Linear regression

Lukáš Lehotský & Petr Ocelík

ESS401 Social Science Methodology / MEB431 Metodologie sociálních věd

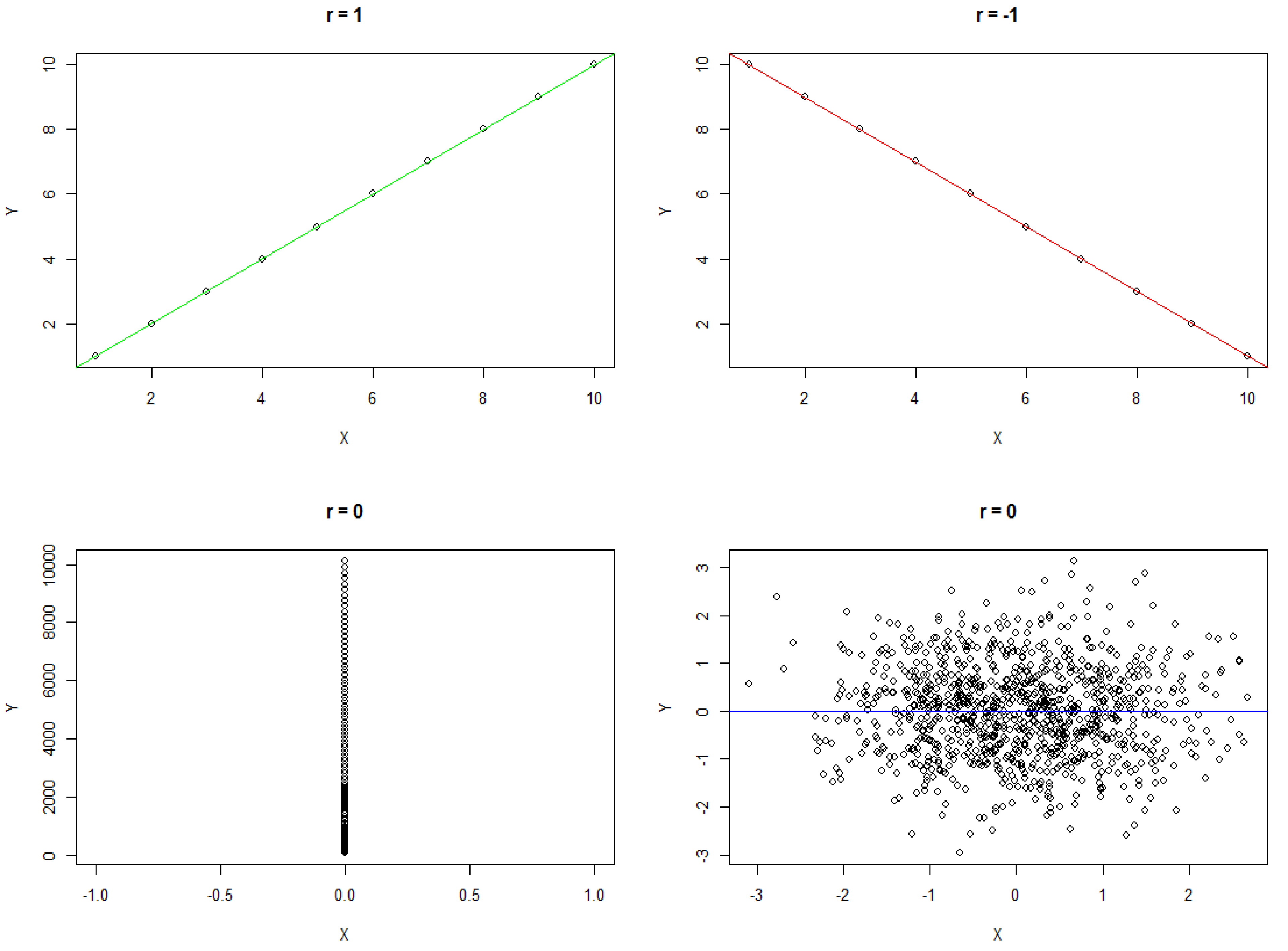
6th February 2017

Outline

- Refresh: Pearson's r correlation
- (Simple) linear regression

Refresh: Pearson’s r

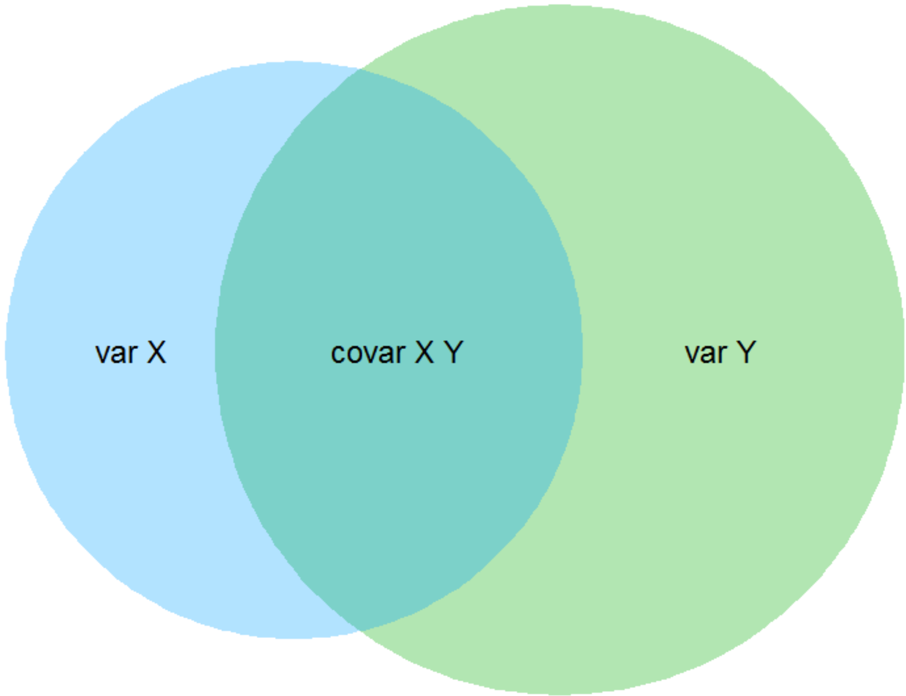
- Pearson’s rho product-moment correlation coefficient (**r**).
- Pearson’s r measures the **strength and direction of the linear relationship between two variables**.
- Ranges within <-1,1>
  - Perfect positive linear relationship = 1
  - Perfect negative linear relationship = -1
  - No linear relationship = 0
- Value does not depend on variables’ units.
- It is a **sample** (aggregative) **statistic.**


r = 1
r = -1
r = 0
r = 0
Y
X
10
8
6
4
2
-1.0
-0.5
0.0
0.5
1.0
10000
8000
6000
4000
2000
0
3
1
-1
-2
-3

Pearson's r: assumptions

- 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

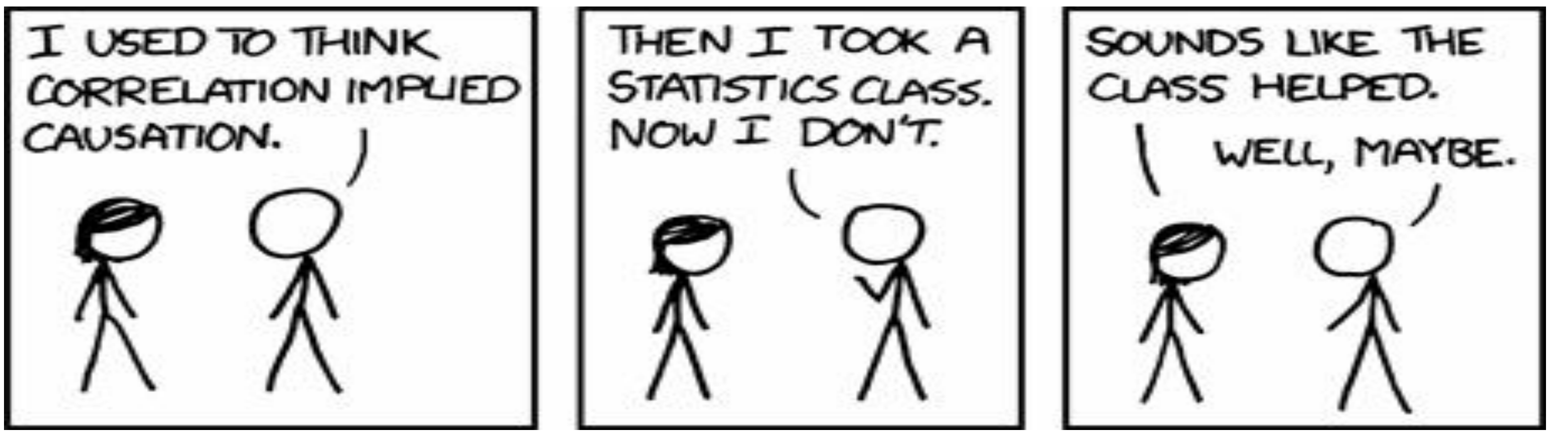
- Correlation = covariance / combined total variance.



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

Association vs. causation

Association does not imply causation!



xkcd.com/552/

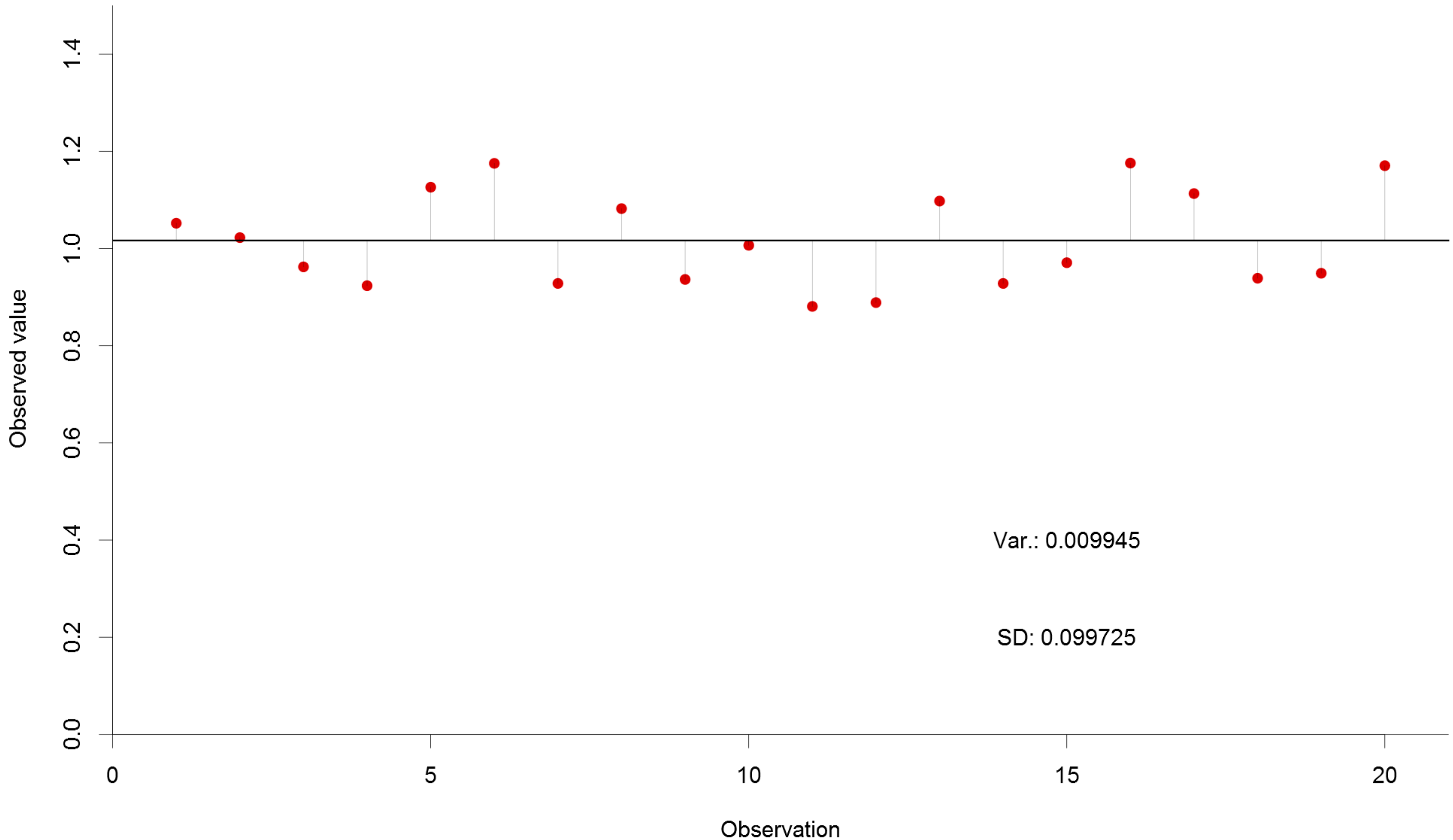
Correlation vs. causation

- 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”.
- More examples here: http://tinyurl.com/85jfu6y

Models

- *All models are wrong; some models are useful* (Box 1976).
- Models (not only mathematical!) reduce and represent the real-world phenomena.

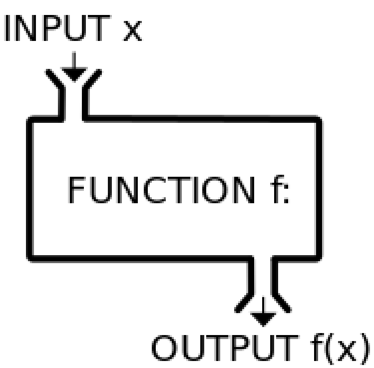
Mean as a model



Statistical models

- We need a mathematical function for statistical prediction.

→ **Function** changes **input** (values of predictor variable) to an **output** (value of outcome variable) according to specific **rule(s)**.



Y = f(X); Y = 2*X

if X = 2, then Y = 4

- For different relationships between quantities, different functions might be used.

wikimedia commons

(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 relationship

- A relationship where two variables are related **in the first degree**; i.e. the **power of variables is 1**.

- Linear relationship is represented by formula:

  outcome (dep. var.) = constant + coefficient*predictor + error

  $Y = \beta 0 + \beta 1X + \varepsilon$; population regression function

  $Y = a + bX + e$; sample regression function

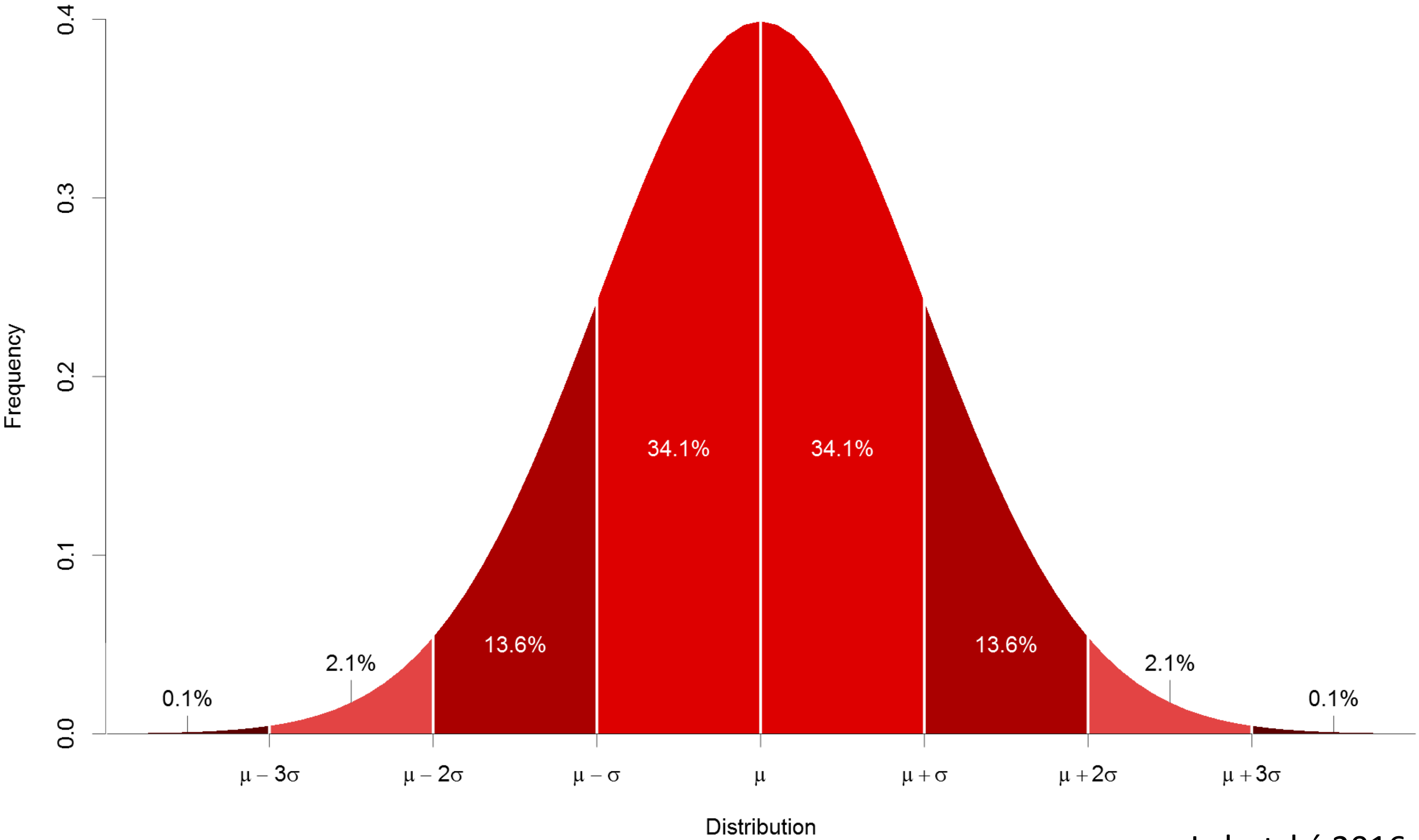
  $Y' = 0.75 + 0.425 * X + 2.791$; sample regression line

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

Linear regression: assumptions

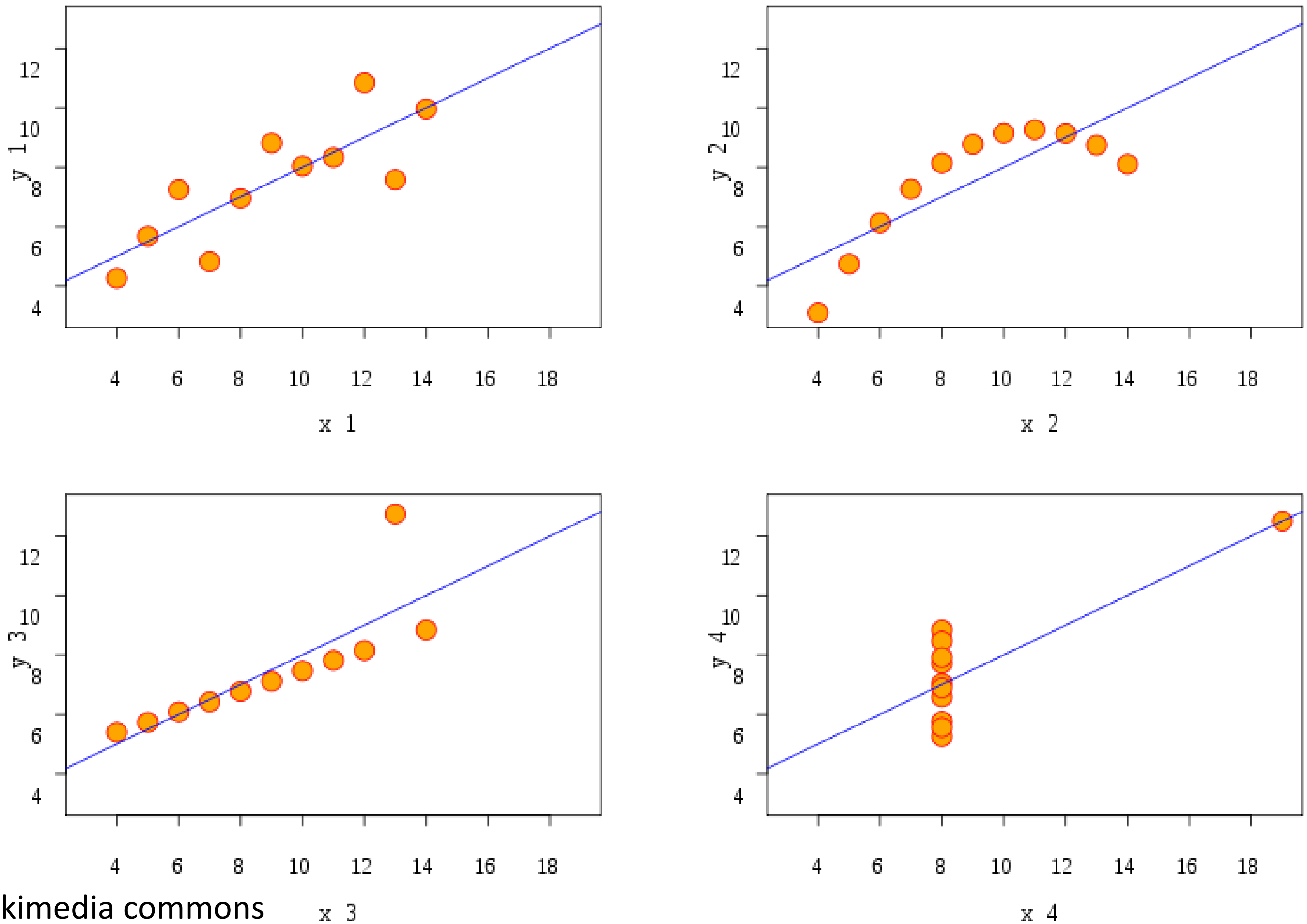
- **Independence of observations** (random sampling).
- Normal distribution of Y.
- Linear relationship between X and Y.
- **Normal distribution of residuals.**
- Homoscedasticity.
- **Independence of residuals (over time).**
- Applicable to metric level of measurement.
- Sensitive to outliers.

The standard normal distribution

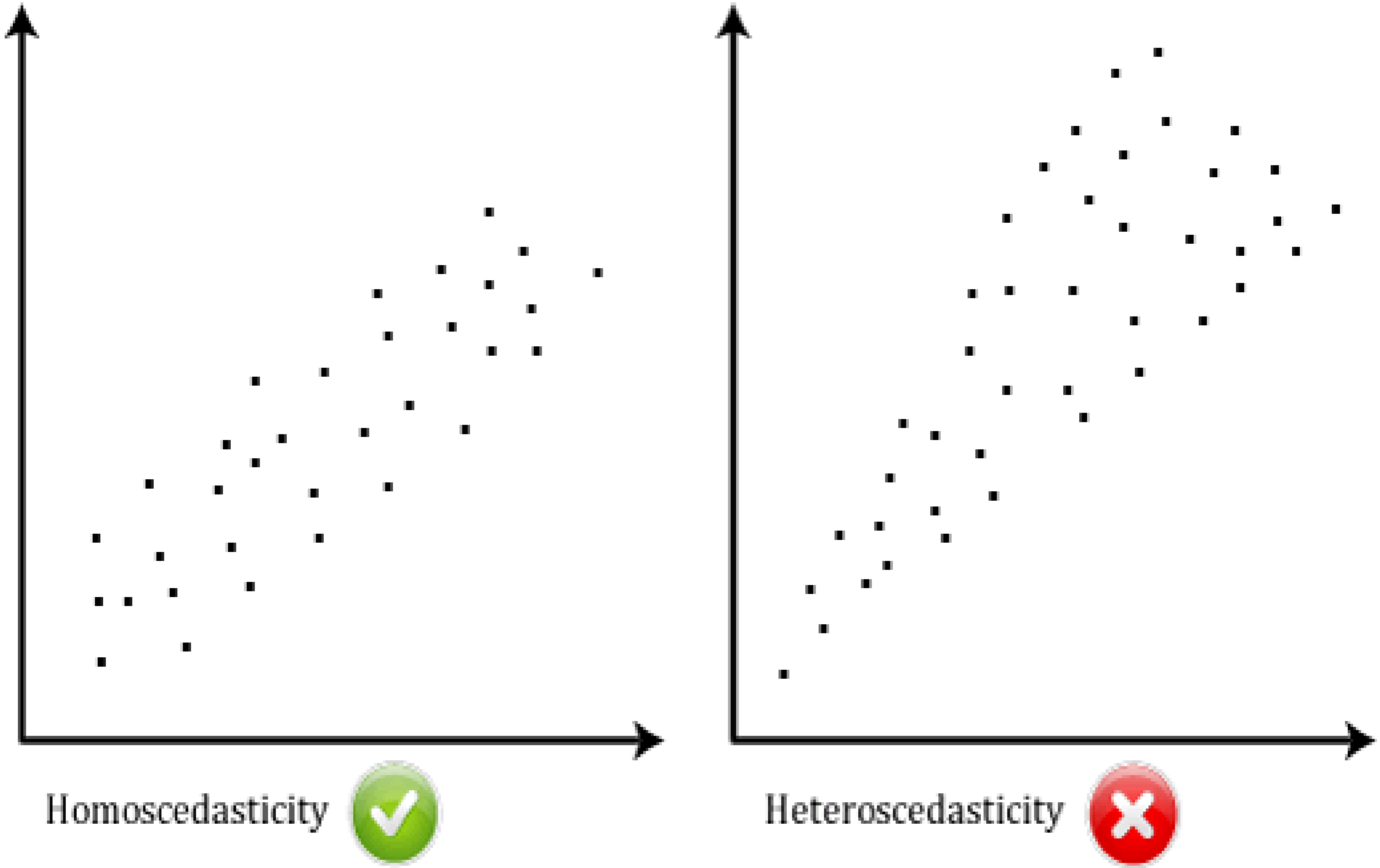


Lehotský 2016

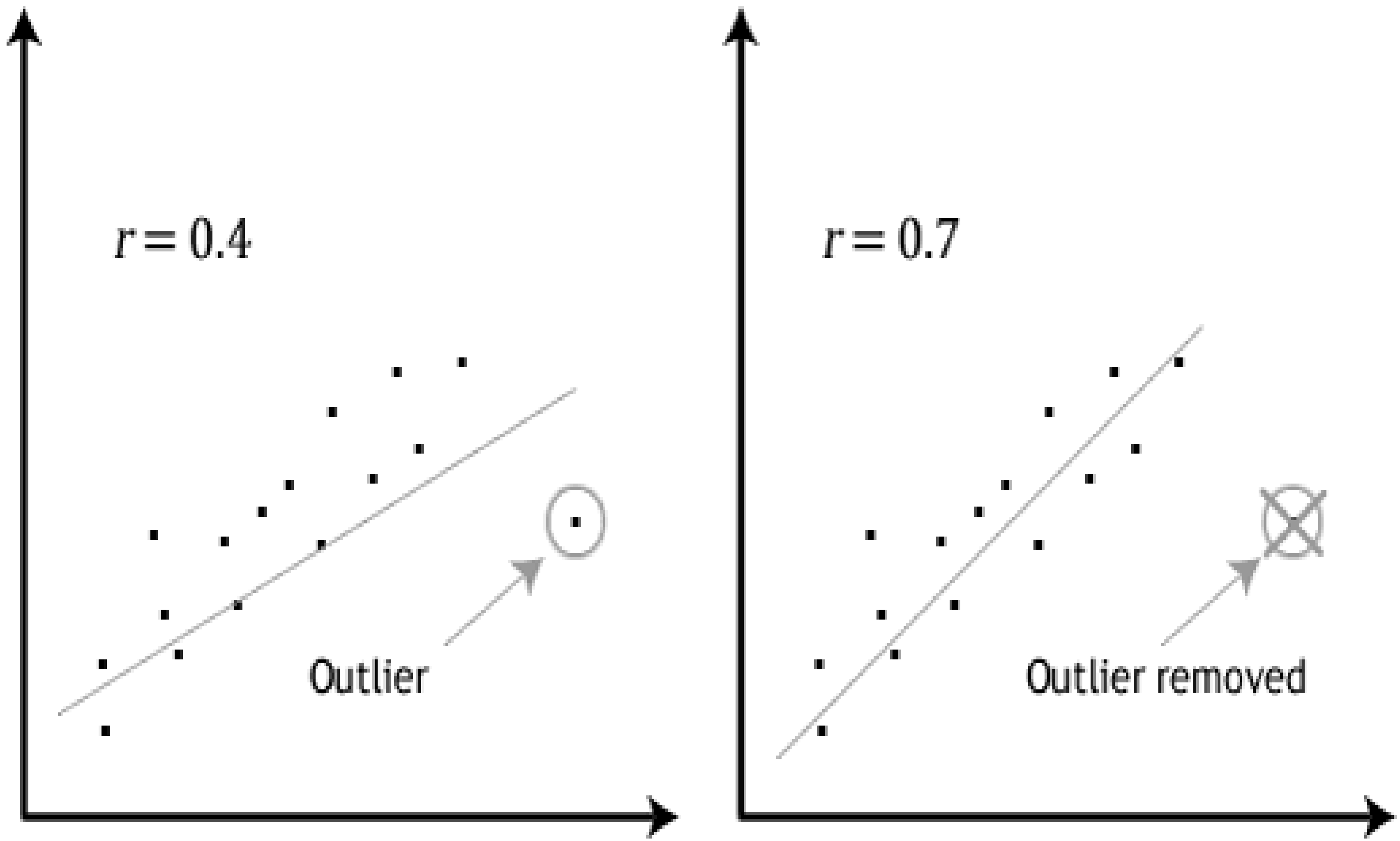
Anscombe's quartet



wikimedia commons

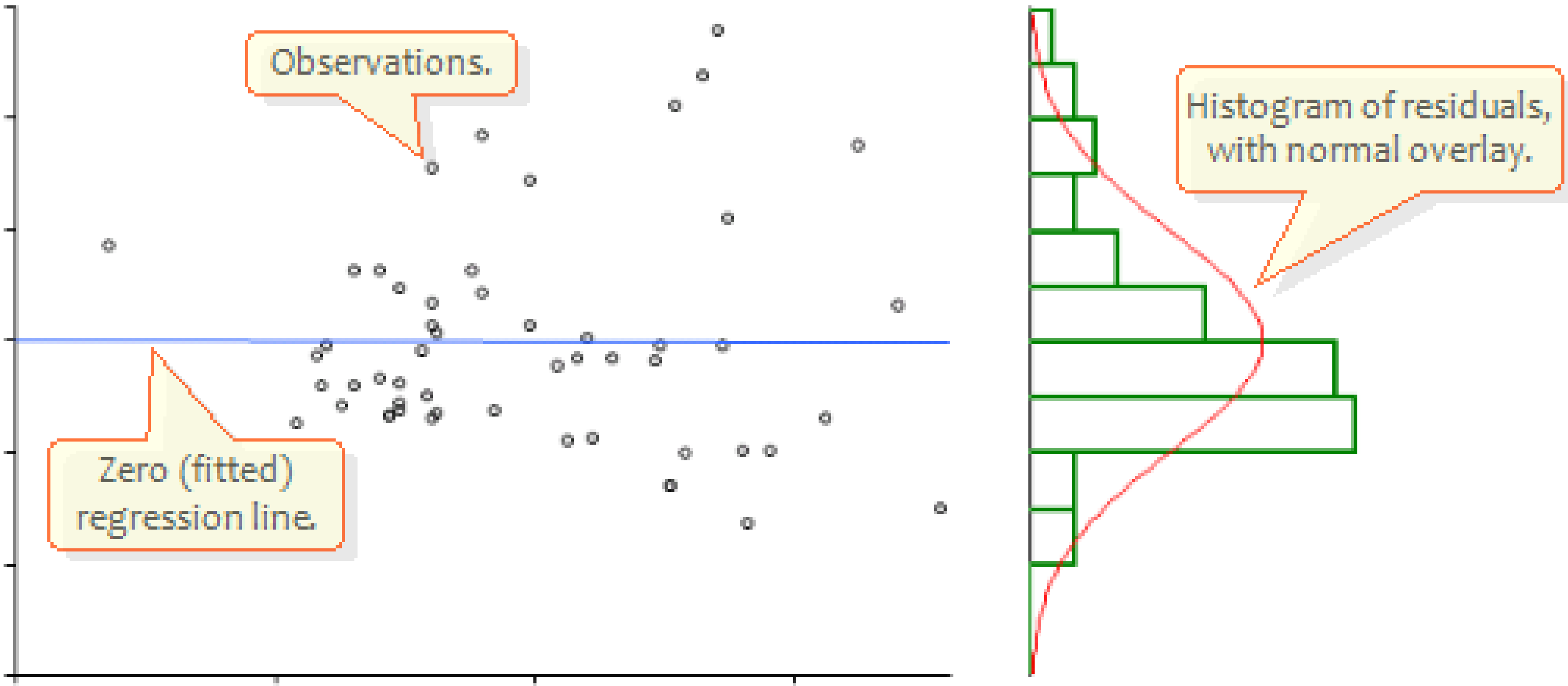


stats.stackexchange.com



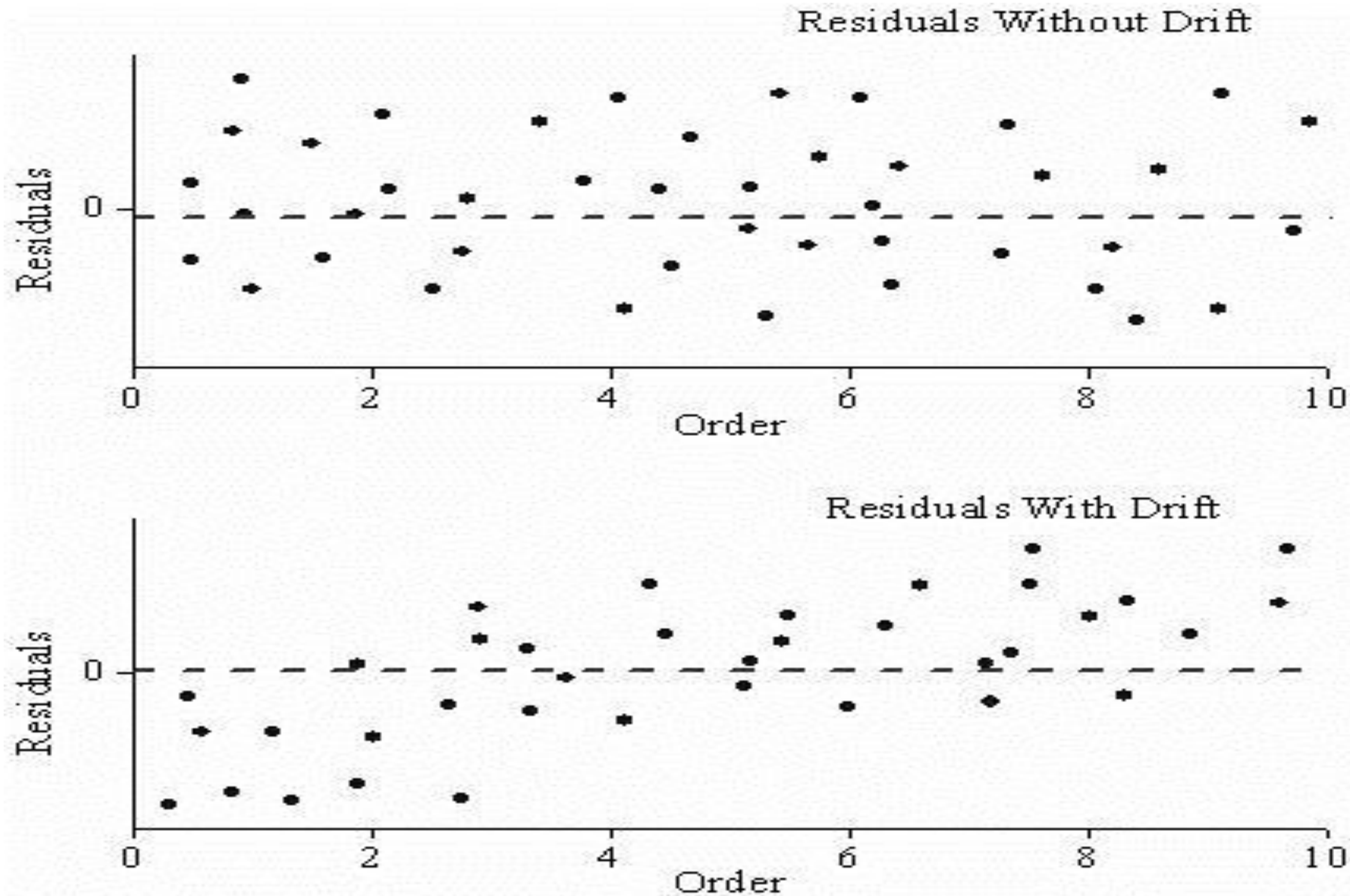
statistics.leard.com

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**; i.e. the **power of variables is 1**.

- Linear relationship is represented by formula:

  outcome (dep. var.) = constant + coefficient*predictor + error

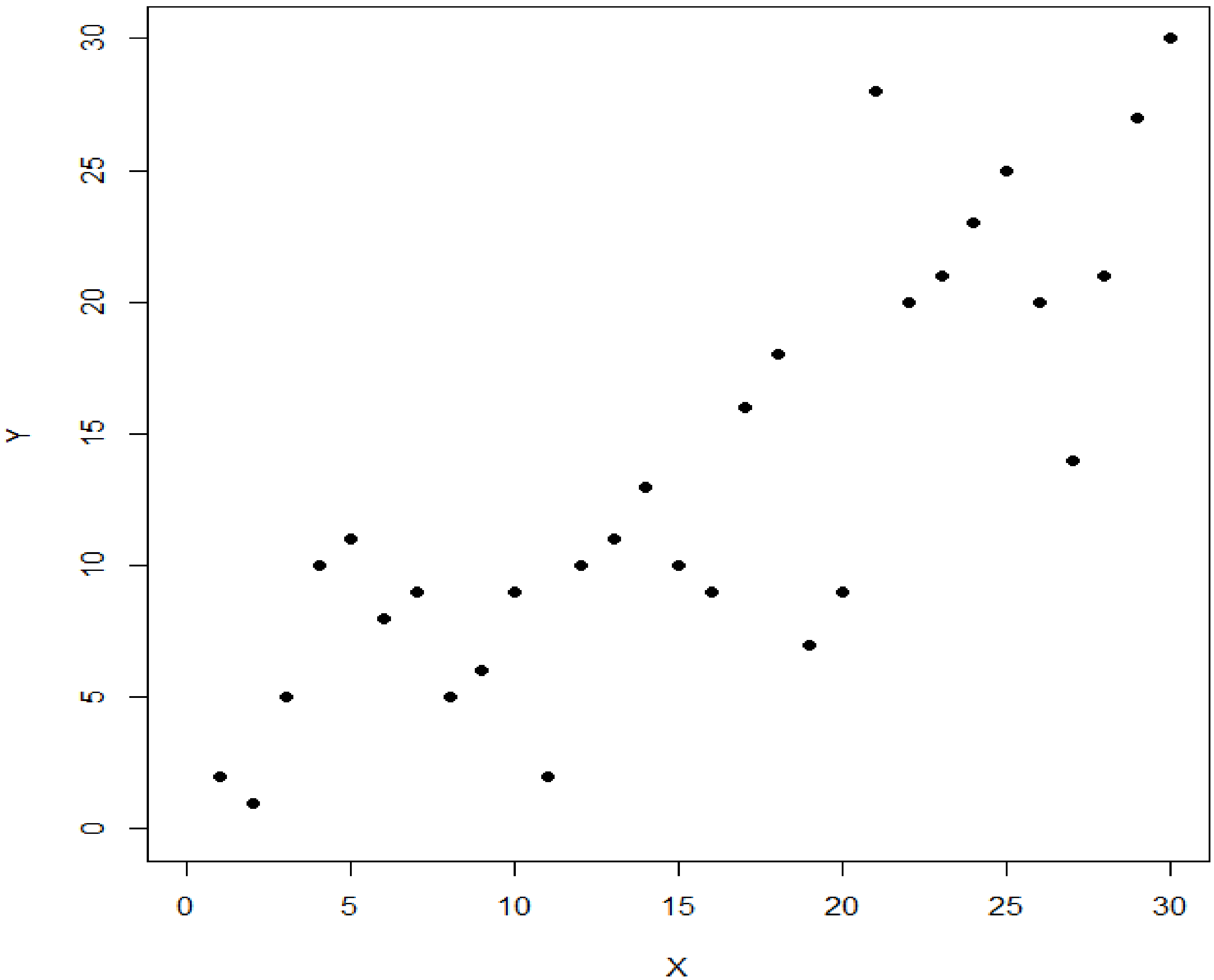
  $Y = \beta 0 + \beta 1X + \varepsilon$; population regression function

  $Y = a + bX + e$; sample regression function

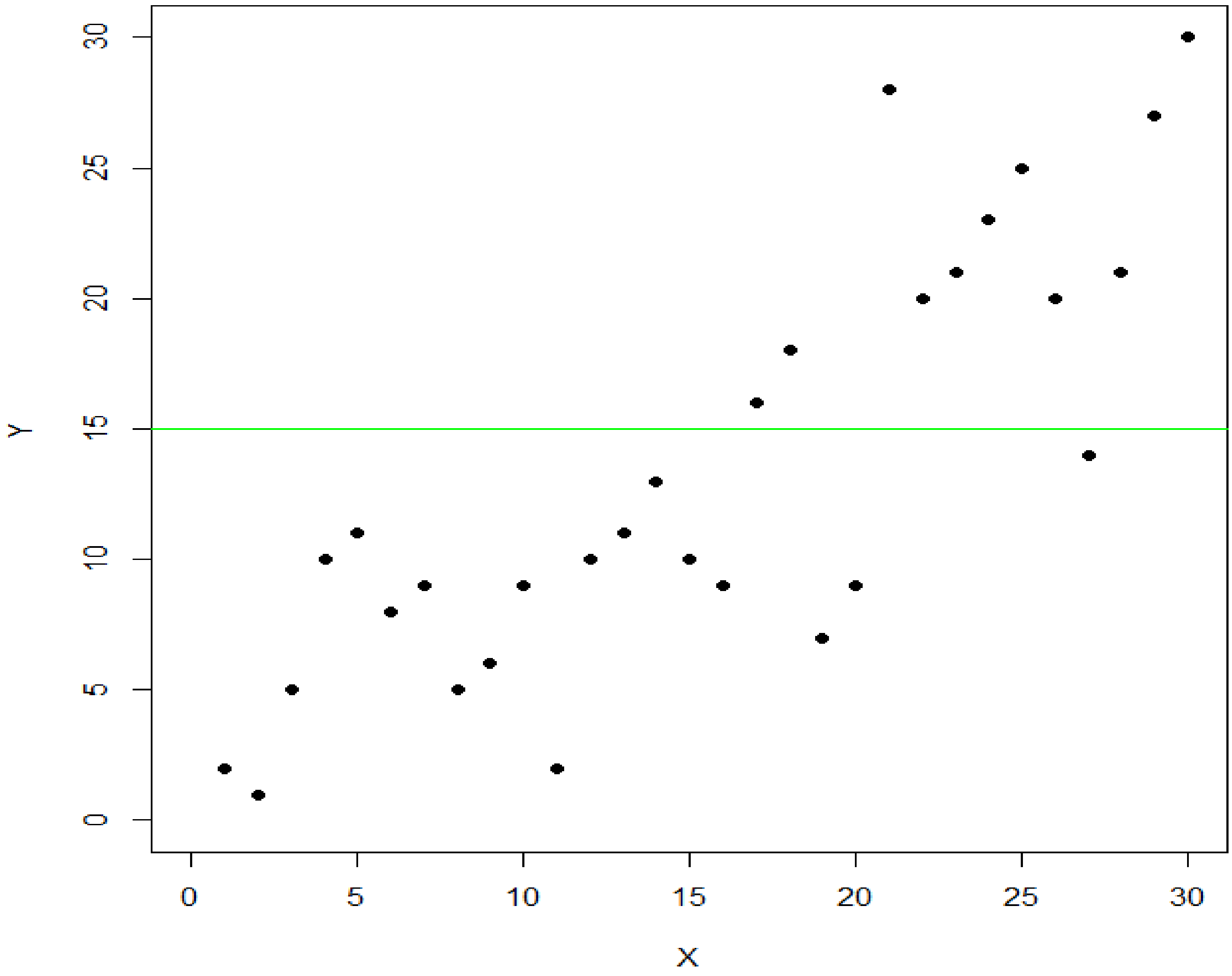
  $Y' = 0.75 + 0.425*X + 2.791$; sample regression line

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

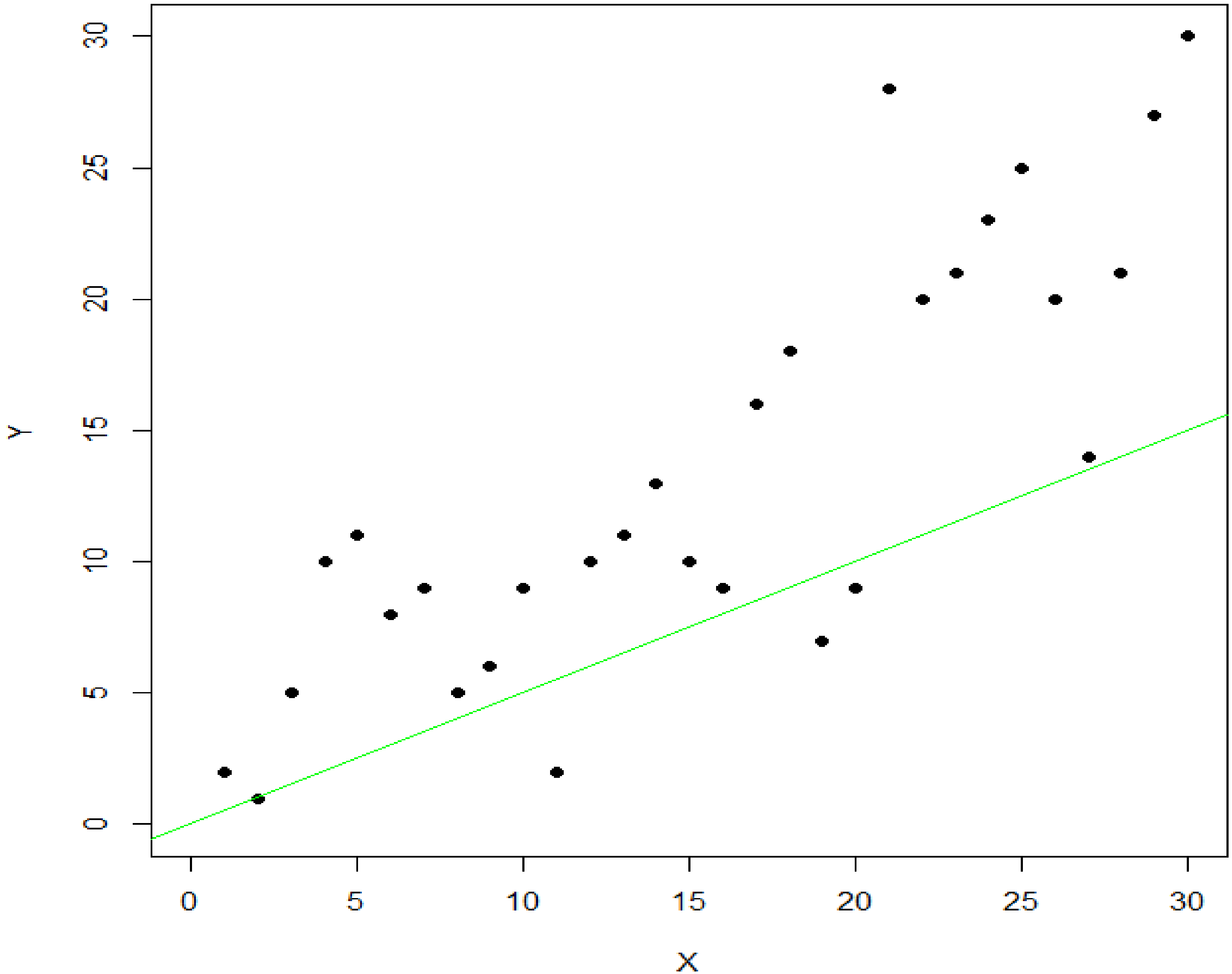
Fitting a straight line



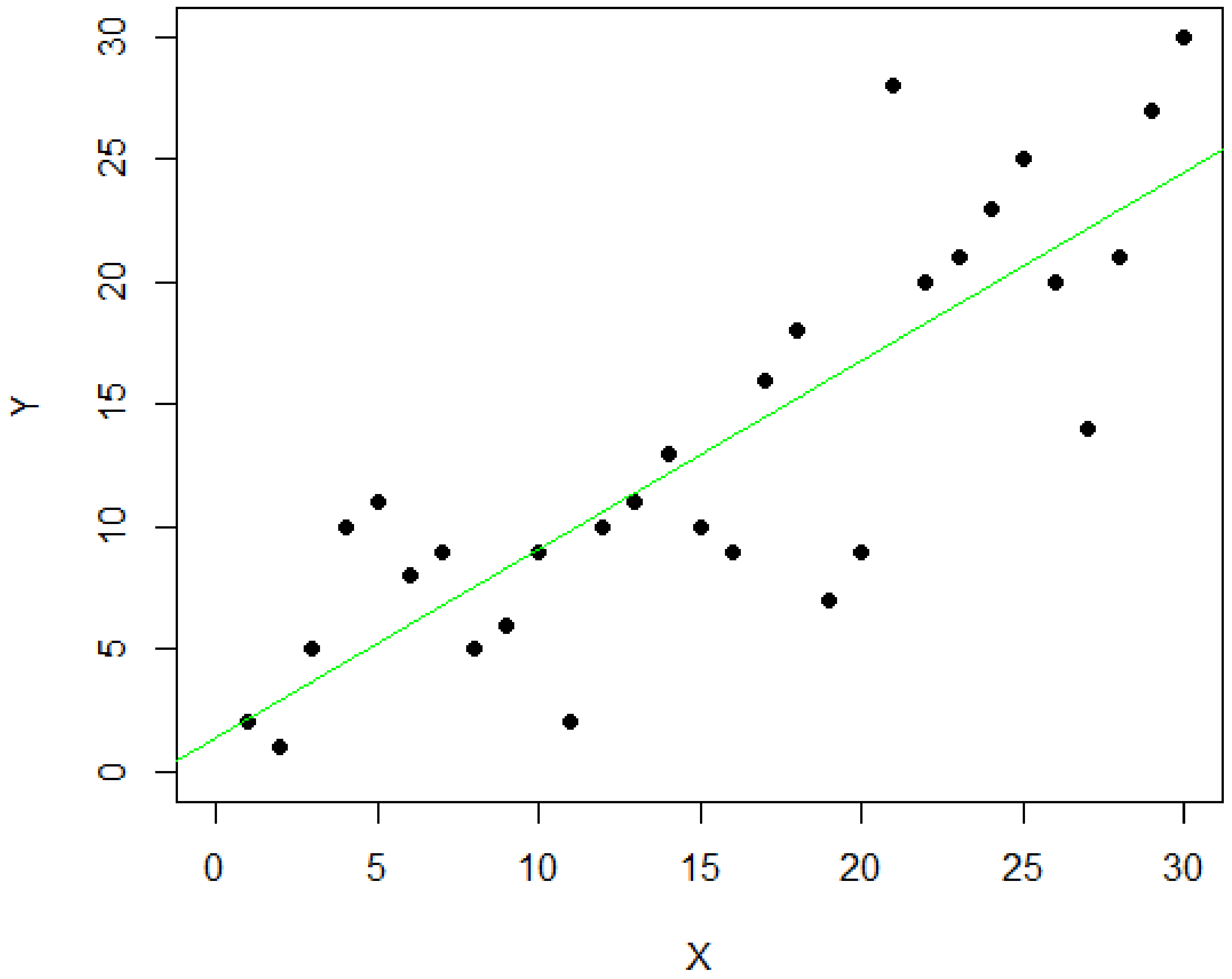
Fitting a straight line



Fitting a straight line



Fitting a straight line



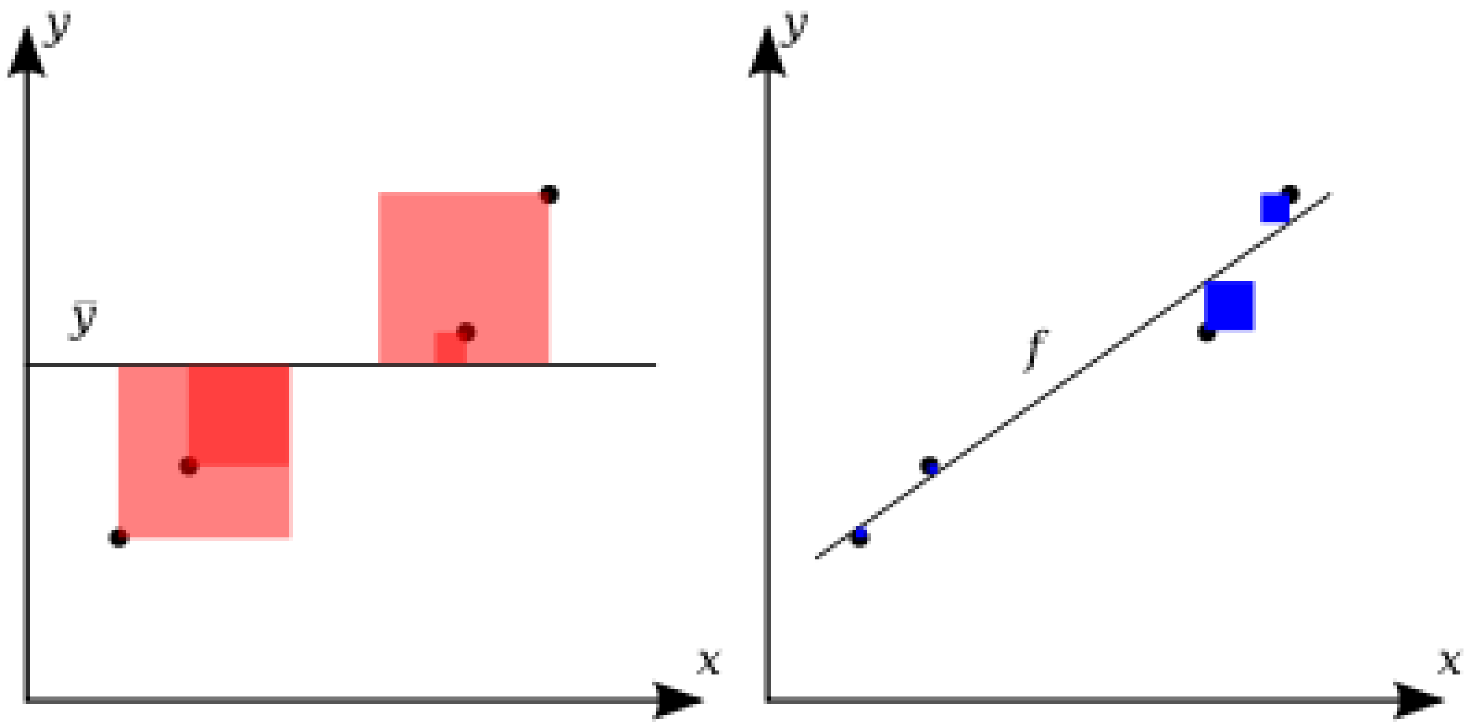
Ordinary least squares

- **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’)**

- ∑ (Y - Y’) = 0 ; ∑ (Y - Y’)^2 >= 0
- **OLS: Y’ = min ∑ (Y - Y’)^2**

Ordinary least squares

- Comparison of mean and OLS estimation.

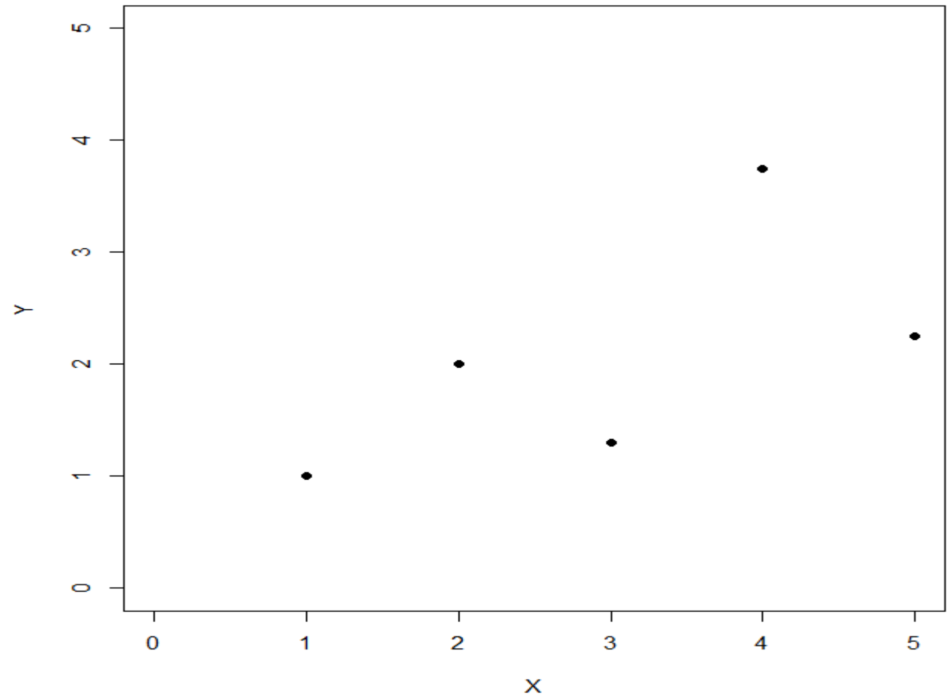


wikimedia commons

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:

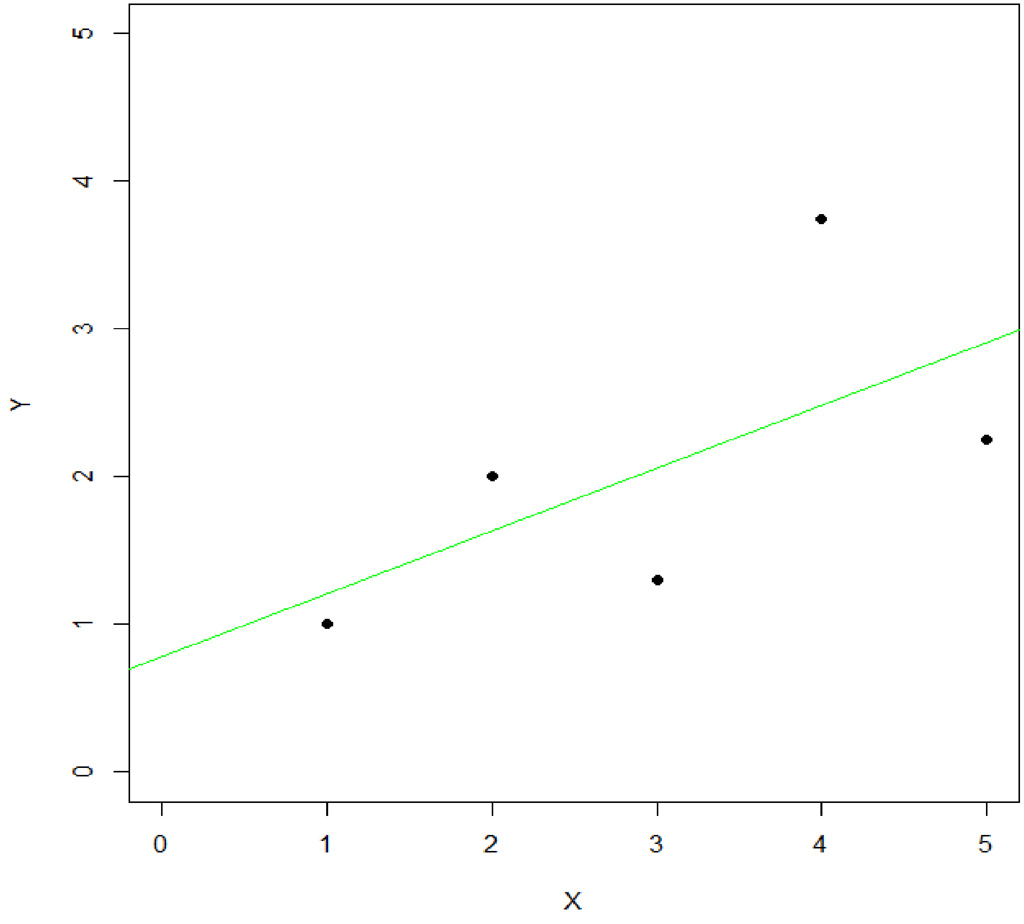
m(X)	m(Y)	s(X)	s(Y)	r(X, Y)
3	2.06	1.581	1.072	0.627

- The **slope (b): r(x, y) * (s(Y)/s(X))** ; same as →
- The **slope (b): ∑(x – m(x))*(y – m(y)) / ∑((x – m(x))^2)**
- The **intercept (a): m(Y) – b*m(X)**

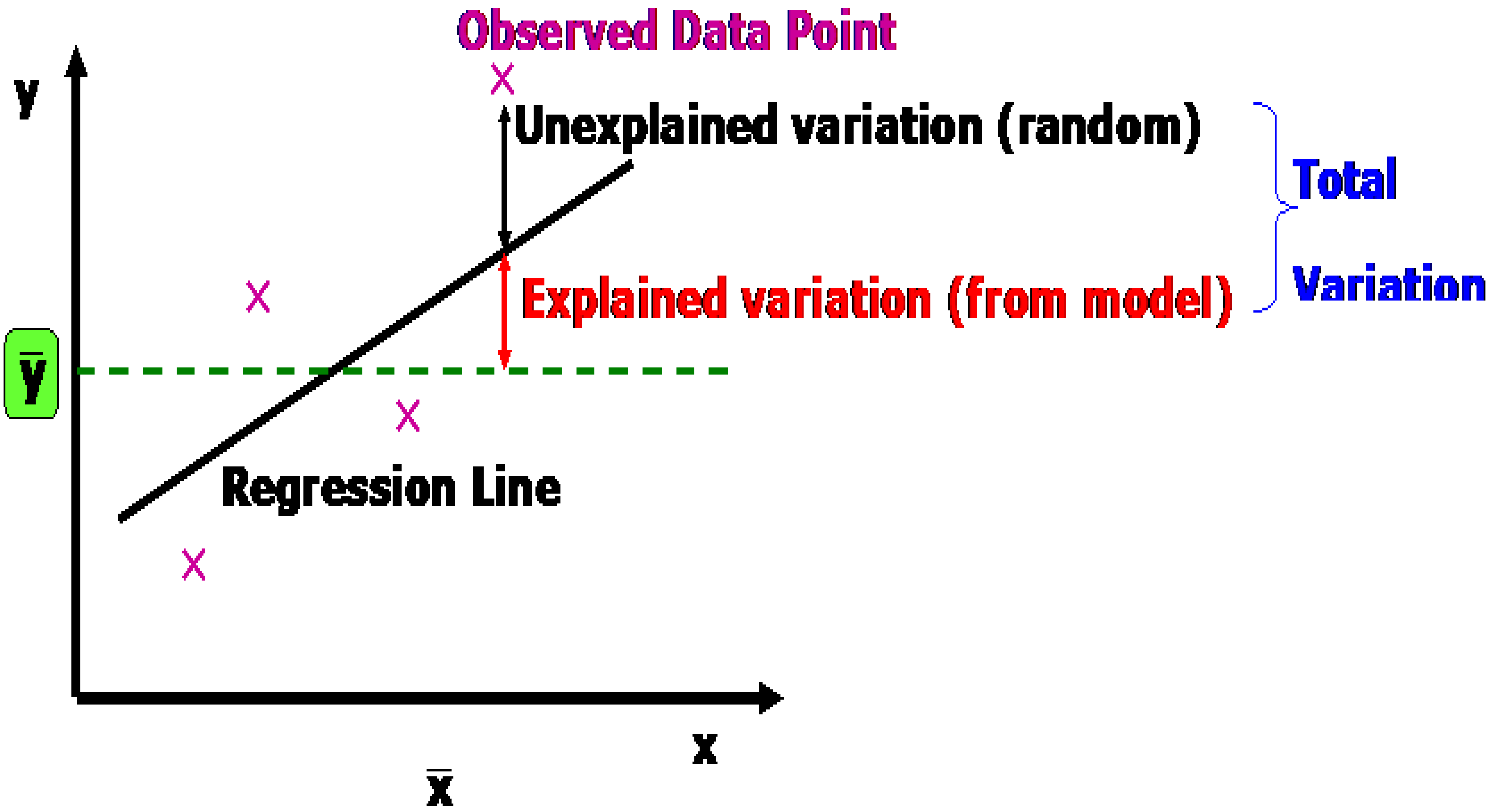
- **b** = 0.627 * 1.072 / 1.581 = **0.425**
- **a** = 2.06 – 0.425 * 3 = **0.75**

Linear regression: example

- Fitting a straight line by using OLS.



Total / unexplained / explained variation



n.a.

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.

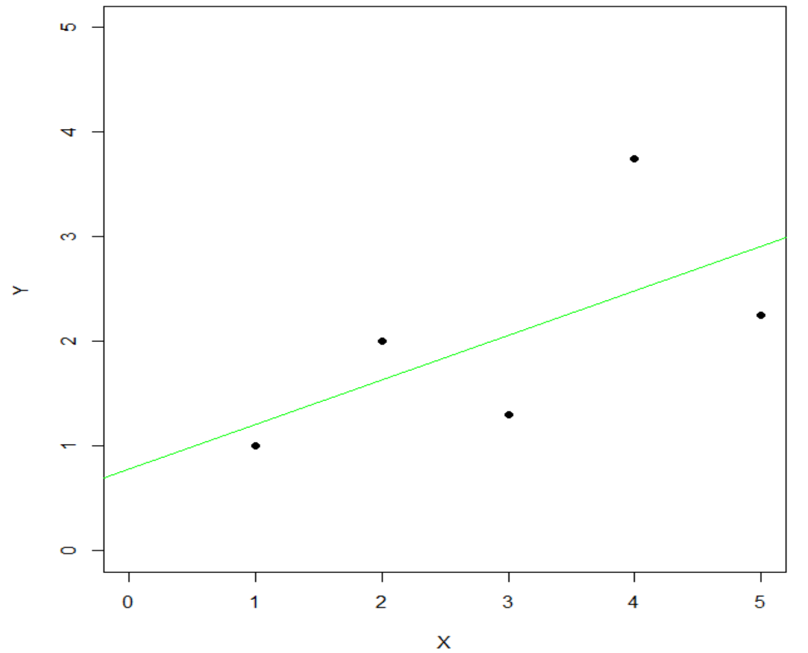
- Model is represented by formula in a form of **linear equation:** Y' = a + bX + e.

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

- R command: *lm()*

Linear regression: interpretation

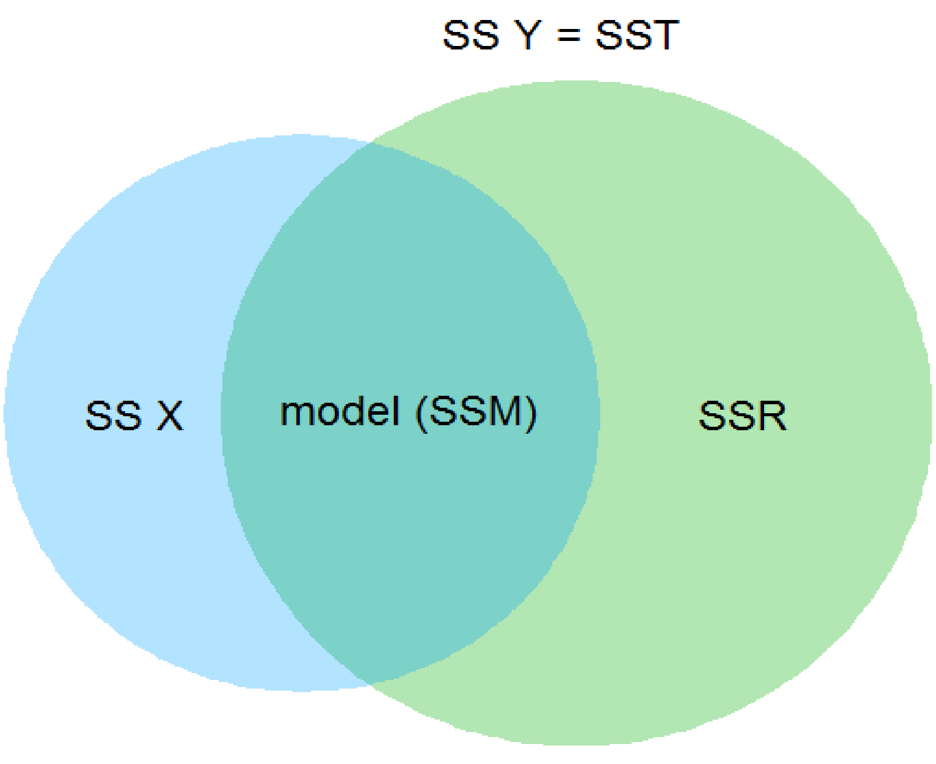
- Model in example: Y = 0.78 + 0.425*X
- **Intercept:** value of Y when value of X = 0.
- **Slope:** change in Y when X increases by 1 unit.
- **Error:** unexplained variance of Y.



- 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

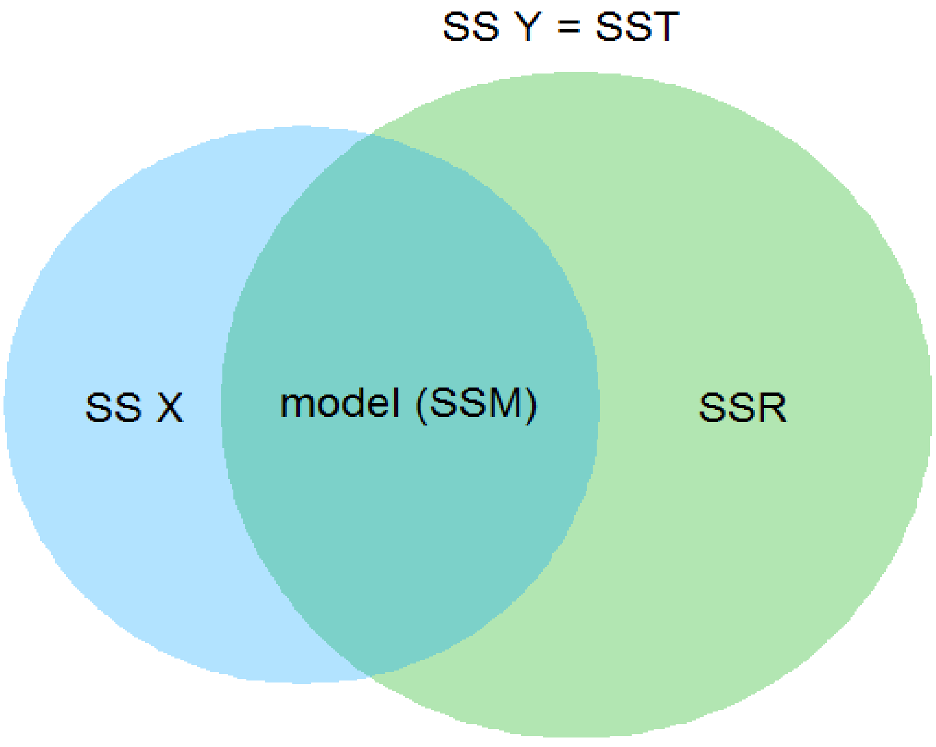
Y’	mean Y	(Y’ – mY)	(Y’ – mY)^2
1.210	2.06	-0.850	0.72
1.653	2.06	-0.425	0.18
2.060	2.06	0	0
2.485	2.06	0.425	0.18
2.910	2.06	0.850	0.72
sum (SSM)			**1.81**

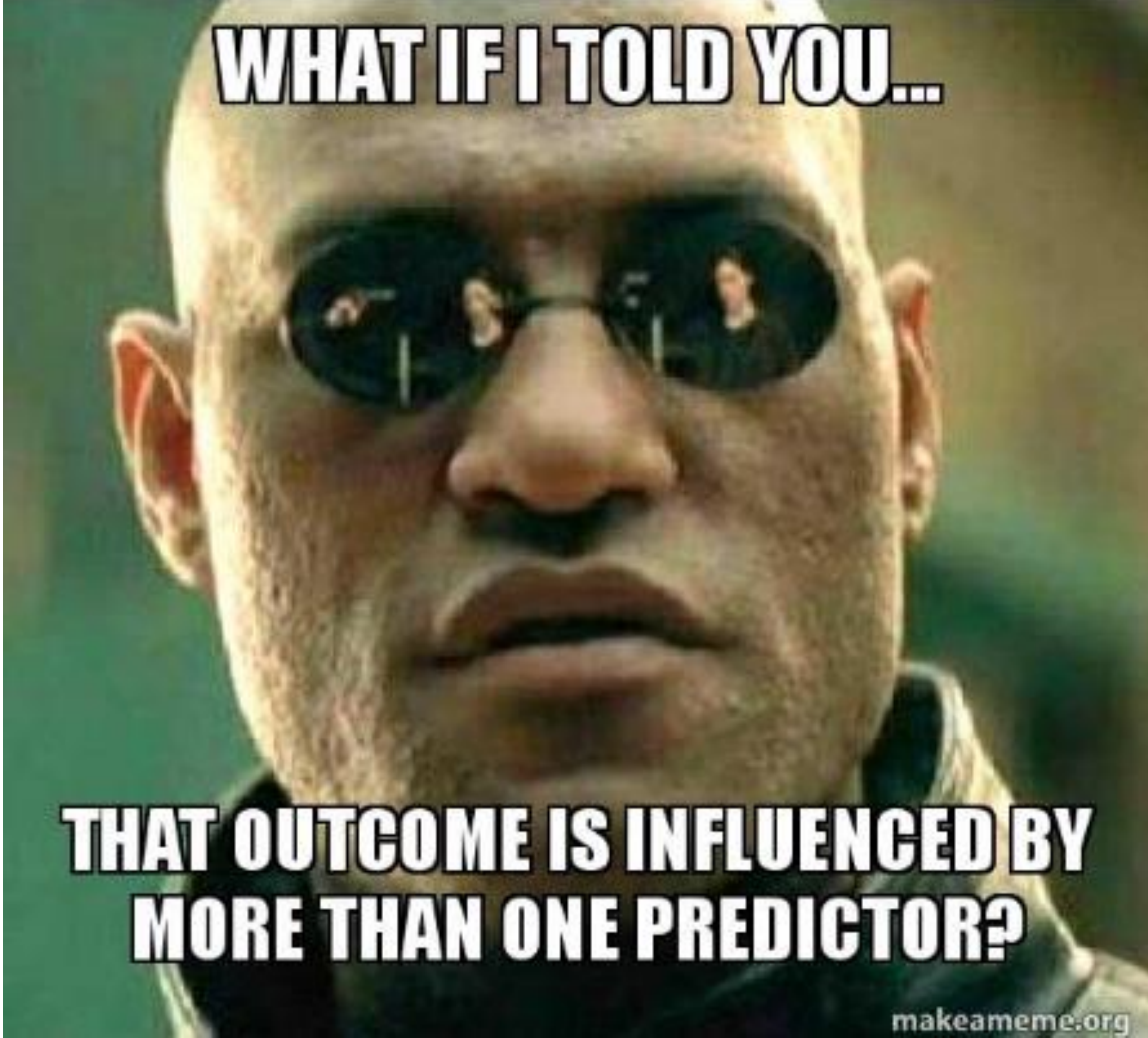
Y	Y’	Y – Y’	(Y – Y’)^2
1	1.210	-0.210	0.044
2	1.653	0.365	0.133
1.3	2.060	-0.760	0.578
3.75	2.485	1.265	1.600
2.25	2.910	-0.660	0.436
sum (SSR)			**2.791**

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

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.)




WHAT IF I TOLD YOU...
THAT OUTCOME IS INFLUENCED BY MORE THAN ONE PREDICTOR?
makeameme.org

Rationale for multiple regression

- **But:** What if the outcome variable is influenced by more than one predictor variable?
- (Always the case...)

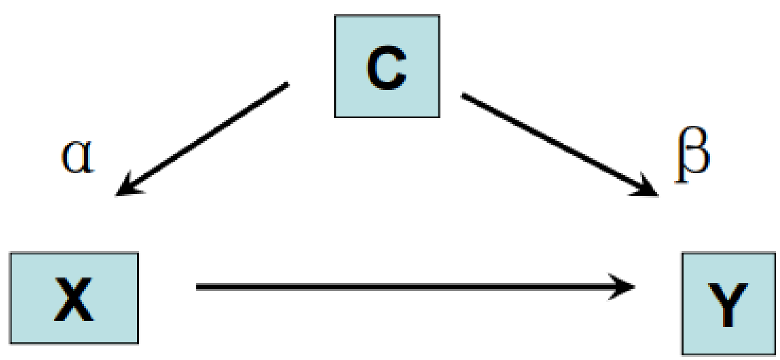
- E.g.: Income can be predicted by completed years of education **and** gender.

→ Idea of **statistical control**

Statistical control: confounding effect

- **Confounding effect:** third variable affects the relationship between predictor(s) and outcome variable.
- A confounder is a variable that correlates both with predictor(s) and outcome variable.

- E.g.: Relationship between income (predictor) and risk of heart attack (outcome) may be confounded by age (confounder).



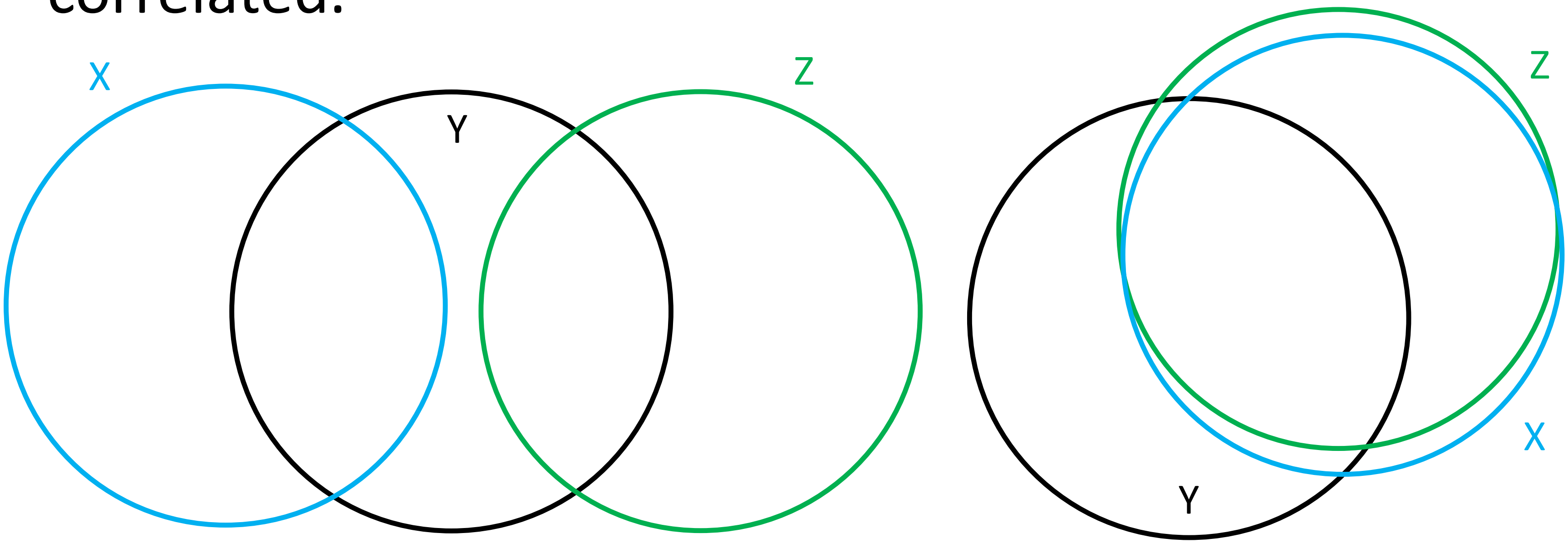
Wu 2010

Multiple regression: assumptions

- Independence of observations (random sampling)
- Normal distribution of Y
- Linear relationship between X and Y
- Normal distribution of residuals
- Homoscedasticity (variance of error is constant)
- Independence of residuals (over time)
- **No high collinearity between predictors**

Collinearity

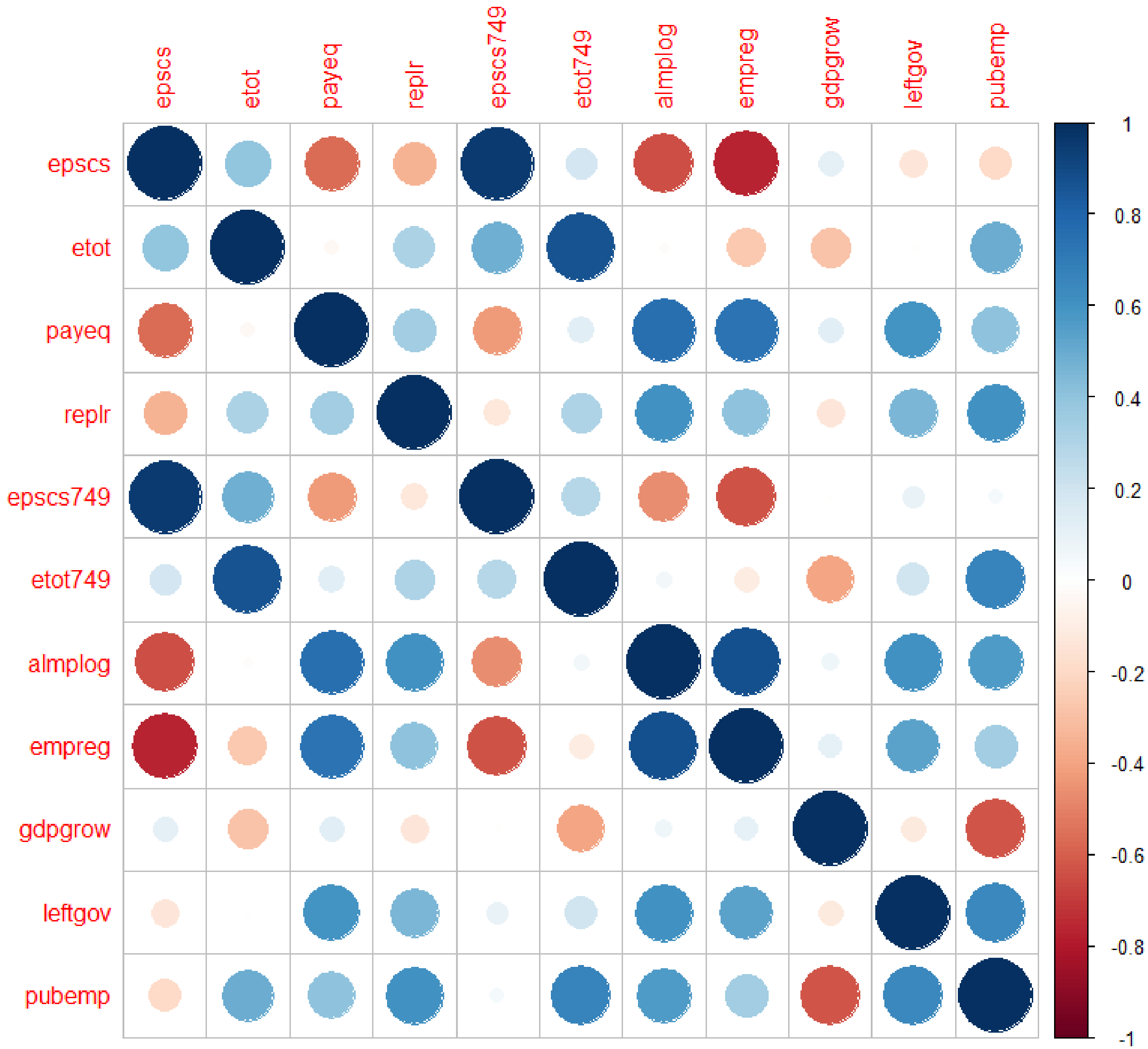
- Collinearity (multicollinearity) = two or more predictors are correlated.



$r_{XZ} = 0$

$r_{XZ} > 0.9$

- Correlation matrix of IVs as a simple diagnostic



Multiple linear relationship

- We add further coefficient*predictor terms into the formula:

outcome (dependent variable) =

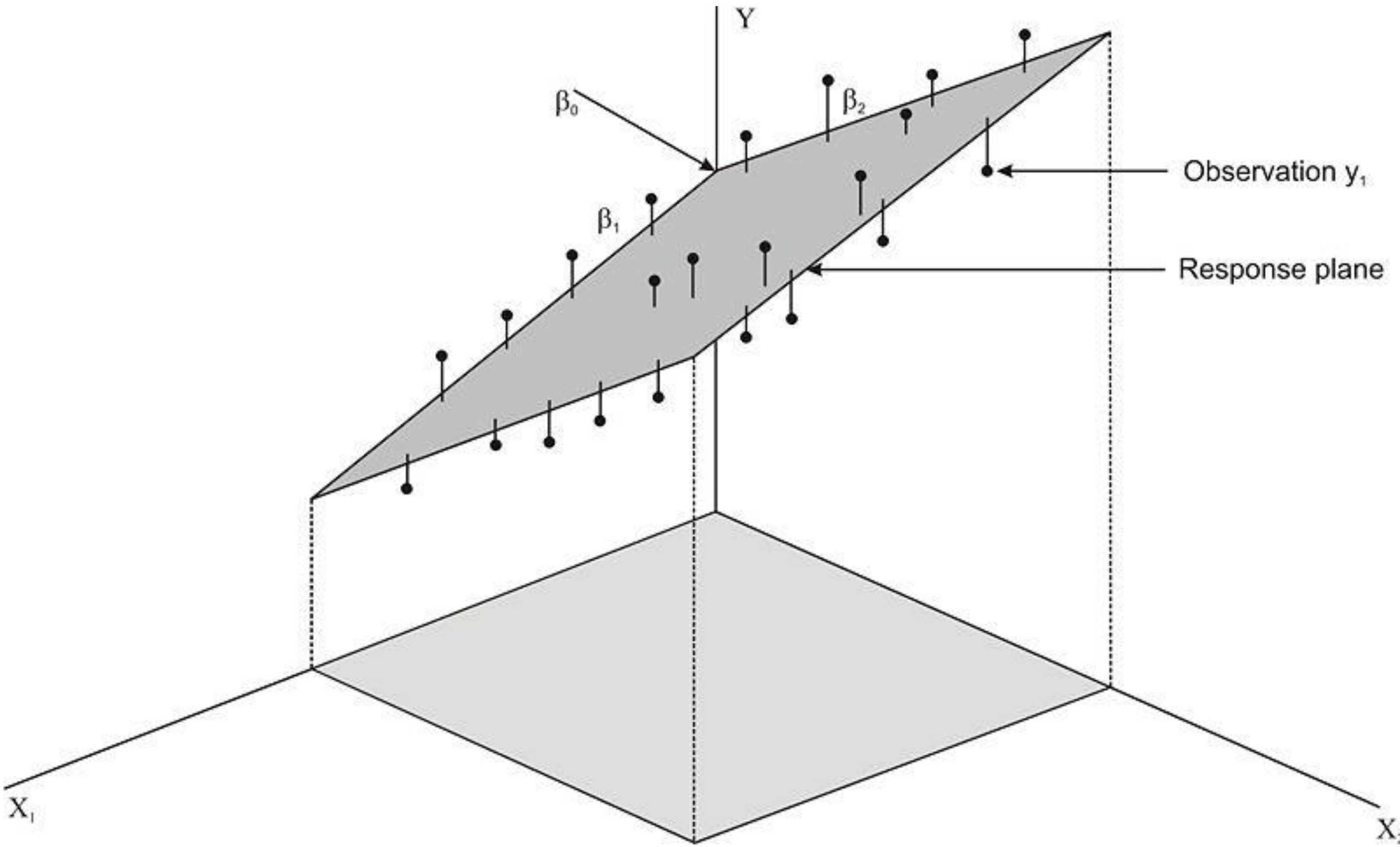
constant + $\boxed{\text{coefficient1*predictor1 + coefficient2*predictor2}}$ + error

Y = β0 + β1X1 + β2X2 + ε ; population regression function

Y = a + b1X1 + b2X2 + e ; sample regression function

Y' = 0.75 + 0.425*X1 + 0.132*X2 + 2.791; sample regression line

Fitting a plane



www.ck12.org

Slope in multiple regression

- Slope gives us information about the **change of the outcome variable** caused by the predictor **while controlling for other predictors** in the model.

- **E.g.:** what is the effect of education (predictor) on income (outcome variable) when we control for age (predictor)?

  income <- 6000 + 500*education + 100*age

- **Interpretation:** for each change in one unit of education (e.g. year), the average unit change of income is 500 unit (i.e. 500 Kč) if age is not changing.

Interpretations

- **If the coefficients are statistically significant:**

- **If X and Z uncorrelated** → reduction to bivariate slopes (X and Z are independent on each other)
- **If X correlates with Y more than Z** → effect of X is stronger (while controlling for Z)
- **If Z correlates with Y more than X** → effect of Z is stronger (while controlling for X)
- **If X and Z (almost) perfectly correlated** → denominator close to 0, resulting values approach infinity (non-interpretable) → problem of **collinearity** (reduction to one variable)

Conclusions

- Linear regression allows us to go beyond associations measurement
  - Prediction
  - Statistical control
- Models are always imprecise!
  - Reduction as well as measurement
- Extensions of regression framework
  - Logistic regression (binary category outcome variable)
  - Multinomial logistic regression (multiple category outcome variable)
  - Ordinal regression (ordinal outcome variable)
  - etc.