of residuals
- Homoscedasticity
  - Same as with linear relationship

Correlation vs. causation

Correlation does not imply causation.



- Correlation is necessary but not sufficient condition for causation.

Correlation vs. causation

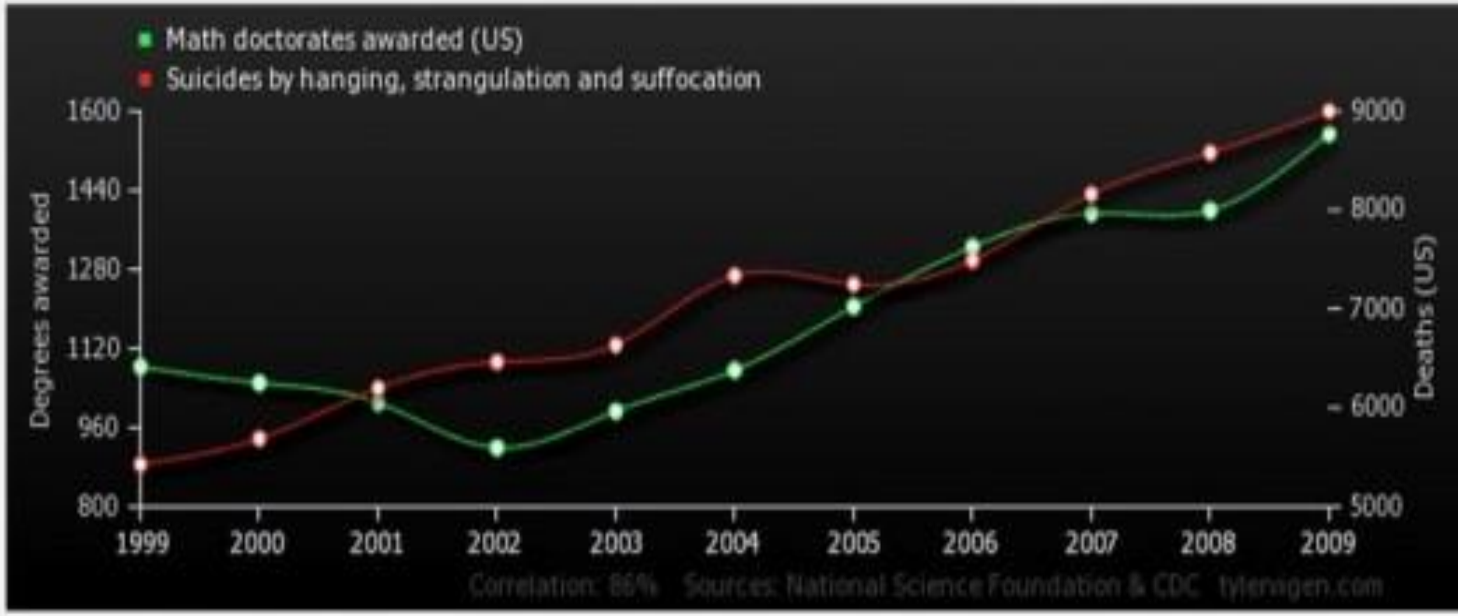
General patterns:

- X causes Y and Y causes X (bidirectional causation):
  - Democracies trade more, therefore trade increases democracy.
- Y causes X (reverse causation):
  - The more firemen is sent to a fire, the more damage is done.
- X and Y are consequences of common cause:
  - There is a correlation between ice cream consumption and street criminality (both more prevalent during summer).
- There is no connection between X and Y (coincidence):
  - Number of meaningless “funny correlations”.

Math doctorates awarded (US)

correlates with

Suicides by hanging, strangulation and suffocation



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Math doctorates awarded (US) *Degrees awarded (National Science Foundation)*	1,083	1,050	1,010	919	993	1,076	1,205	1,325	1,393	1,399	1,554
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

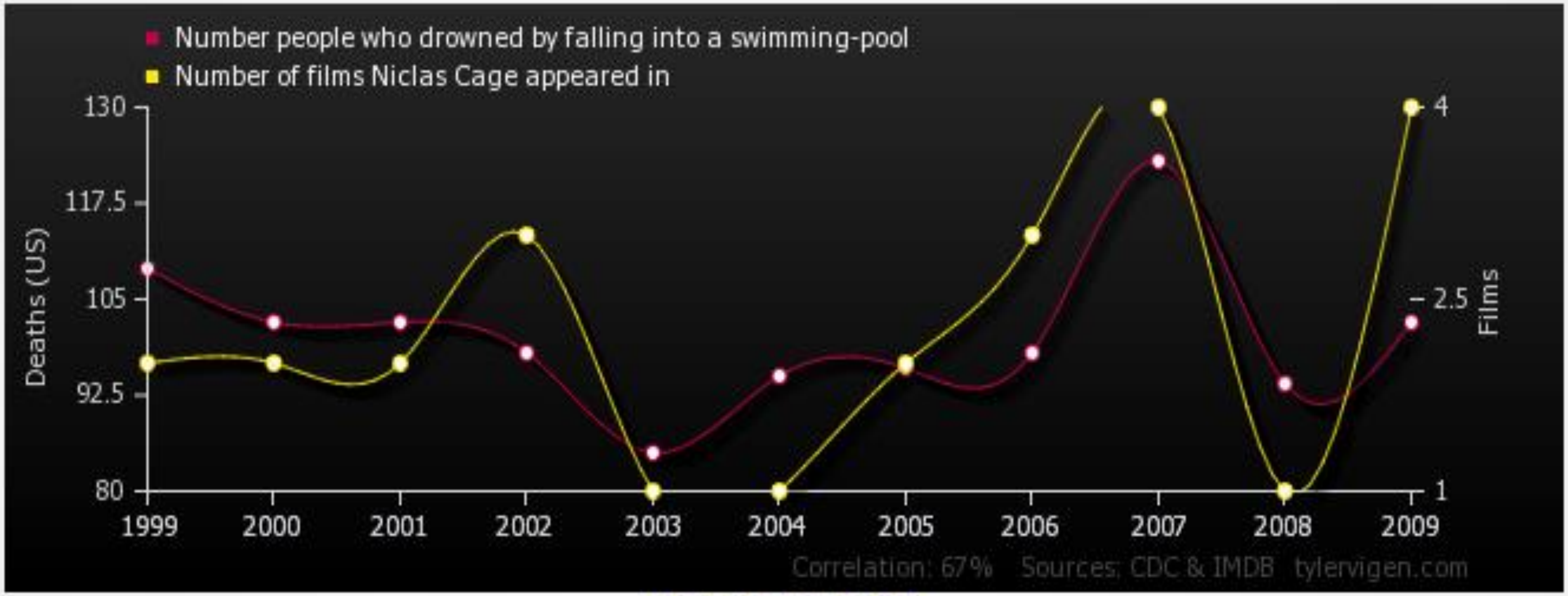
Correlation: 0.860176

Permalink · Not interesting

Number people who drowned by falling into a swimming-pool

correlates with

Number of films Nicolas Cage appeared in

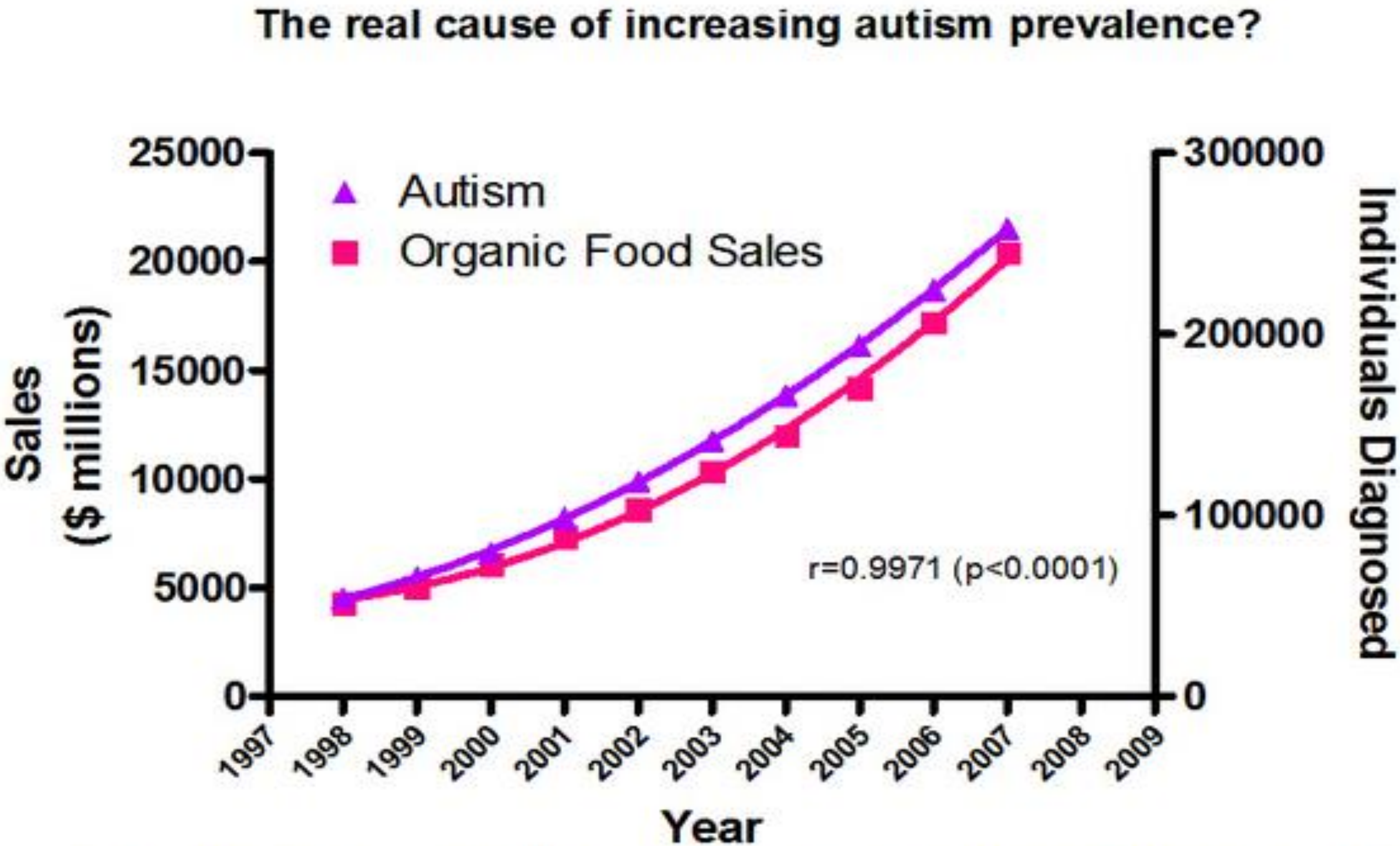


Upload this image to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool Deaths (US) (CDC)	109	102	102	98	85	95	96	98	123	94	102
Number of films Nicolas Cage appeared in Films (IMDB)	2	2	2	3	1	1	2	3	4	1	4

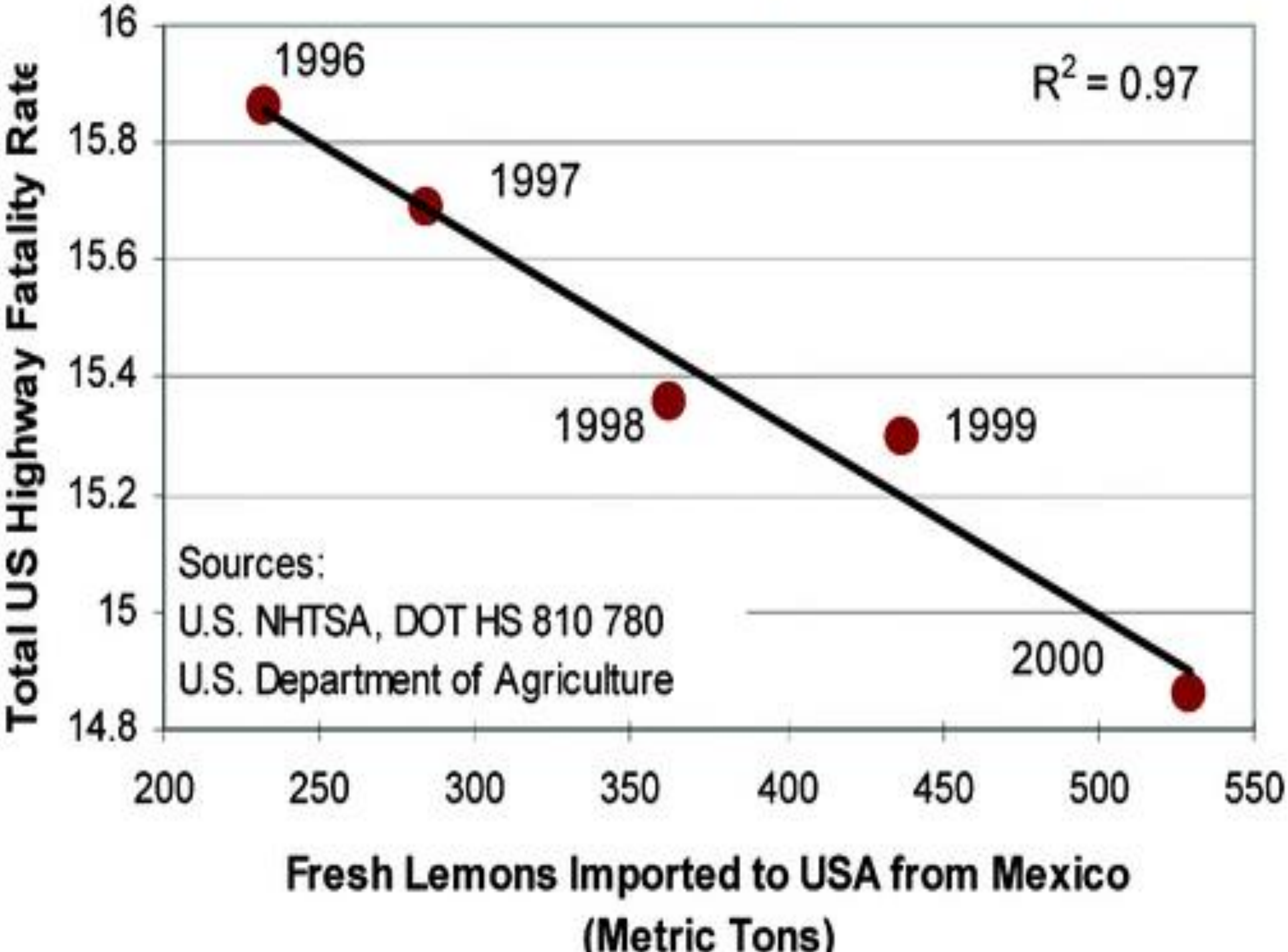
Correlation: 0.666004

4. Eating organic food causes autism.



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

7. Mexican lemon imports prevent highway deaths.



pubs.acs.org

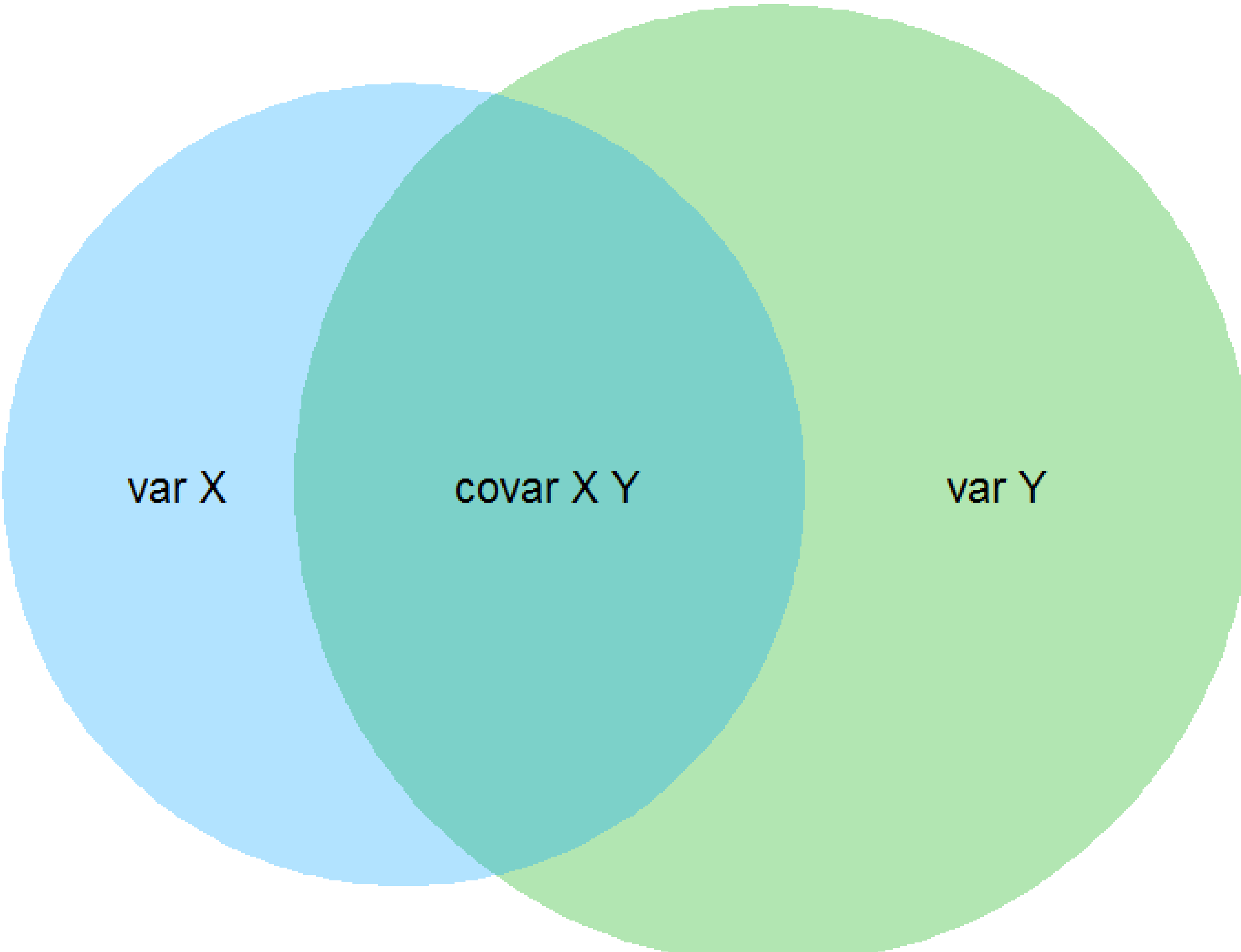
Correlation: example

- Assume we have 2 variables: X and Y.

X	Y
1	0
2	1
1	4
6	8
7	4

- What is correlation (r) of these two variables?

- Correlation = covariance / combined total variance.



- First: we calculate **variance of variables**.
- *mean(x)* = 3.4; *mean(y)* = 3.4
- R command = *var()*

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

X	(x – m)	dev.	dev.^2
1	(1 – 3.4)	-2.4	5.76
2	(2 – 3.4)	-1.4	1.96
1	(1 – 3.4)	-2.4	5.76
6	(6 – 3.4)	2.6	6.76
7	(7 – 3.4)	3.6	12.96
sum	0	0	33.2

Y	(y – m)	dev.	dev.^2
0	(0 – 3.4)	-3.4	11.56
1	(1 – 3.4)	-2.4	5.76
4	(4 – 3.4)	0.6	0.36
8	(8 – 3.4)	4.6	21.16
4	(4 – 3.4)	0.6	0.36
sum	0	0	39.2

- **s^2(X)** = 33.2 / 4 = **8.3**; **s^2(Y)** = 39.2 / 4 = **9.8**

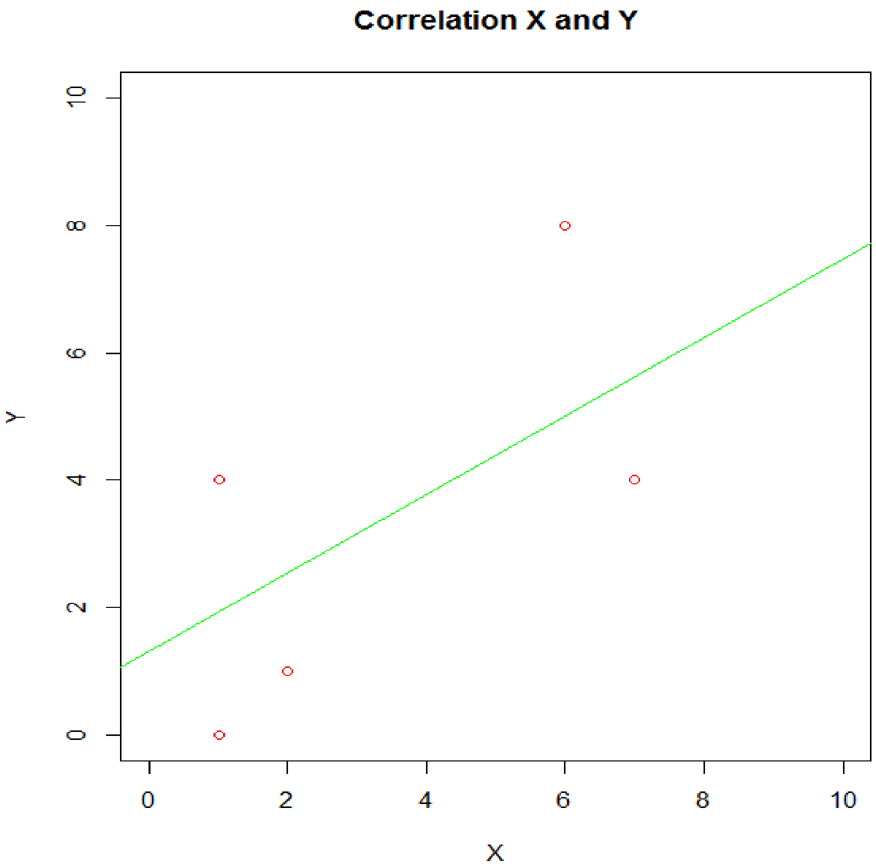
- Second: we calculate **covariance of variables**.
- Covariance is a sum of deviation products of two variables divided by n–1.

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

(x – m)	**(y – m)**	**cross-prod.**
(1 – 3.4)	(0 – 3.4)	8.16
(2 – 3.4)	(1 – 3.4)	3.36
(1 – 3.4)	(4 – 3.4)	-1.44
(6 – 3.4)	(8 – 3.4)	11.96
(7 – 3.4)	(4 – 3.4)	2.16
0	0	**24.2**

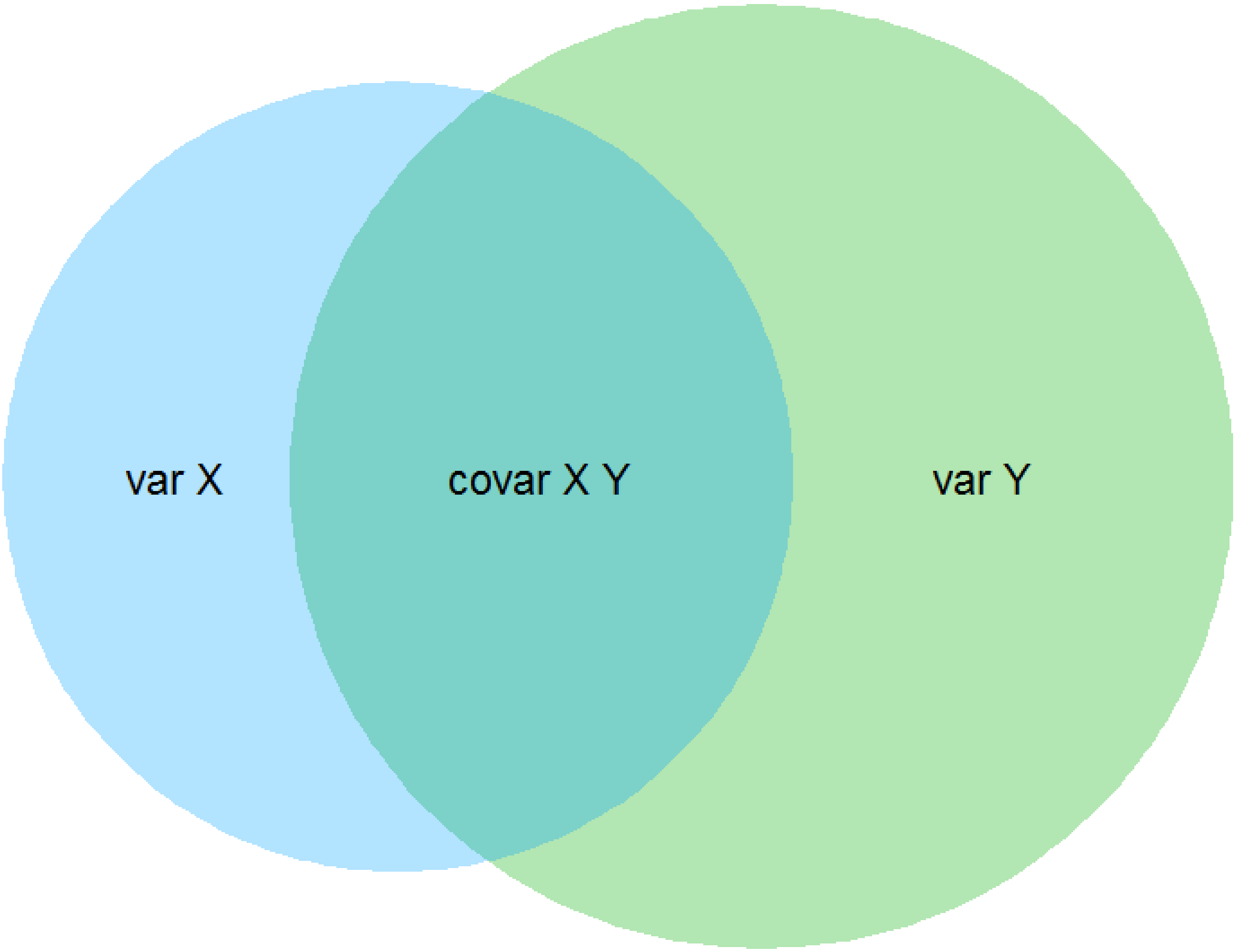
cov(X, Y) = 24.2 / 4 = **6.05;** R command = *cov()*

- Third: we divide X, Y covariance by square rooted product of X and Y variances.
  - **r = cov(X, Y) / sqrt(var(X) * var(Y))**
  - **r** = 6.05 / sqrt(8.3 * 9.8) = **0.67**
  - R command: *cor()*



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2}}$$

- Correlation = covariance / combined total variance.



(Linear) regression

- Regression is a statistical method used to **predict scores on an outcome variable based on scores of one ore more predictor variables**.
- Linear regression: models linear relationship.
- Bivariate (simple) linear regression: uses only one predictor variable.
- Multivariate (multiple) linear regression: uses more than one predictor variable.

Regression: terminology / notation

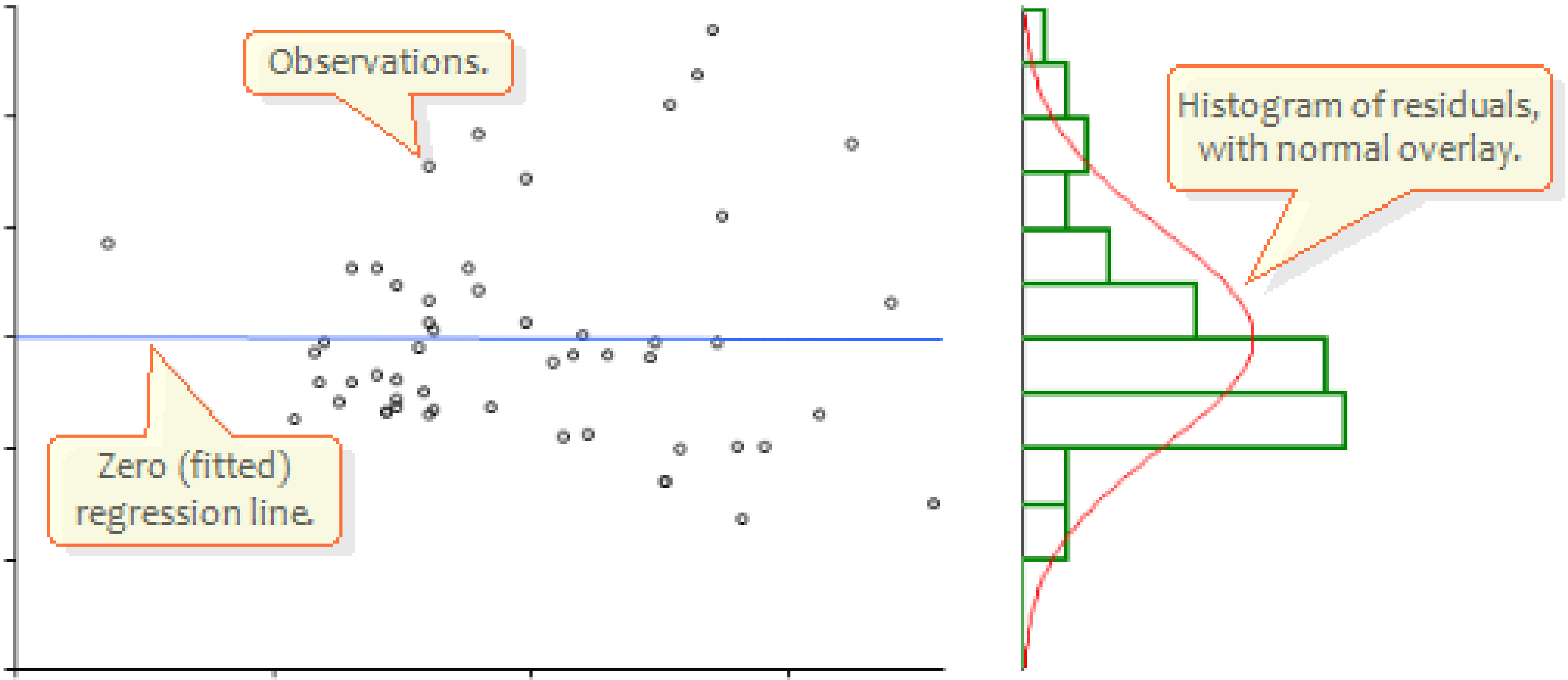
X	Y
cause	effect
independent variable	dependent variable
predictor variable	outcome variable
explanatory variable	response variable

α, a, b, β0, B0, m	β, B, b	ε, e
intercept	slope	error / residual
constant	coefficient	
alpha	Beta	

Linear regression: assumptions

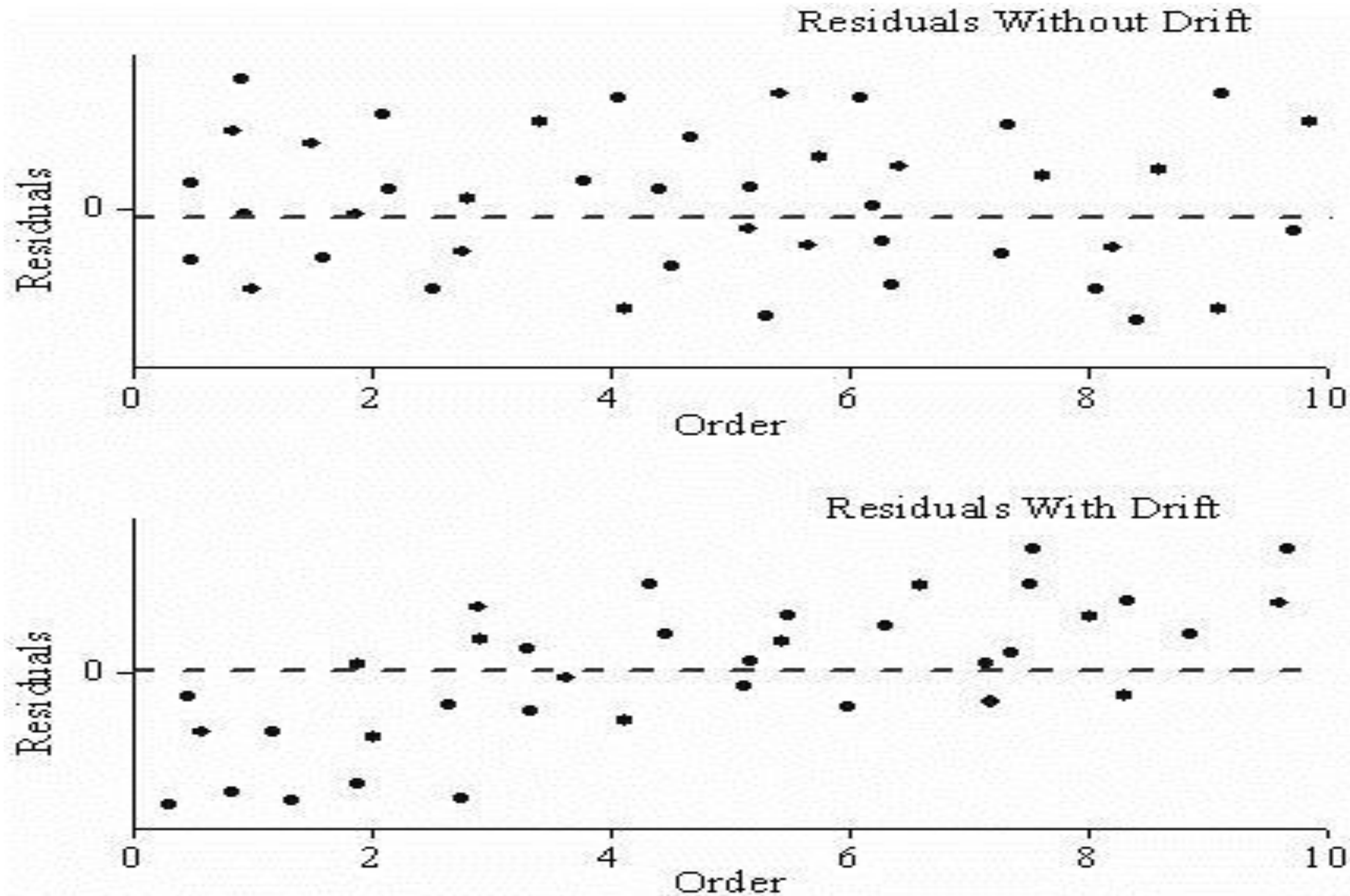
- **Independence of observations** (random sampling).
- Normal distribution of Y.
- Linear relationship between X and Y.
- **Normal distribution of residuals.**
- Homoscedasticity.
- **Independence of residuals (over time).**
- Applicable for continuous variables.
- Sensitive to outliers.

Normal distribution of residuals



Draper & Smith 1998

Independence of residuals



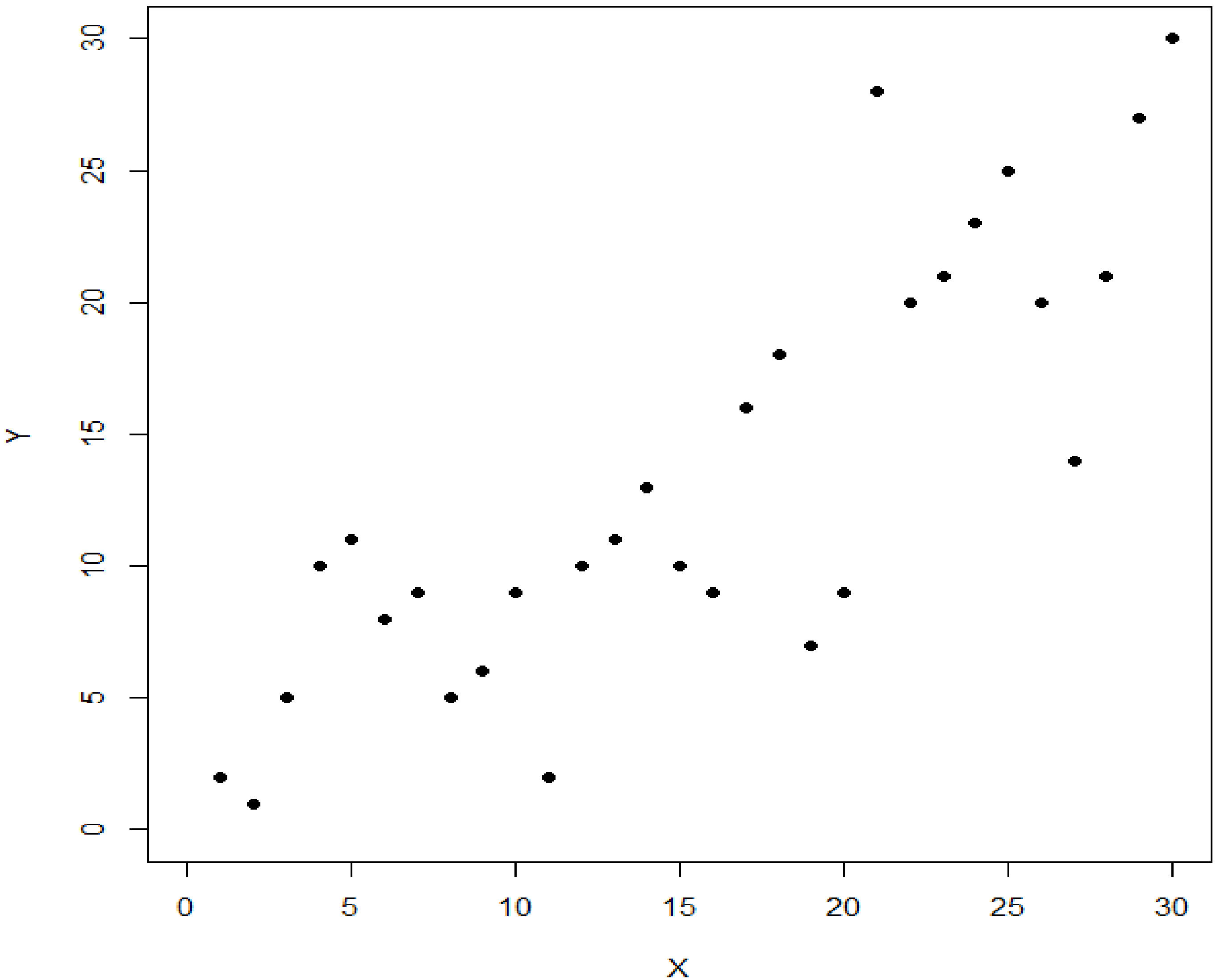
OriginLab 2015

Linear relationship

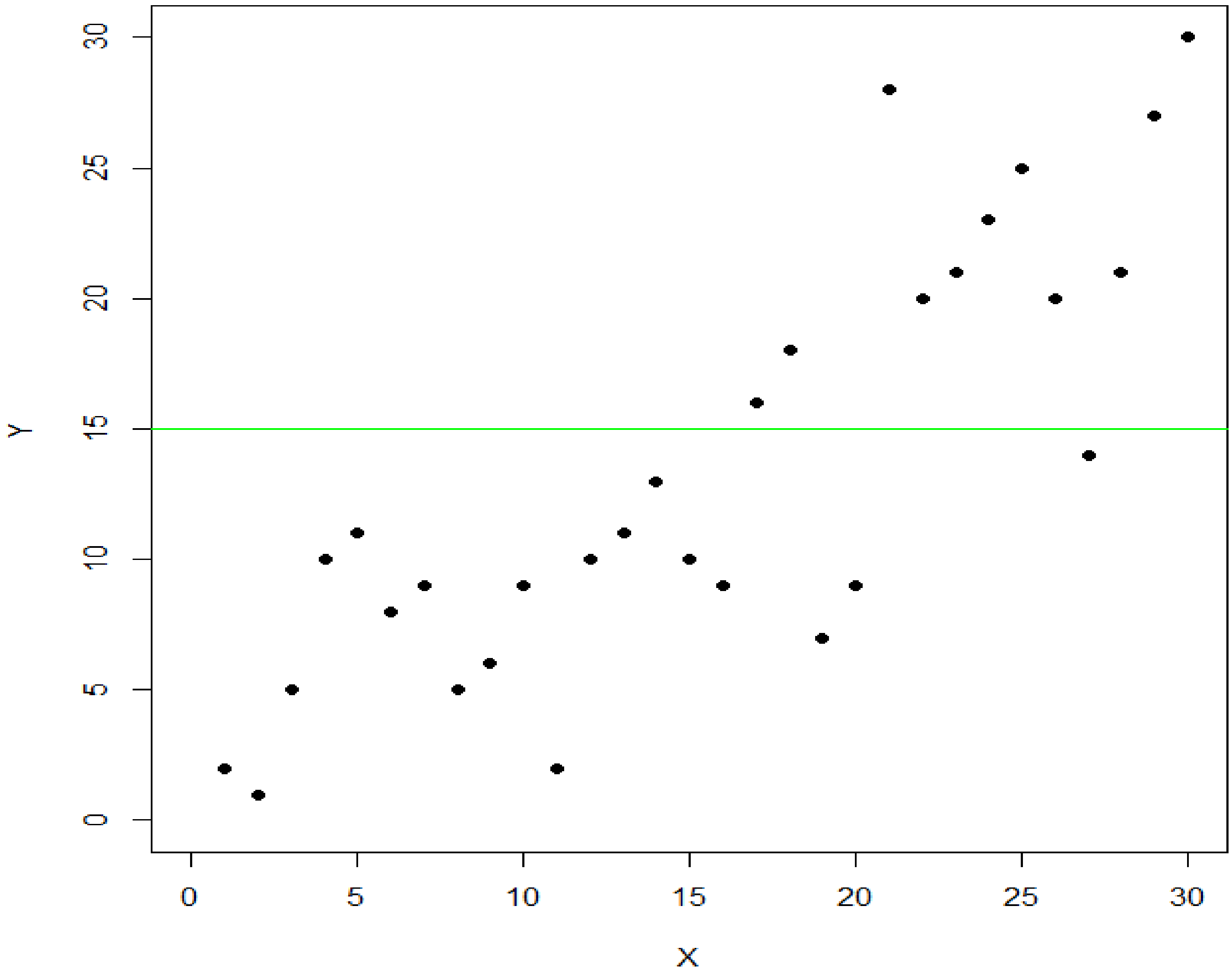
- A relationship where two variables are related **in the first degree**.
- Meaning the **power of variables is 1**.
- Linear relationship is represented by formula:
- **Y = a + bX**
- $Y = \beta 0 + \beta 1X + \varepsilon$; population regression function
- Y = a + bX + e ; sample regression function
- Y' = 0.75 + 0.425*X + 2.791; sample regression line

- Linear relationship is graphically represented by **straight line**.

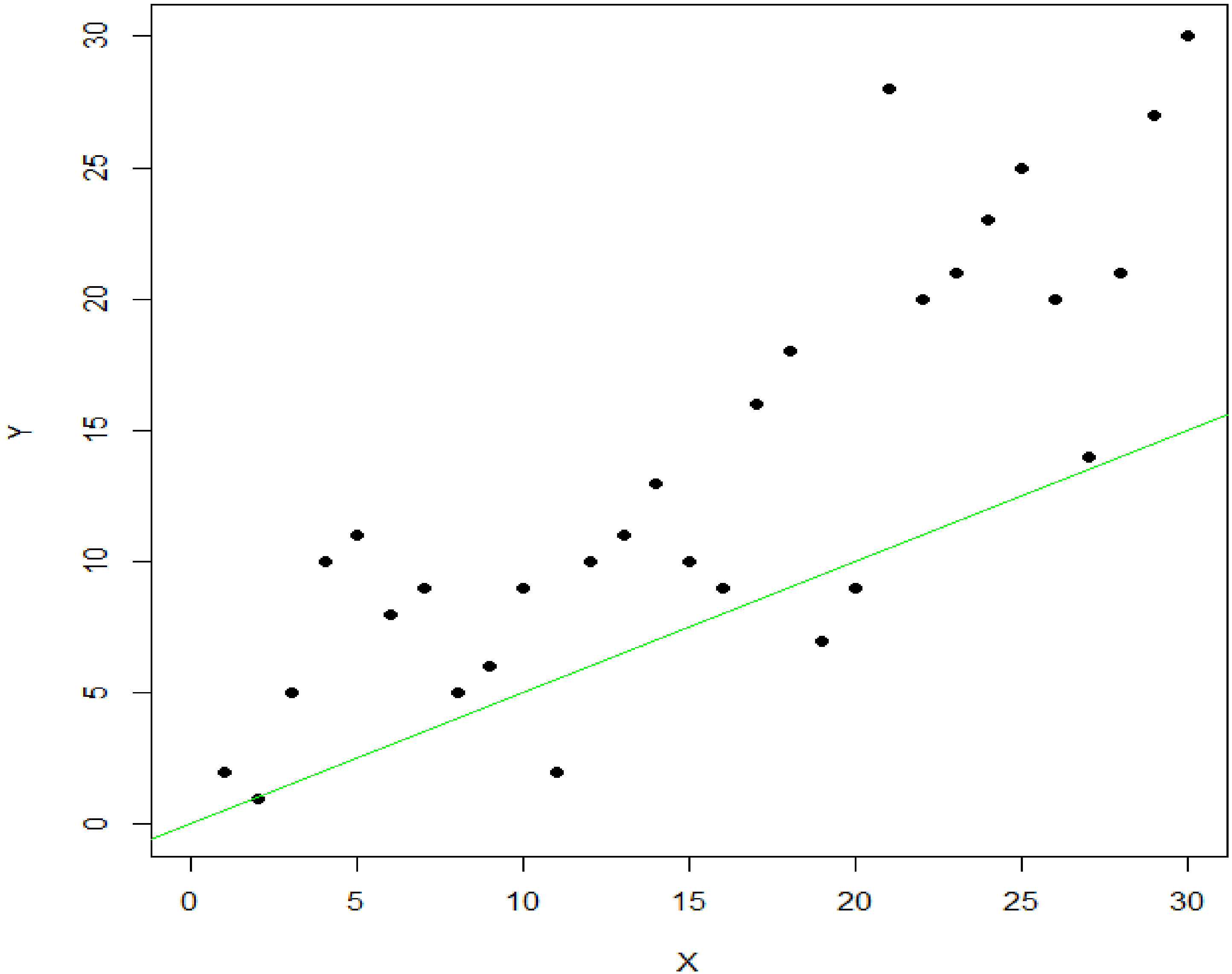
Fitting a straight line



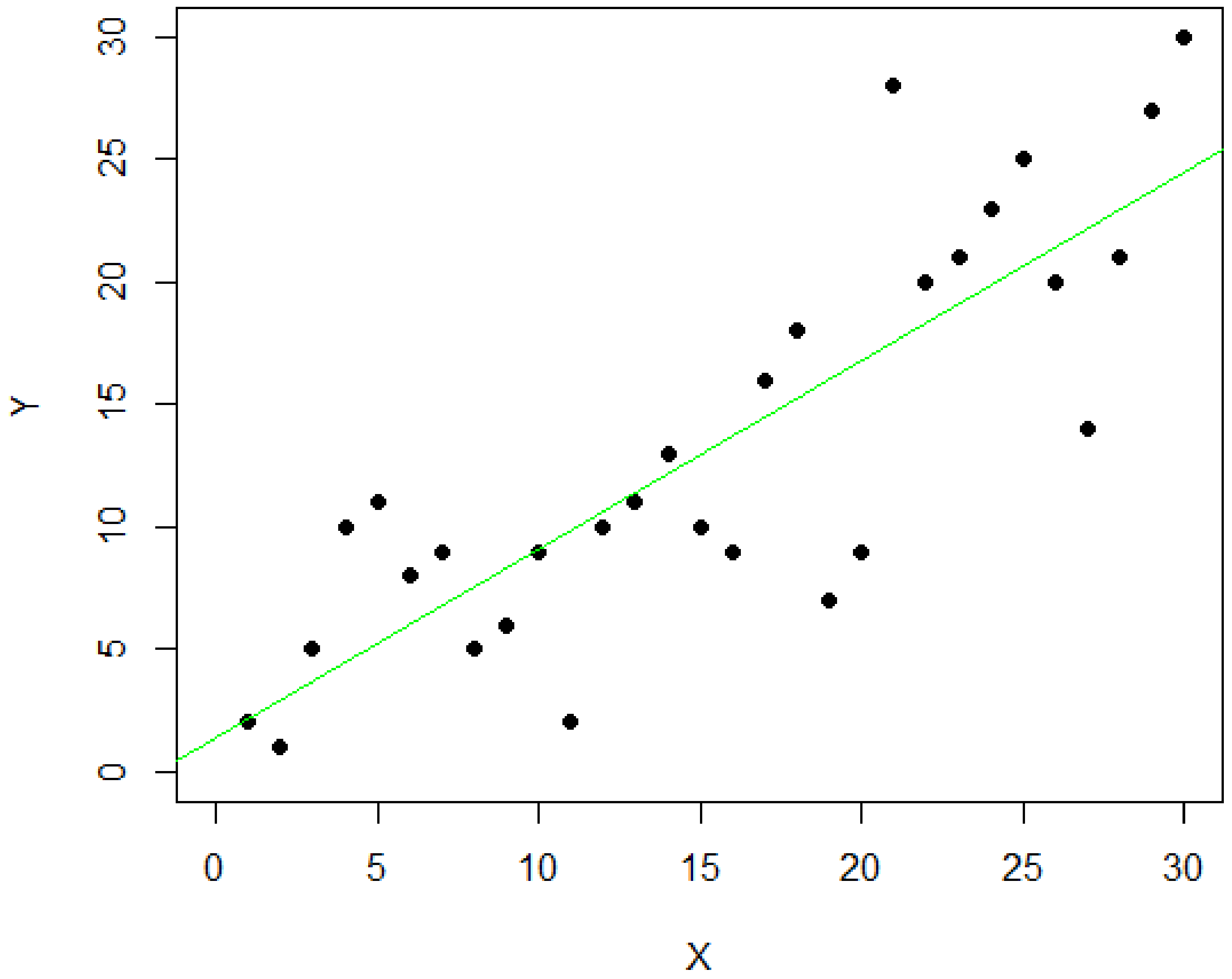
Fitting a straight line



Fitting a straight line



Fitting a straight line

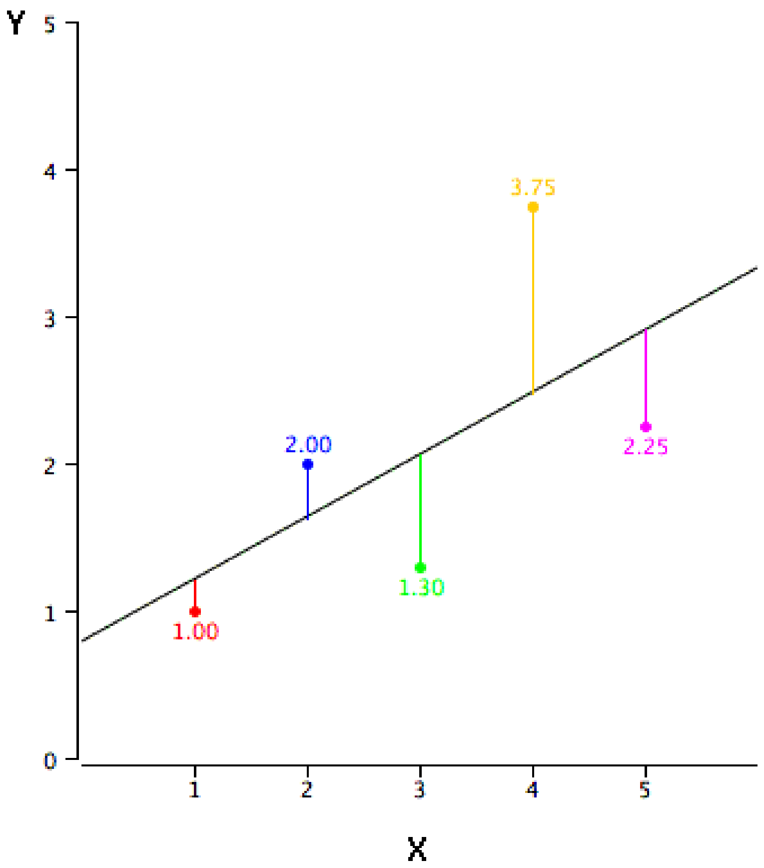


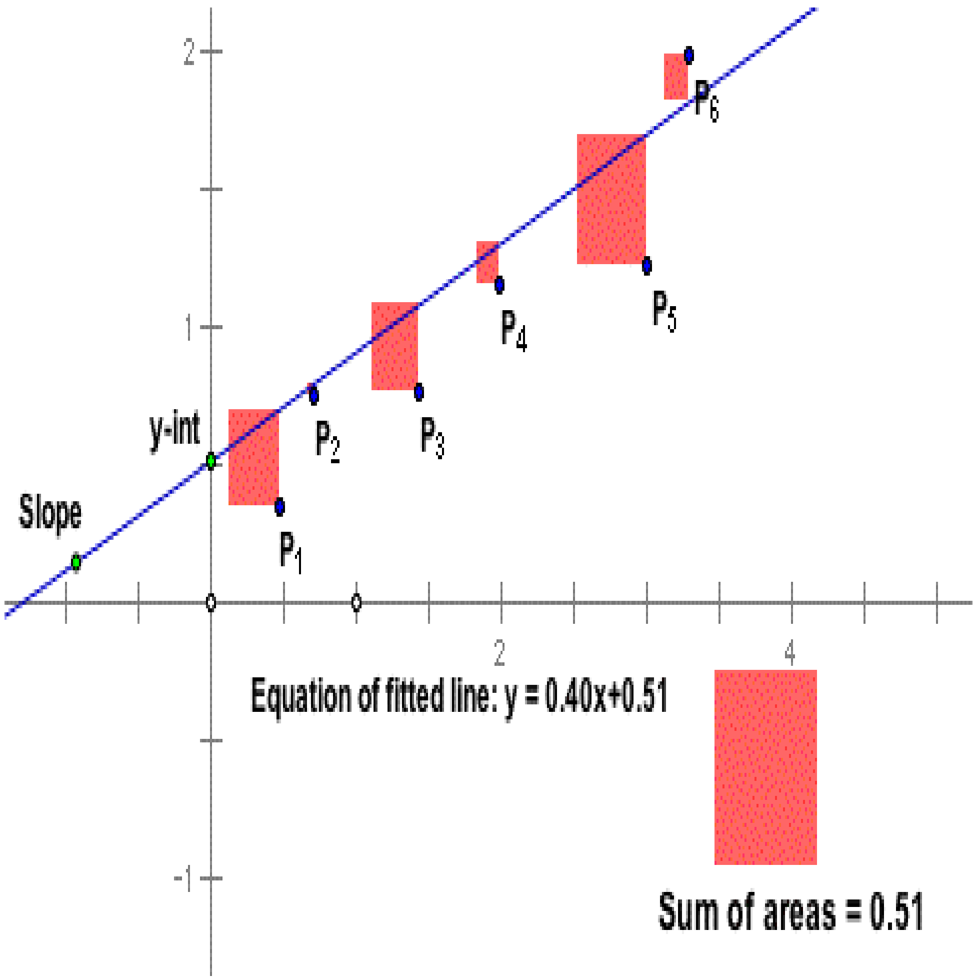
Ordinary least squares

- **Ordinary least squares (OLS):** estimates parameters (intercept and slope) in a linear regression model.
- **Minimizes squared vertical distances** between the observations (Y) and the straight line (predicted value of Y = Y’).
- **Residual = (Y - Y’)**

- $\sum (Y - Y') = 0$; $\sum (Y - Y')^2 >= 0$
- **OLS: Y’ = min** $\sum (Y - Y')^2$

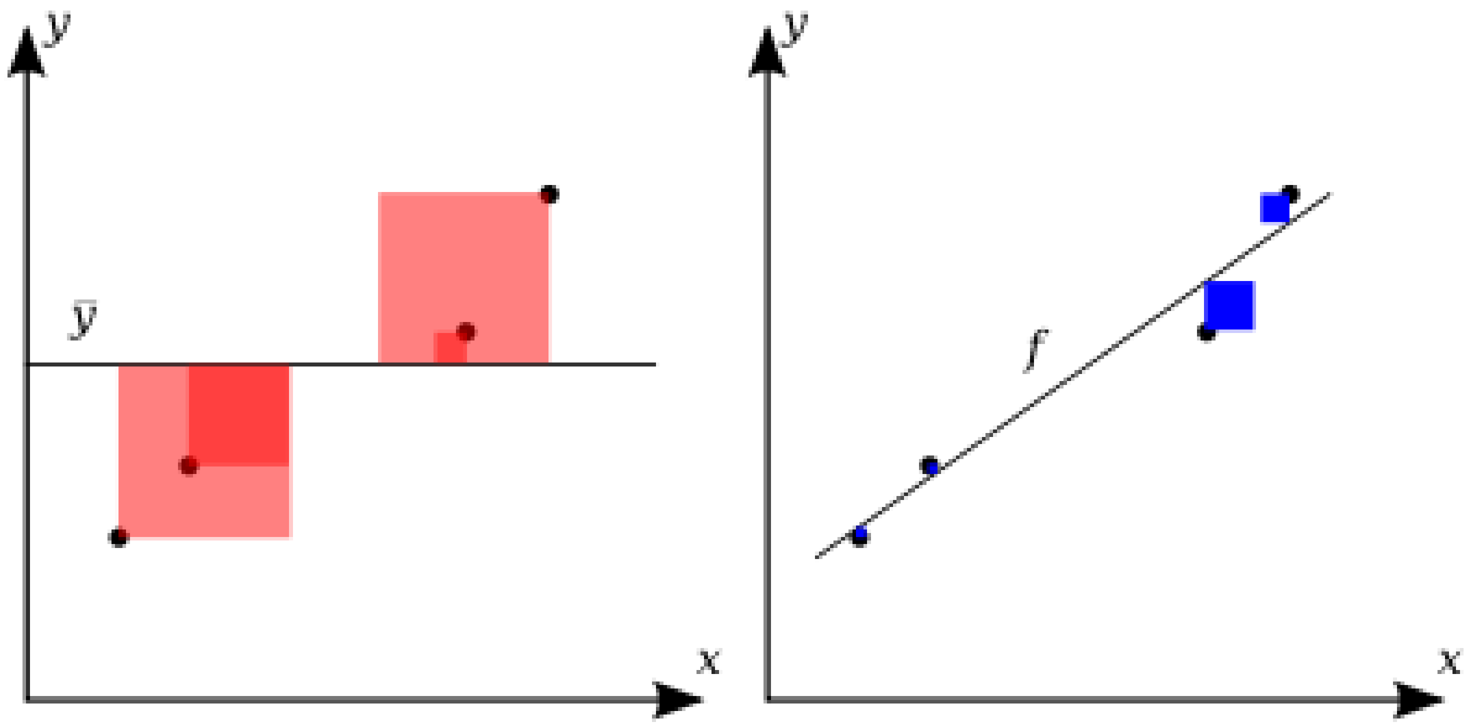
Ordinary least squares





Ordinary least squares

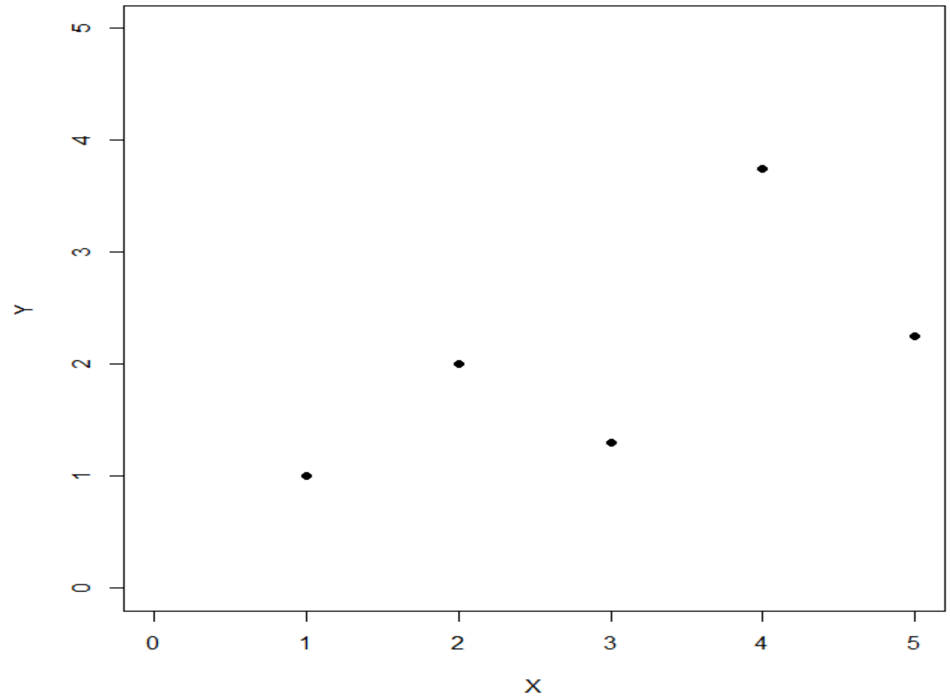
- Comparison of mean and OLS estimation.



Linear regression: example

- Assume we have two variables: X and Y.

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25



- To what extent X explains Y?

Linear regression: example

- Statistics for calculating regression line:

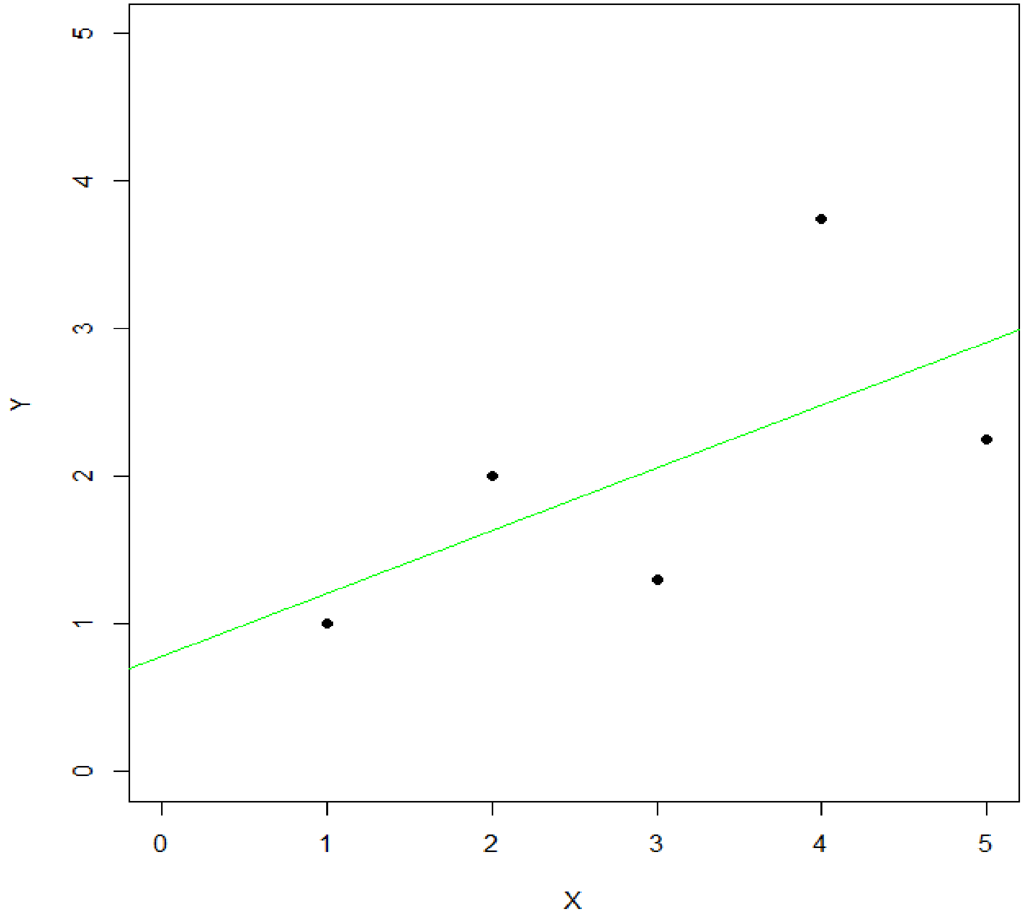
m(X)	m(Y)	s(X)	s(Y)	r(X, Y)
3	2.06	1.581	1.072	0.627

- The **slope (b): r(X, Y) * s(Y) / s(X)**
- The **intercept (a): m(Y) – b*m(X)**

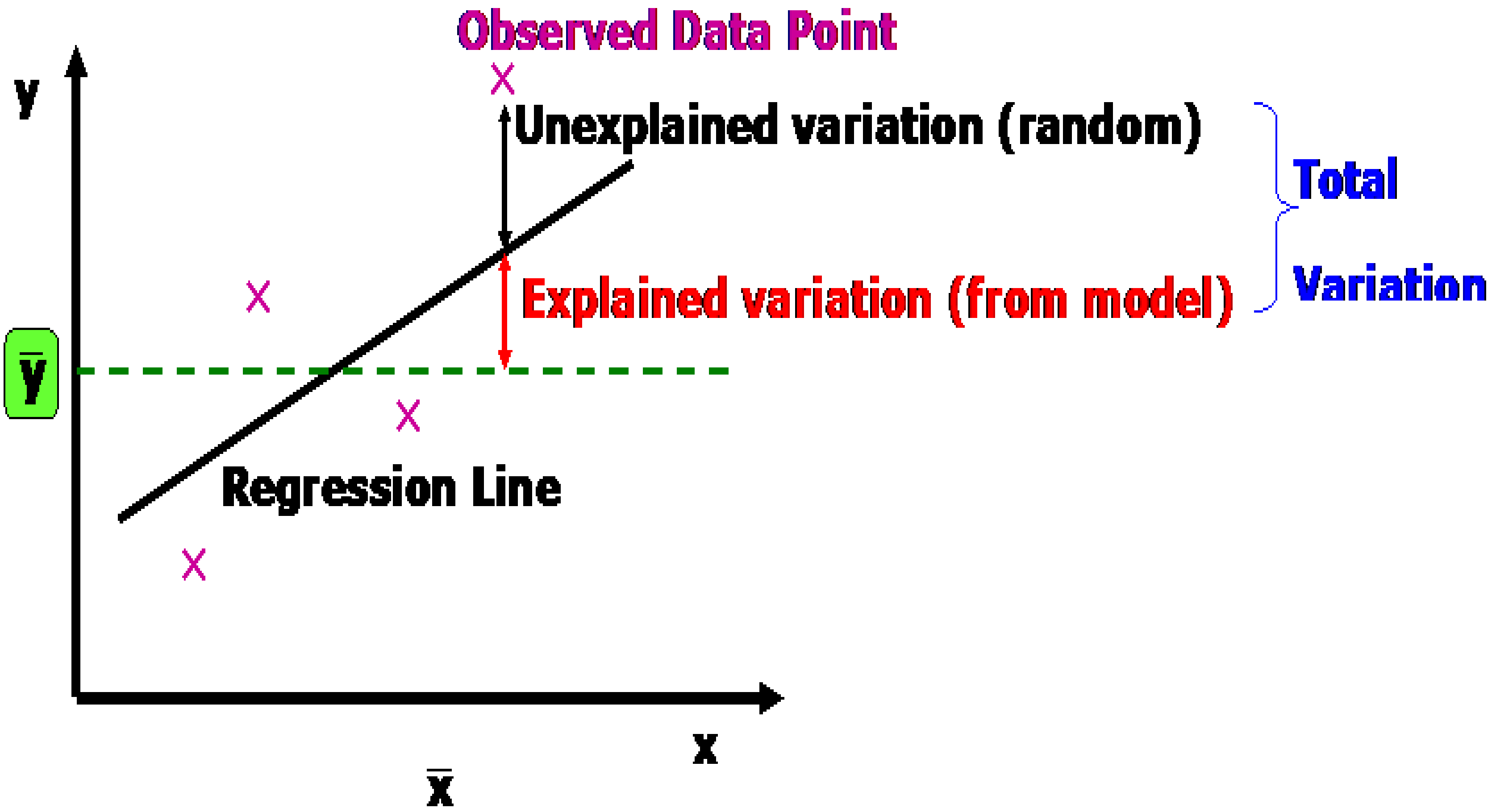
- **b** = 0.627 * 1.072 / 1.581 = **0.425**
- **a** = 2.06 – 0.425 * 3 = **0.75**

Linear regression: example

- Fitting a straight line by using OLS.



Total / unexplained / explained variation



Linear regression: example

- **Residual:** difference between observed values Y and predicted values Y' .

X	Y	Y'	Y – Y'	(Y – Y')^2
1	1	1.21	-0.210	0.044
2	2	1.653	0.365	0.133
3	1.3	2.060	-0.760	0.578
4	3.75	2.485	1.265	1.600
5	2.25	2.910	-0.660	0.436
sum			0	2.791

Linear regression: example

- **Model** is a representation of the relationship between variables. Linear regression model predicts (models) values of Y based on values of X.

- Model is represented by formula in a form of **linear equation:** $Y' = a + bX + e$.

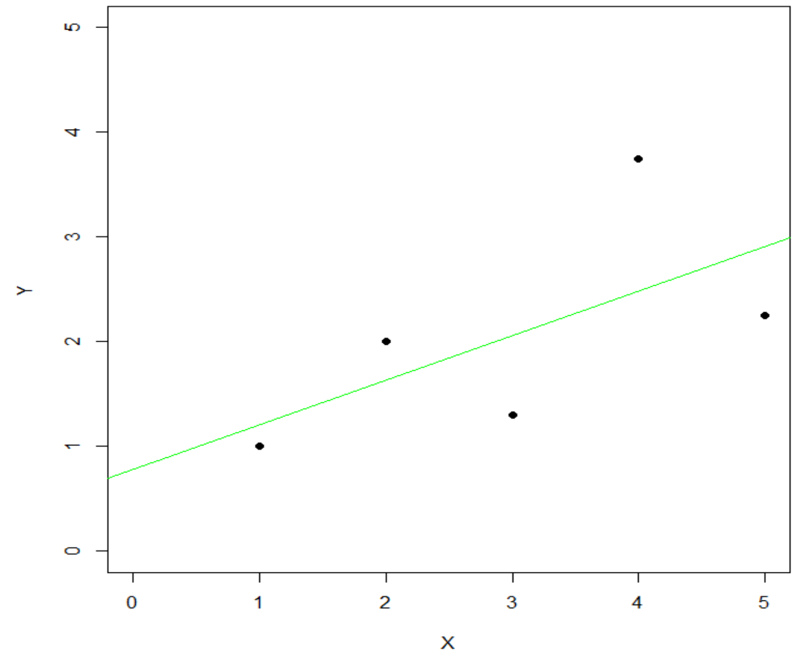
- Model in example: $Y' = 0.75 + 0.425*X + 2.791$.

- R command: *lm()*

Linear regression: interpretation

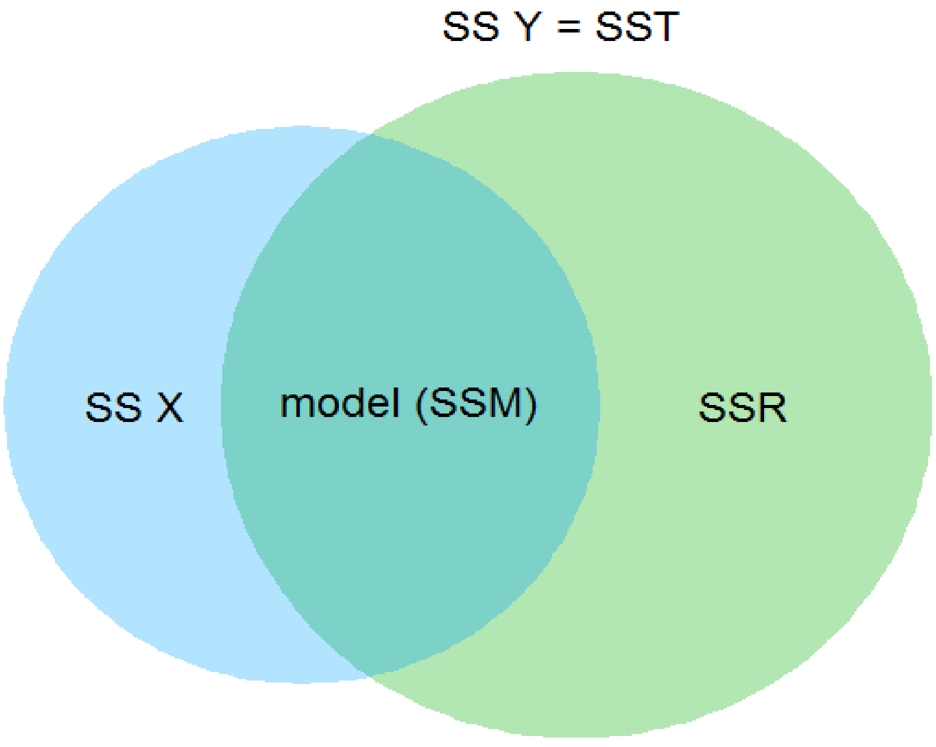
- Model in example: Y = 0.78 + 0.425*X
- **Intercept:** value of Y when value of X = 0.
- **Slope:** change in Y when X increases by 1 unit.
- **Error:** unexplained variance of Y.

- What is the Y’ for X = 2?
- Y’ = 0.75 + (0.425)*2
- Y’ = 0.75 + 0.850 = 1.6



Coefficient of determination

- CoD (**R^2**) indicates proportion of Y explained variation (SSM) to Y total variation (SST) = **SSM / SST**.
- SST = SSM (explained var.) + SSR (unexplained var.)



Coefficient of determination

- **Unexplained variation = difference between observed values of Y and predicted values of Y'** (regression line) = sum of squares of residuals (**SSR**).
- **Explained variation = difference between predicted values of Y' and mean of Y =** sum of squares of model (**SSM**).
- **Total variation = difference between observed values of Y and mean of Y** = SSE + SSR = sum of squares of total variation (**SST**).

- Explained variation (%) = SSM / SST = **coefficient of determination = R^2**

Coefficient of determination: example

Y’	mean Y	(Y’ – mY)	(Y’ – mY)^2
1.210	2.06	-0.850	0.72
1.653	2.06	-0.425	0.18
2.060	2.06	0	0
2.485	2.06	0.425	0.18
2.910	2.06	0.850	0.72
sum (SSM)			**1.81**

Y	Y’	Y – Y’	(Y – Y’)^2
1	1.210	-0.210	0.044
2	1.653	0.365	0.133
1.3	2.060	-0.760	0.578
3.75	2.485	1.265	1.600
2.25	2.910	-0.660	0.436
sum (SSR)			**2.791**

- SST = SSM + SSR = 1.81 + 2.791 = 4.59
- **R^2** = SSM / SST = 1.81 / 4.59 = 0.39 = **39 %**