Biostatistics
Week VIII

Ege Ülgen, M.D.

25 November 2021

ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

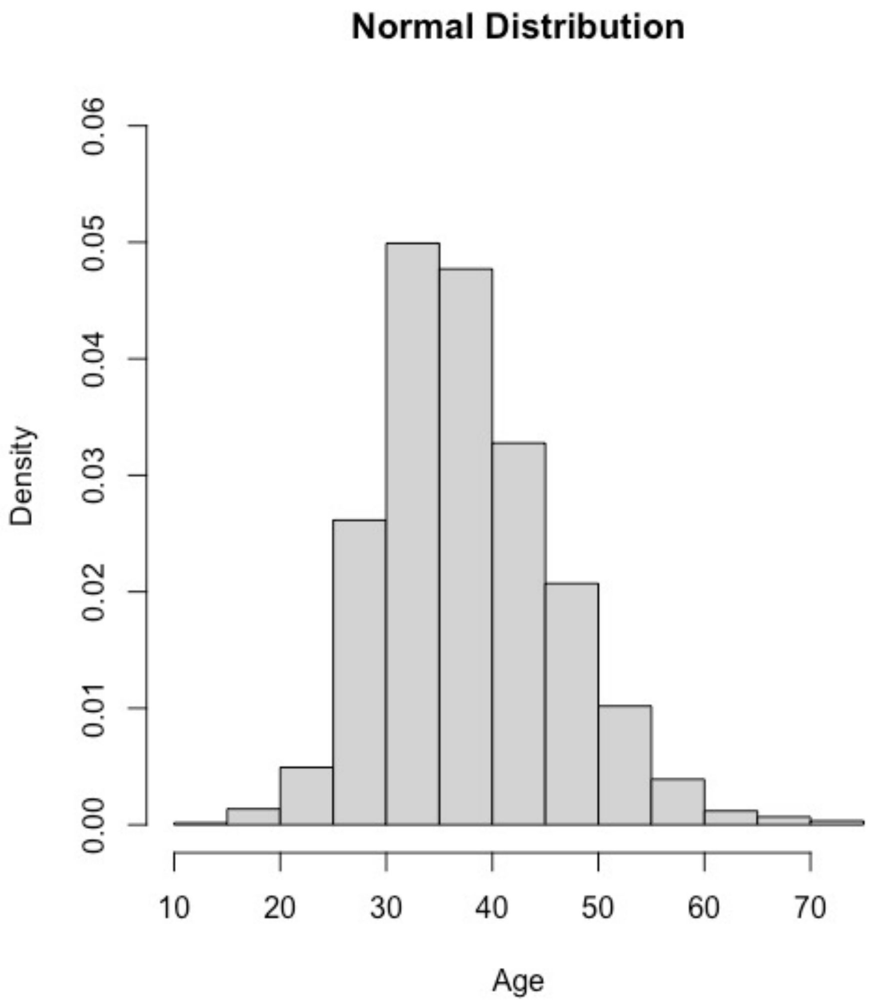
General Assumptions of Parametric Tests

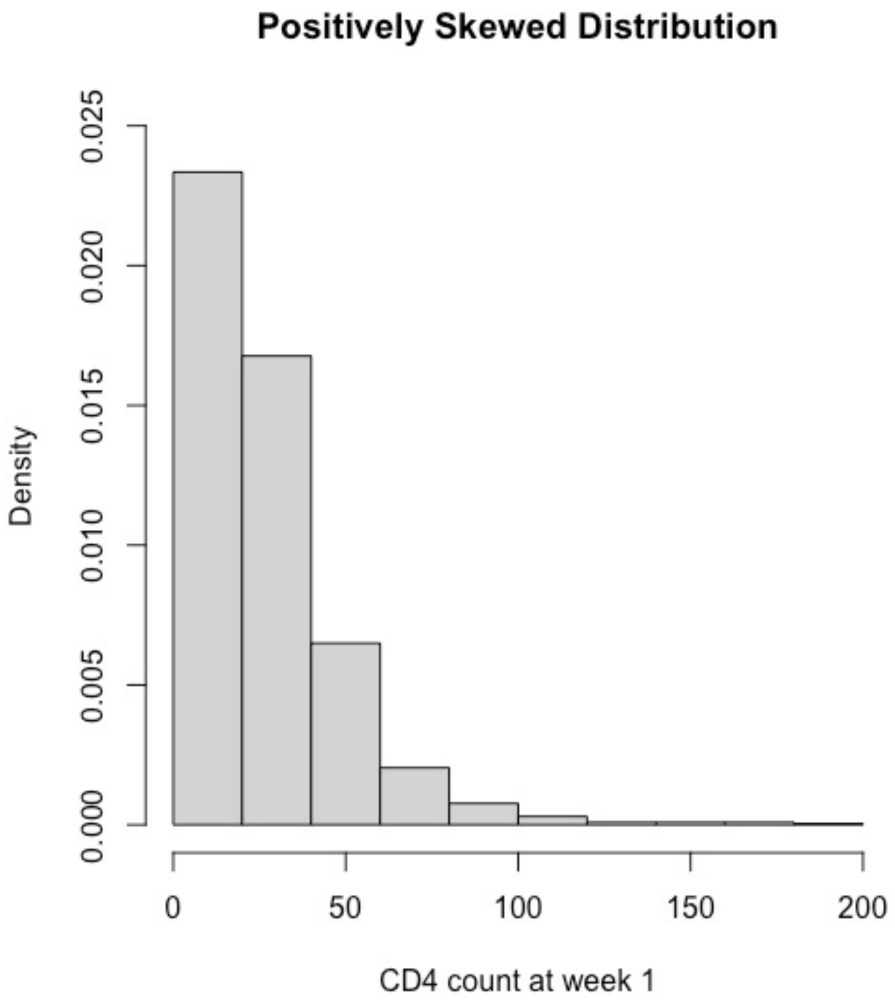
- The population(s) are **normally distributed**
- The selected sample is **representative of general population**
- The data is **continuous**

Assessing Normality

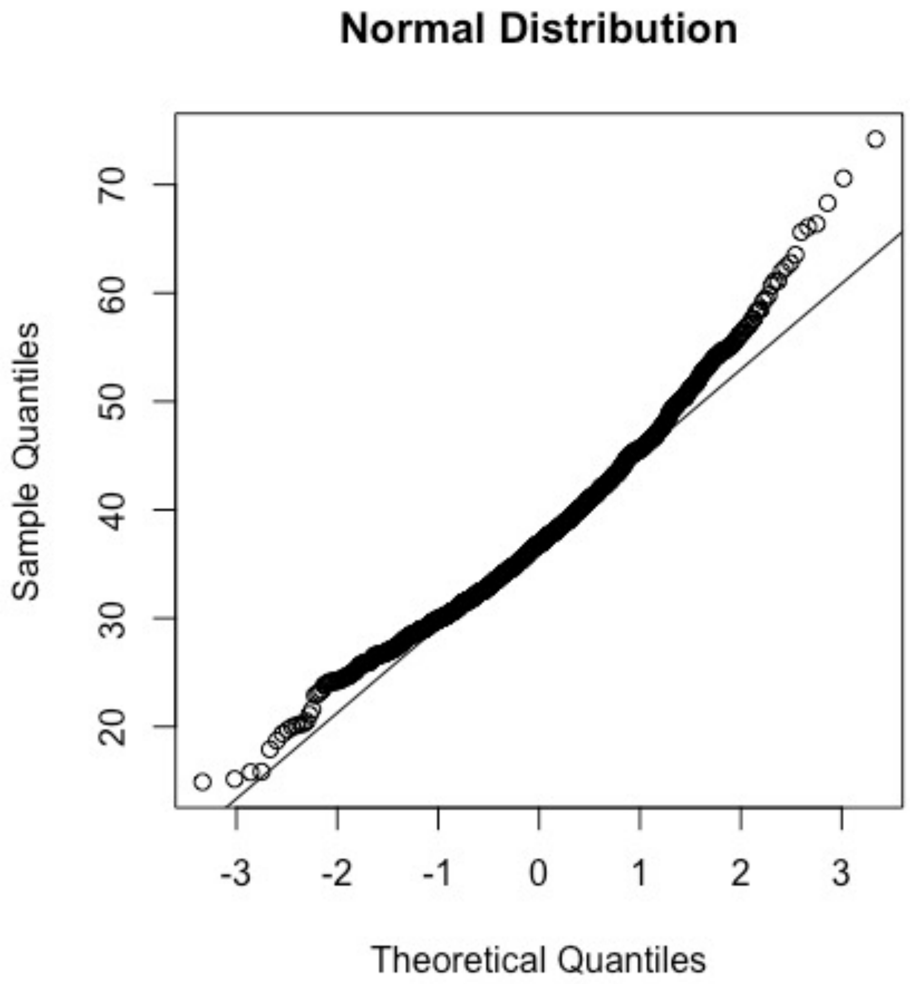
- Inspecting the **histogram** of the variable
- **Quantile-quantile plots**
- **Shapiro-Wilk test**
  - $p < 0.05$ indicates normal distribution
- …

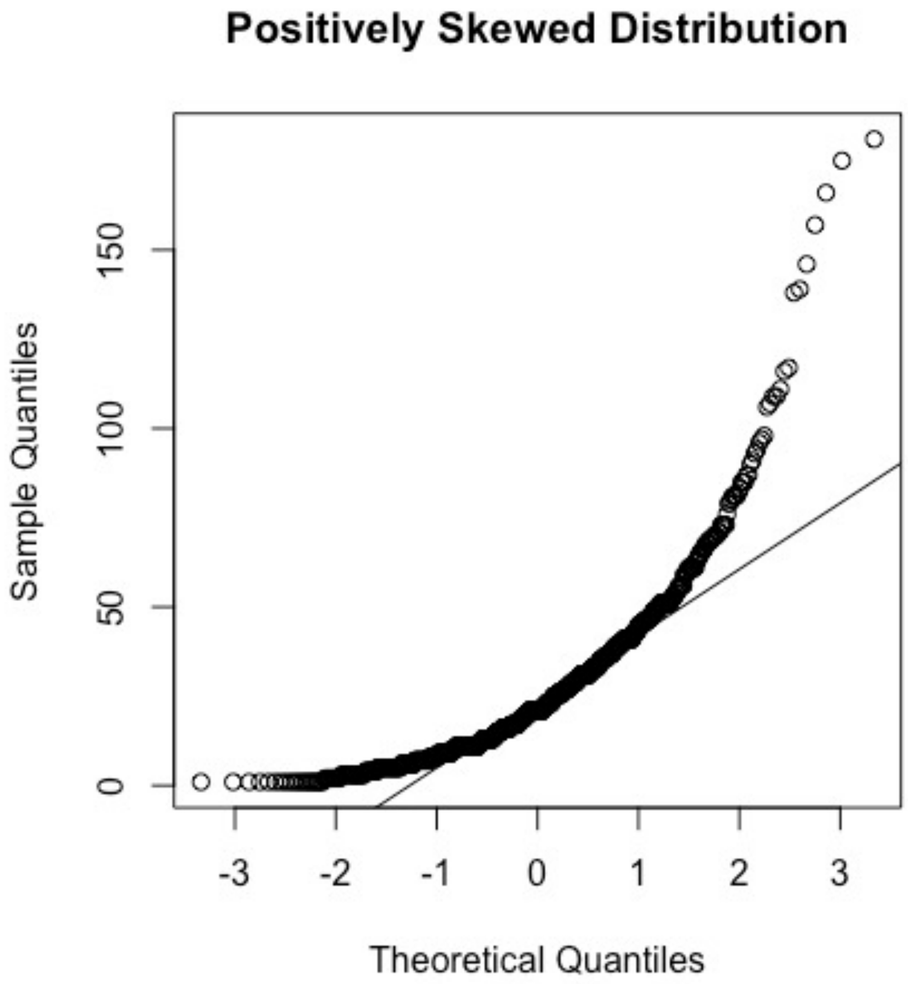
Inspecting Histogram





Quantile-Quantile Plots





Non-parametric Tests

- Used when assumptions of parametric tests are not met
- **Not dependent on the distribution**
- **Less assumptions**
    - e.g., they do not depend on the assumption of normality

- **Less statistical power** compared to parametric tests
    - Higher risk of type II errors (e.g., high probability of accepting there is no difference between the groups where there is a difference)

Non-parametric Tests

- **χ^2 test**
- **Wilcoxon rank-sum test (Mann–Whitney U test)** ~ t-test
- **Kruskal-Wallis test** ~ANOVA
- **Spearman's rank correlation test** ~ Pearson correlation test
- …

Brief Summary

- Normality of a variable can be assessed using
  - Histogram
  - Q-Q plot
  - Shapiro-Wilk test
- Non-parametric tests have **fewer assumptions** but also have **less statistical power** compared to parametric tests

Biostatistics
Week VIII – part II

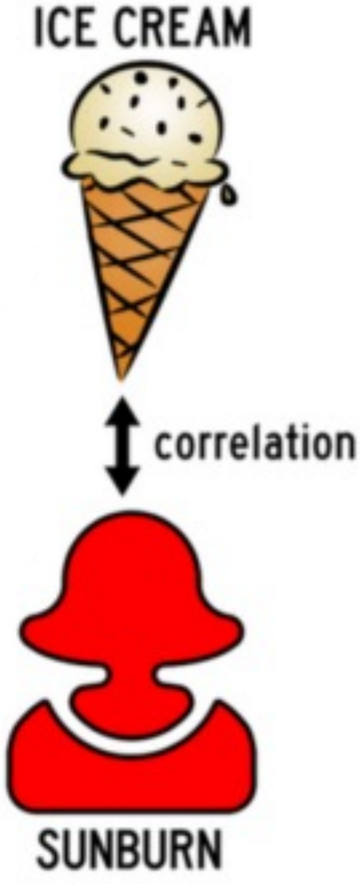
Ege Ülgen, M.D.

25 November 2021

Correlation

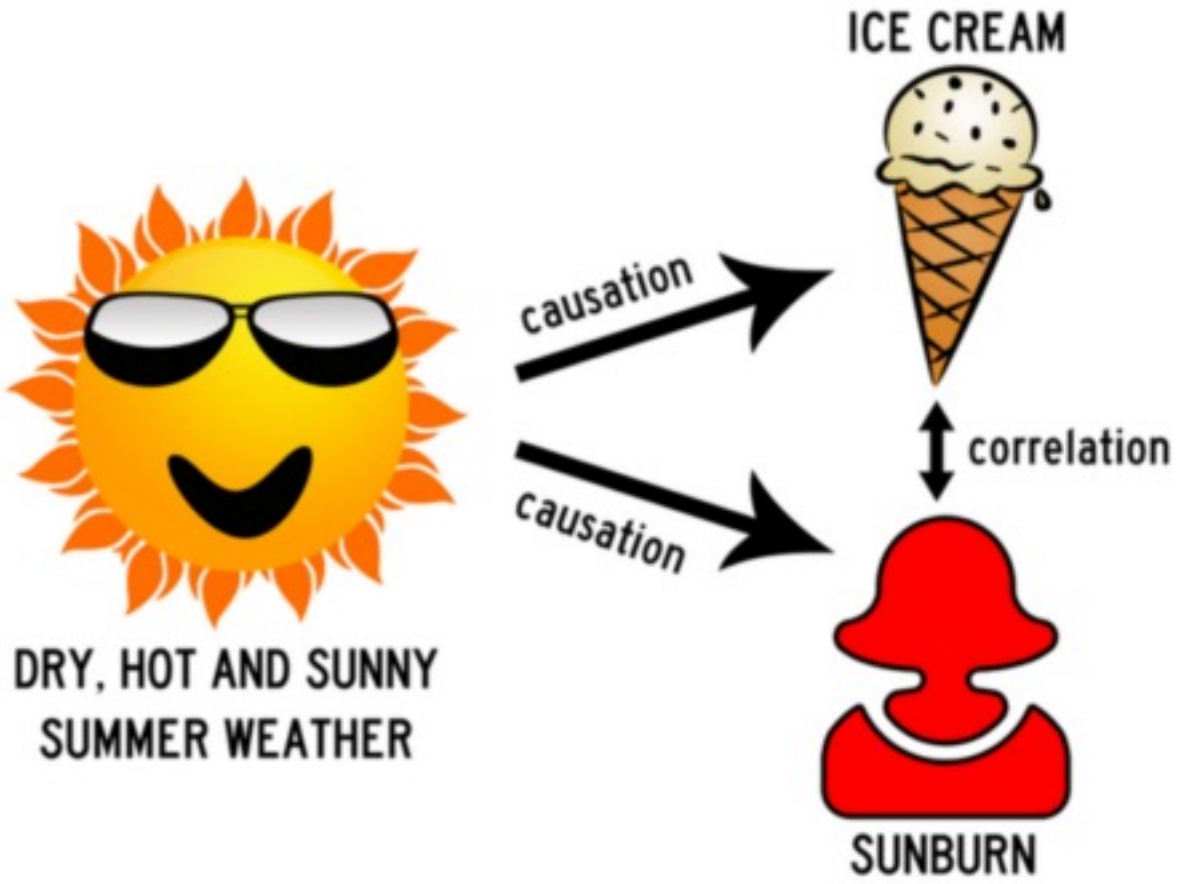
- Correlation is a bivariate analysis that measures **the strength of association** between two variables and **the direction** of the relationships
- In terms of the strength of relationship, the value of the correlation coefficient varies **between +1 and -1**
- **Correlation does not mean causation**

Correlation does not mean causation



Figueroa A. Correlation is not causation [Internet]. Medium. 2019 [cited 2021 Oct 8]. Available from: https://towardsdatascience.com/correlation-is-not-causation-ae05d03c1f53

Correlation does not mean causation



Figueroa A. Correlation is not causation [Internet]. Medium. 2019 [cited 2021 Oct 8]. Available from: https://towardsdatascience.com/correlation-is-not-causation-ae05d03c1f53

Correlation Coefficient

- A statistic that measures the relationship between two variables

- Pearson’s r
  - Measures **linear** relationship
  - Both variables have to be normally distributed
- Spearman’s ρ
  - Measures **monotonic** relationship
  - Based on rank – non-parametric

Pearson Correlation Coefficient

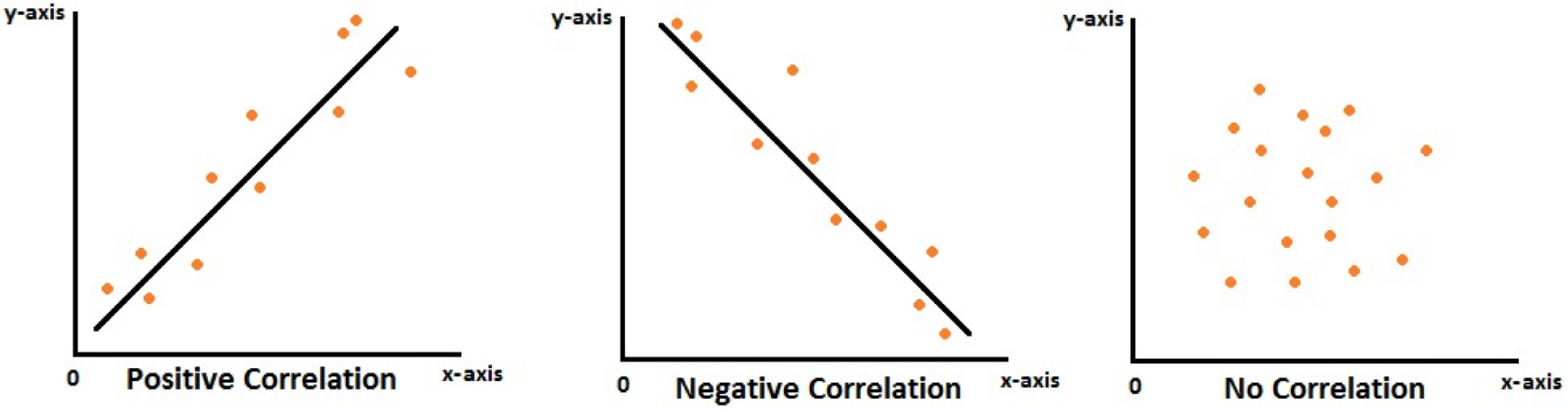
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- A measure of the **linear** correlation between two variables X and Y
- takes values between -1 and 1
- unitless
- $r_{X,Y} = r_{Y,X}$
- $r_{X,Y} = 0$ means **no linear relationship**

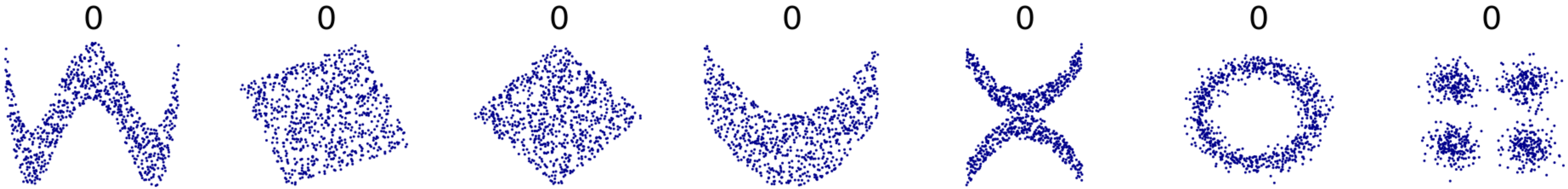
Pearson Correlation Coefficient

Cohen's (1988) conventions to interpret effect size:

- $|r| = 0.10 - 0.29$: Weak
- $|r| = 0.30 - 0.49$: Moderate
- $|r| \geq 0.50$: Strong

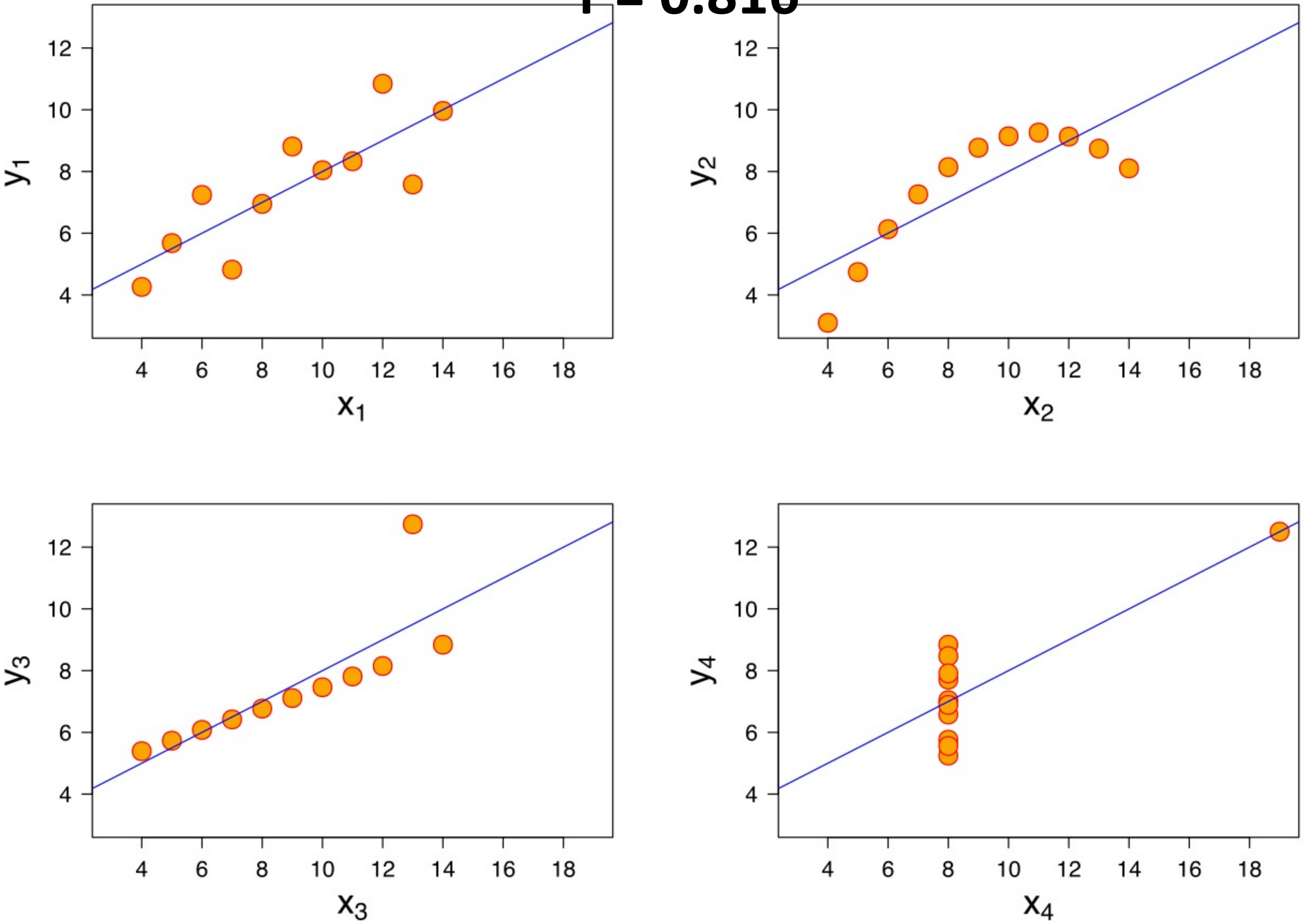


Suresh S. From correlation to causation [Internet]. Medium. 2020 [cited 2021 Oct 8]. Available from: https://towardsdatascience.com/from-correlation-to-causation-49f566eea954



https://en.wikipedia.org/wiki/Correlation_and_dependence

r = 0.816

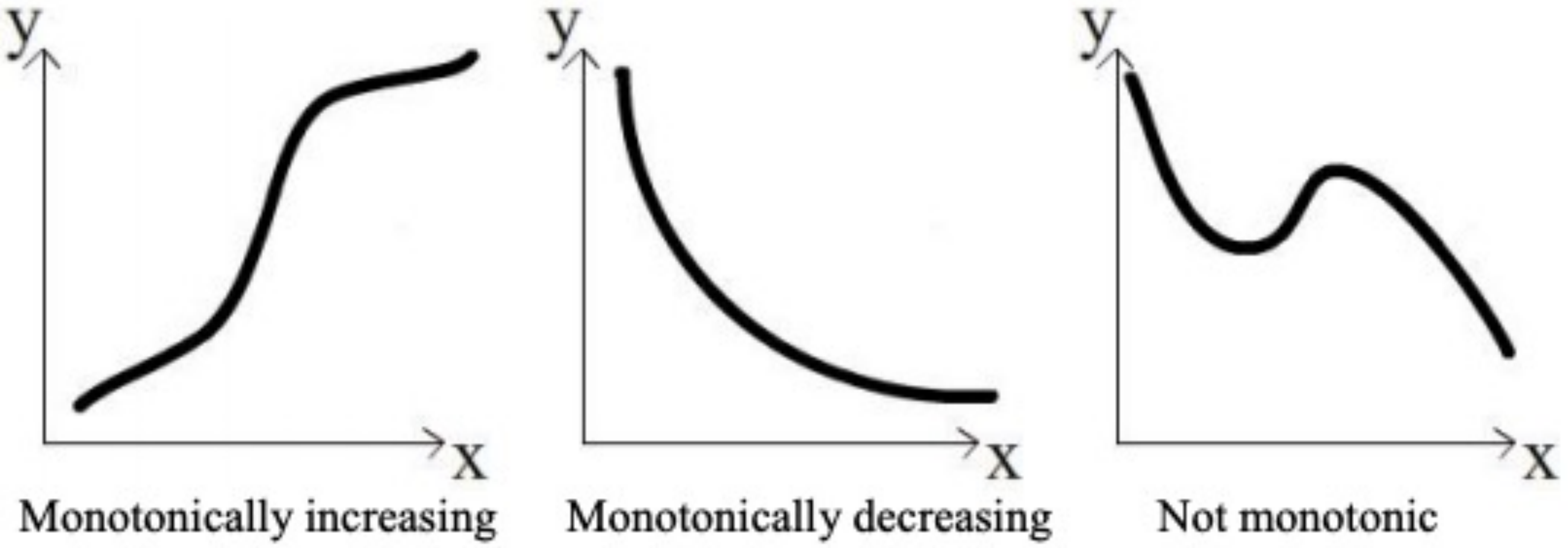


https://en.wikipedia.org/wiki/Correlation_and_dependence

Spearman Rank Correlation

- It assesses how well the relationship between two variables can be described **using a monotonic function**
- It **does not carry any assumptions about the distribution** of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal

Spearman Rank Correlation



Spearman Rank Correlation

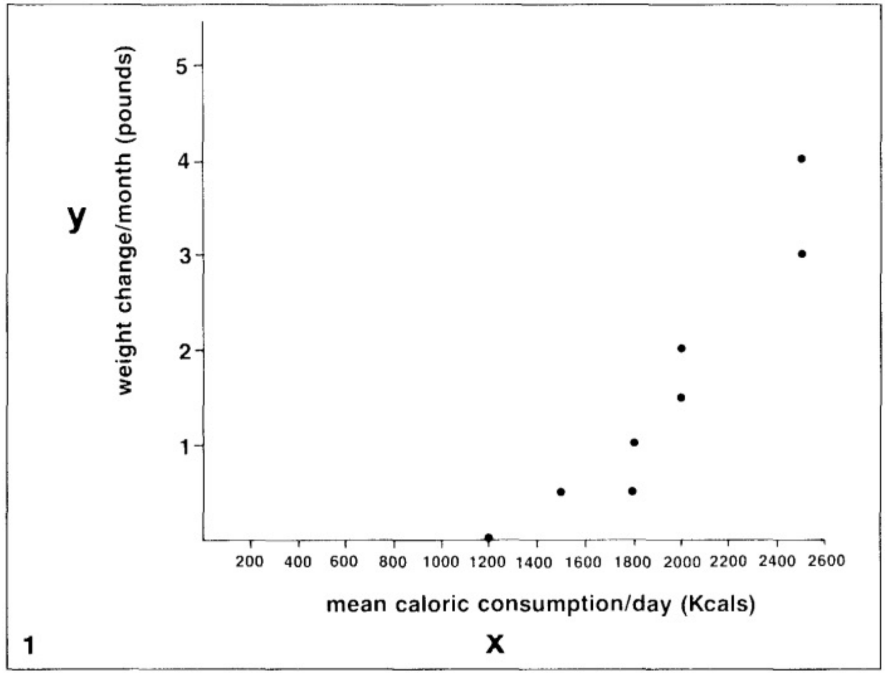
$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

- d_i := the difference between the ranks of corresponding variables (i.e., $d = X_i - Y_i$)
- n := number of observations

TABLE 1. *Sample data: Caloric consumption versus weight change*

Patient	(X) Mean Caloric Consumption/Day	(Y) Weight Change/ Month
1	1,200	0.0
2	1,500	0.5
3	1,800	0.5
4	2,000	1.5
5	2,500	4.0
6	1,800	1.0
7	2,500	3.0
8	2,000	2.0

FIGURE 1. *Scatter diagram for sample data given in Table 1 (caloric consumption vs weight change).*



There is a strong positive relationship between mean caloric consumption/day and weight change/month

$r = 0.94$ or
$\rho = 0.97$

Gaddis ML, Gaddis GM. Introduction to biostatistics: Part 6, Correlation and regression. Ann Emerg Med. 1990 Dec;19(12):1462–8.

Regression Analysis

- Regression analysis is used primarily to **model causality** and **provide prediction**
- Predict the values of a **dependent** (response) variable based on values of at least one **independent** (explanatory) variable
- Explain the **effect** of the independent variables on the dependent variable

Regression Analysis

- Regression can be used to
  - Understand the relationship between variables
  - Predict the value of one variable based on other variables
- Examples:
  - Quantifying the relative impacts of age, gender, and diet on BMI
  - Predicting whether the treatment will be successful or not

Regression Analysis

- The variable to be predicted is called the **dependent variable**
  - Also called the **response variable**
- The value of this variable depends on the value of the **independent variable(s)**
  - Also called the **explanatory** or **predictor variable(s)**

Dependent variable = Independent variable + Independent variable + … + Independent variable

Simple Linear Regression

E.g., quantifying the impact of age on BMI

- Linear regression is a method for estimating the **linear relationship** between the dependent and independent variables

- Relationship between variables is described by a linear function

Intercept slope residual

$$\underbrace{Y_i}_{\text{Dependent variable}} = \overbrace{\beta_0}^{\text{Intercept}} + \overbrace{\beta_1}^{\text{slope}} \underbrace{X_i}_{\text{Independent variable}} + \overbrace{\varepsilon_i}^{\text{residual}}$$

Dependent variable

Independent variable

- The coefficients are estimated by minimizing the sum of the squared errors/residuals (Least squares)

James G, Witten D, Hastie T, Tibshirani R. Linear regression. In: James G, Witten D, Hastie T, Tibshirani R, editors. An Introduction to Statistical Learning: with Applications in R [Internet]. New York, NY: Springer; 2013 [cited 2021 May 7]. p. 59–126. (Springer Texts in Statistics). Available from: https://doi.org/10.1007/978-1-4614-7138-7_3

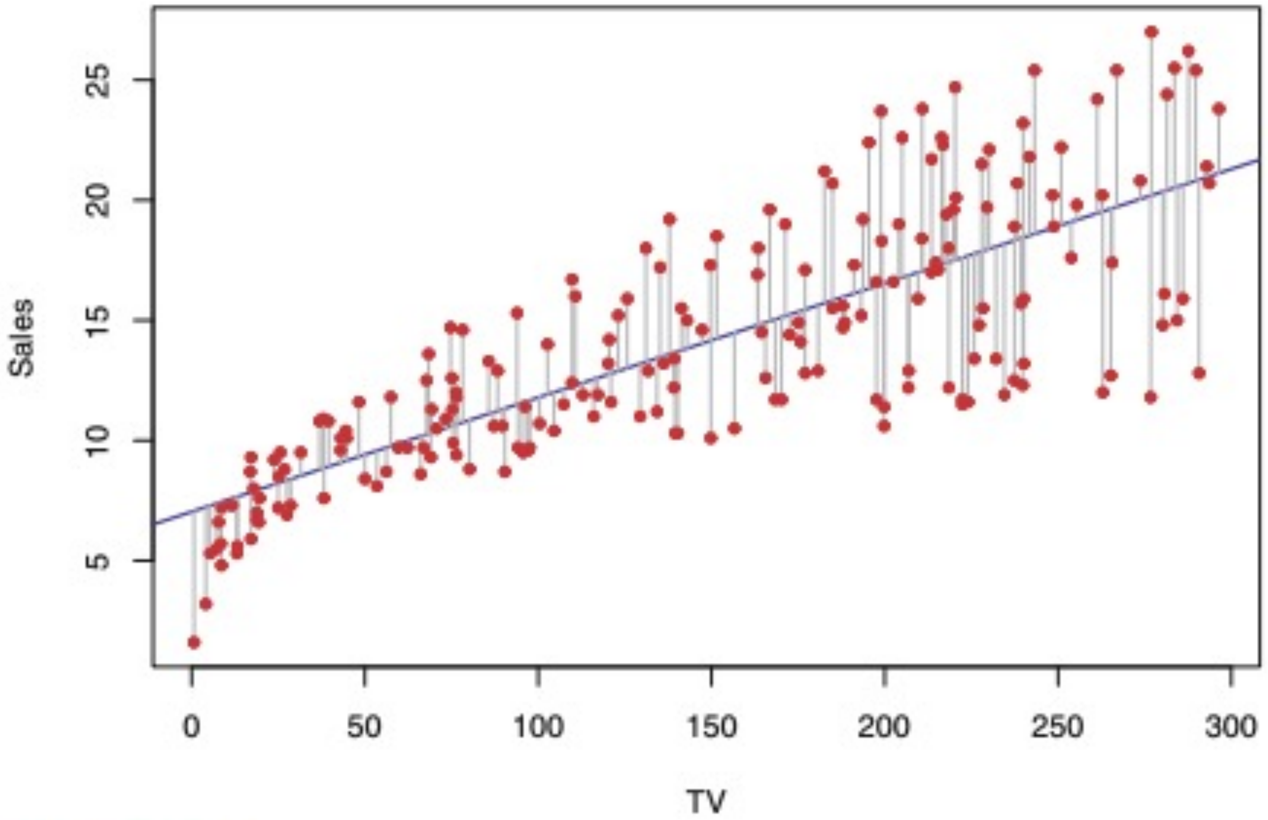
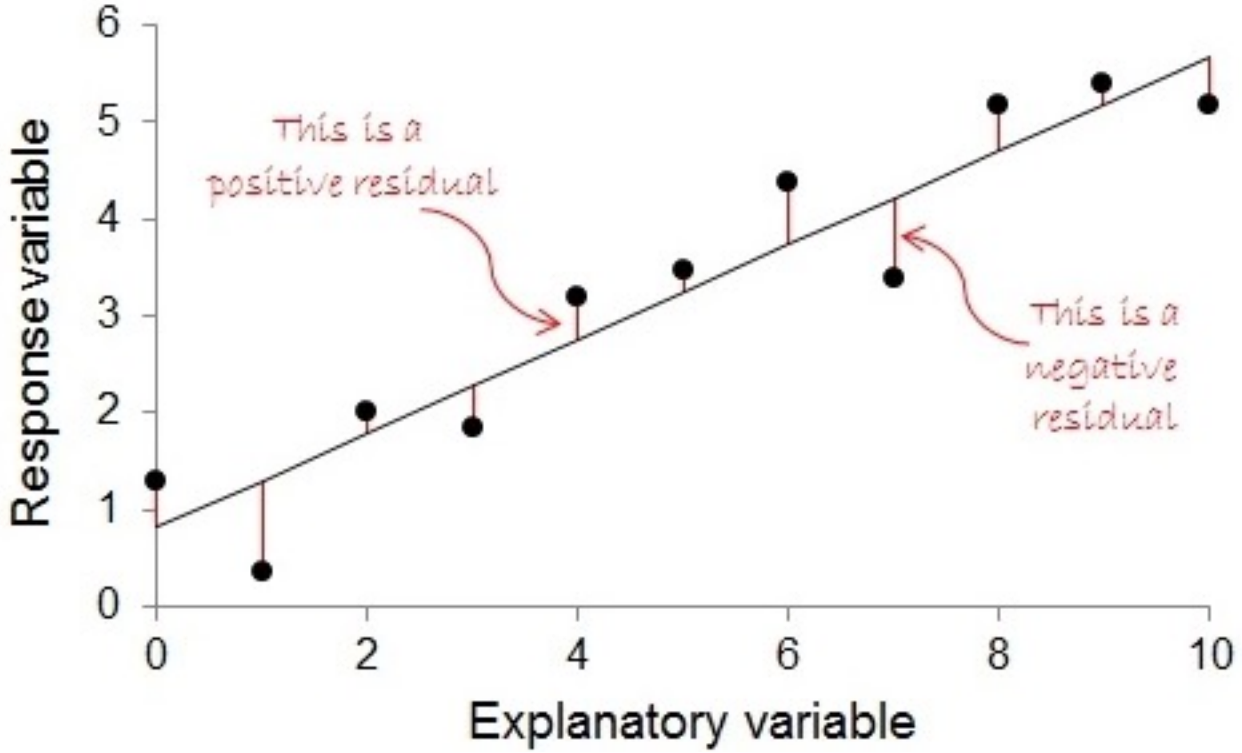


FIGURE 3.1. *For the* `Advertising` *data, the least squares fit for the regression of* `sales` *onto* `TV` *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.*

Error/residual = Actual value – Predicted value



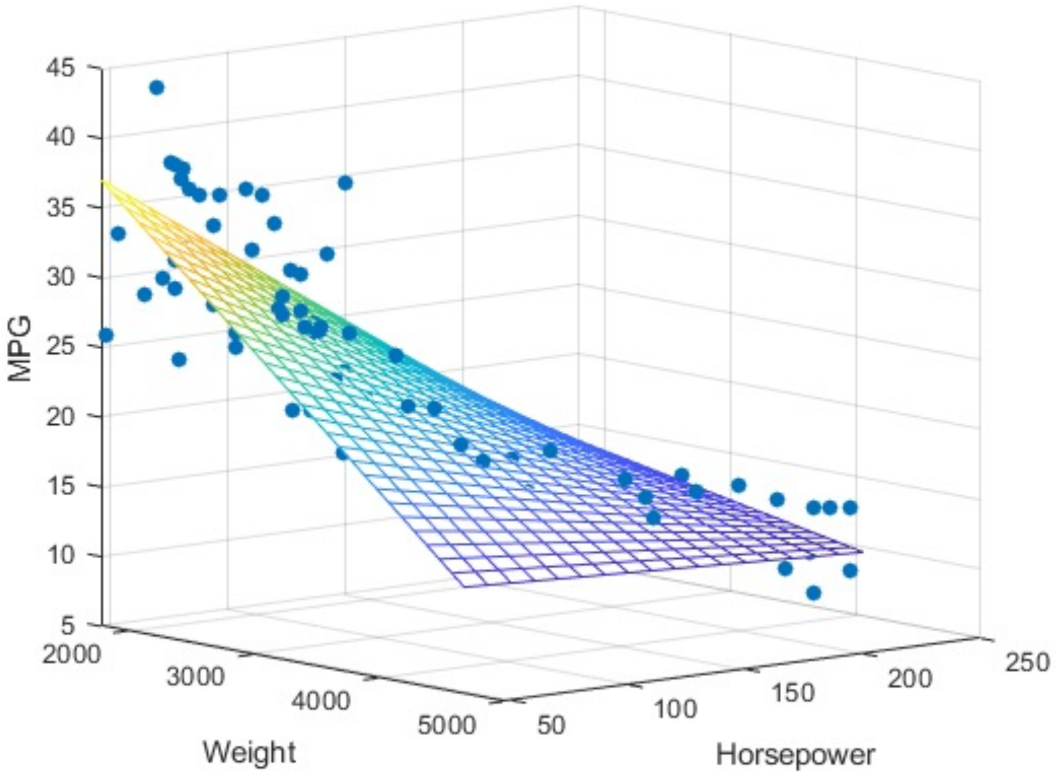
Kumar A. Linear regression explained with python examples [Internet]. Data Analytics. 2020 [cited 2021 Oct 9]. Available from: https://vitalflux.com/linear-regression-explained-python-sklearn-examples/

Multiple Linear Regression

E.g., quantifying the relative impacts of age, gender, and diet on BMI

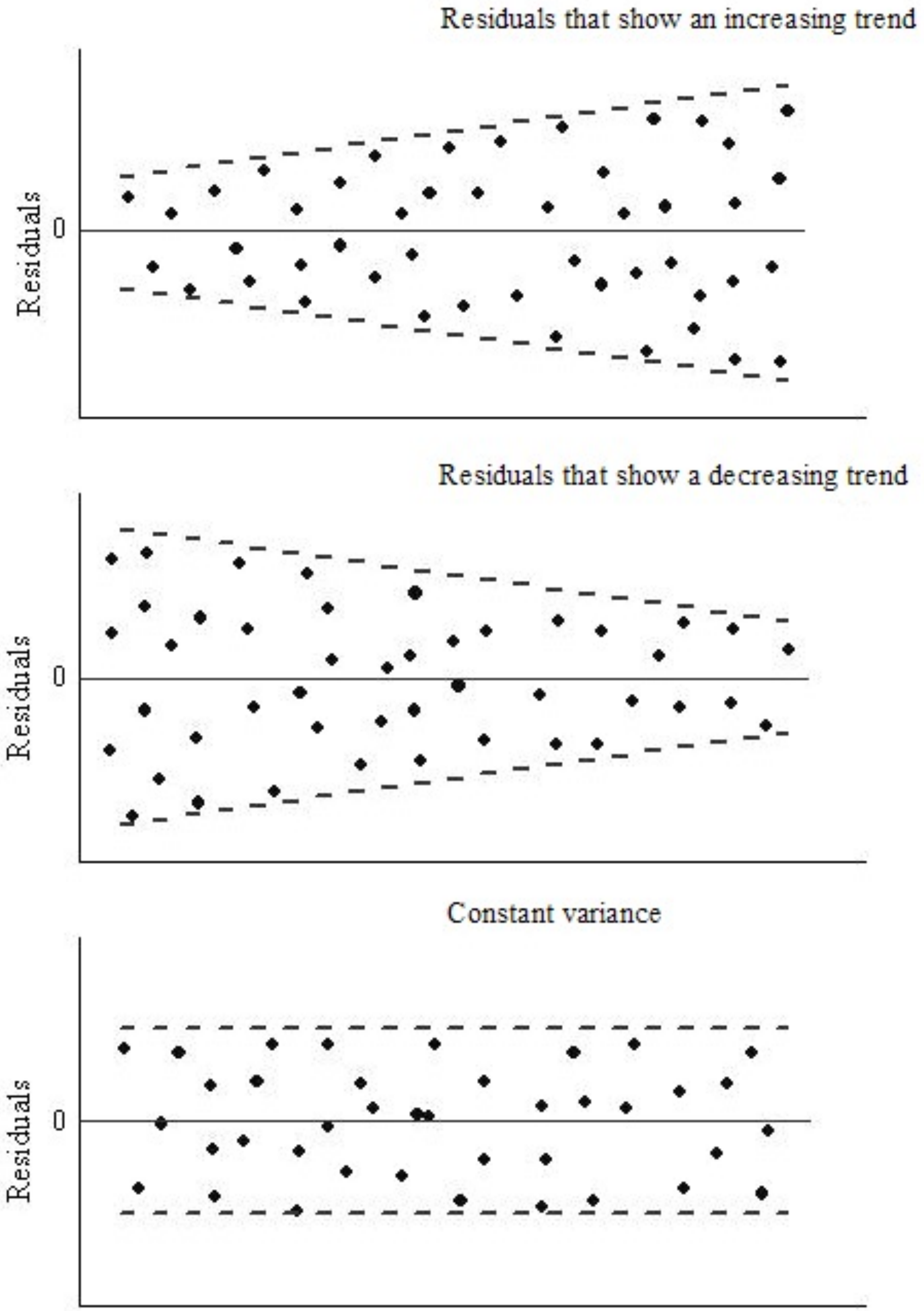
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

where Y is the dependent variable, X_1 to X_p are p independent variables, β_0 to β_p are the coefficients, and ε is the error term



Linear Regression Assumptions

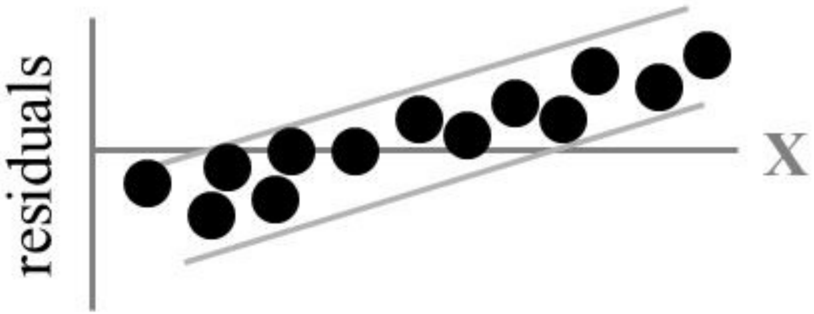
- There is a **linear relationship** between the independent and dependent variables
- **Normality** – (Q-Q plot / Shapiro-Wilk test)
  - Y values are normally distributed for each X
  - Residuals are normally distributed
- Homoscedasticity (**constant variance**) of the residuals
- **Independence of observations**

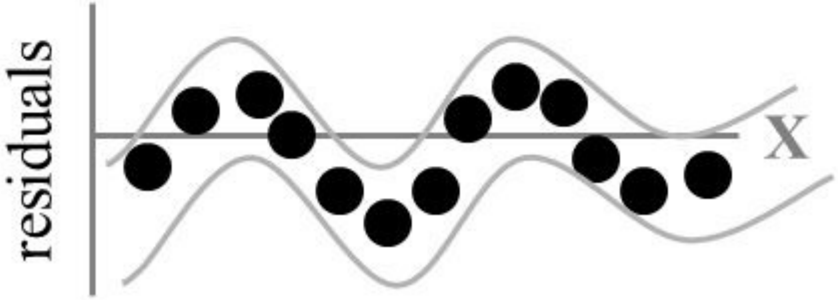


ww.originlab.com/doc/Origin-Help/Residual-Plot-Analysis

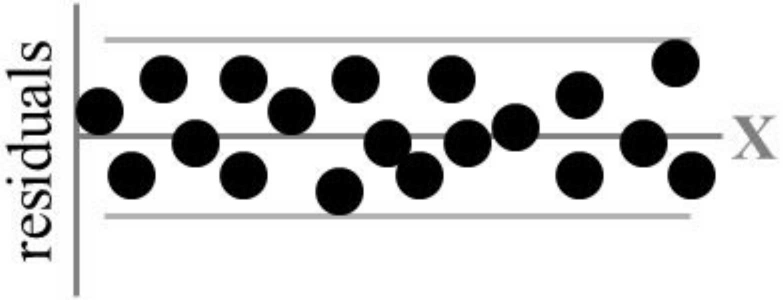
Residual Analysis for Independence

Not Independent





Independent



Linear Regression - Example

Prognostic factors for body fat

- Number of observed individuals: n = 241
- Dependent variable: body fat = percental body fat
- We are interested in the influence of three independent variables:
  - BMI in kg/m2
  - Waist circumference (abdomen) in cm.
  - Waist/hip-ratio

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffffacc68d53/slides-04.pdf https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffffacc68d53/slides-04.pdf

Prognostic factors for body fat - Simple Linear Regression Moedls

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.617	2.939	-9.398	0.000
bmi	1.844	0.116	15.957	0.000

BMI: $R^2 = 0.516$, $R^2_{\text{adj}} = 0.514$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-42.621	2.869	-14.855	0.000
abdomen	0.668	0.031	21.570	0.000

Abdomen: $R^2 = 0.661$, $R^2_{\text{adj}} = 0.659$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-78.066	5.318	-14.680	0.000
waist_hip_ratio	104.976	5.744	18.275	0.000

Waist/hip-ratio: $R^2 = 0.583$, $R^2_{\text{adj}} = 0.581$

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffffacc68d53/slides-04.pdfc

Prognostic factors for body fat - Multiple Linear Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-60.045	5.365	-11.192	0.000
bmi	0.123	0.236	0.519	0.605
abdomen	0.438	0.105	4.183	0.000
waist_hip_ratio	38.468	10.262	3.749	0.000

$R^2 = 0.681,\ R^2_{\text{adj}} = 0.677$

Prognostic factors for body fat - Multiple Linear Regression

Elimination of the non-significant variable bmi:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-59.294	5.158	-11.496	0.000
abdomen	0.484	0.057	8.526	0.000
waist_hip_ratio	36.455	9.486	3.843	0.000

$R^2 = 0.680$, $R^2_{\text{adj}} = 0.678$

Brief Summary

- The relationship between two continuous variables can be visualized using scatter plots
- The relationship between two variables can be assessed using correlation
  - Pearson
  - Spearman
- Regression
  - Understand the relationship between variables
  - Predict the value of one variable based on other variables
- Linear regression is a method for estimating the linear relationship between the dependent and independent variables