Biostatistics
Week XII – part I

Ege Ülgen, MD, PhD

22 December 2022

Conflicting Results

- Researcher A conducts a study comparing the effects an intervention vs. placebo on reducing weight
  - 5 kg reduction among the intervention group ($p = 0.01$)
- Researcher B conducts a similar study comparing the effects an intervention vs. placebo on reducing weight
  - 5 kg reduction among the intervention group ($p = 0.35$)

Statistical Power

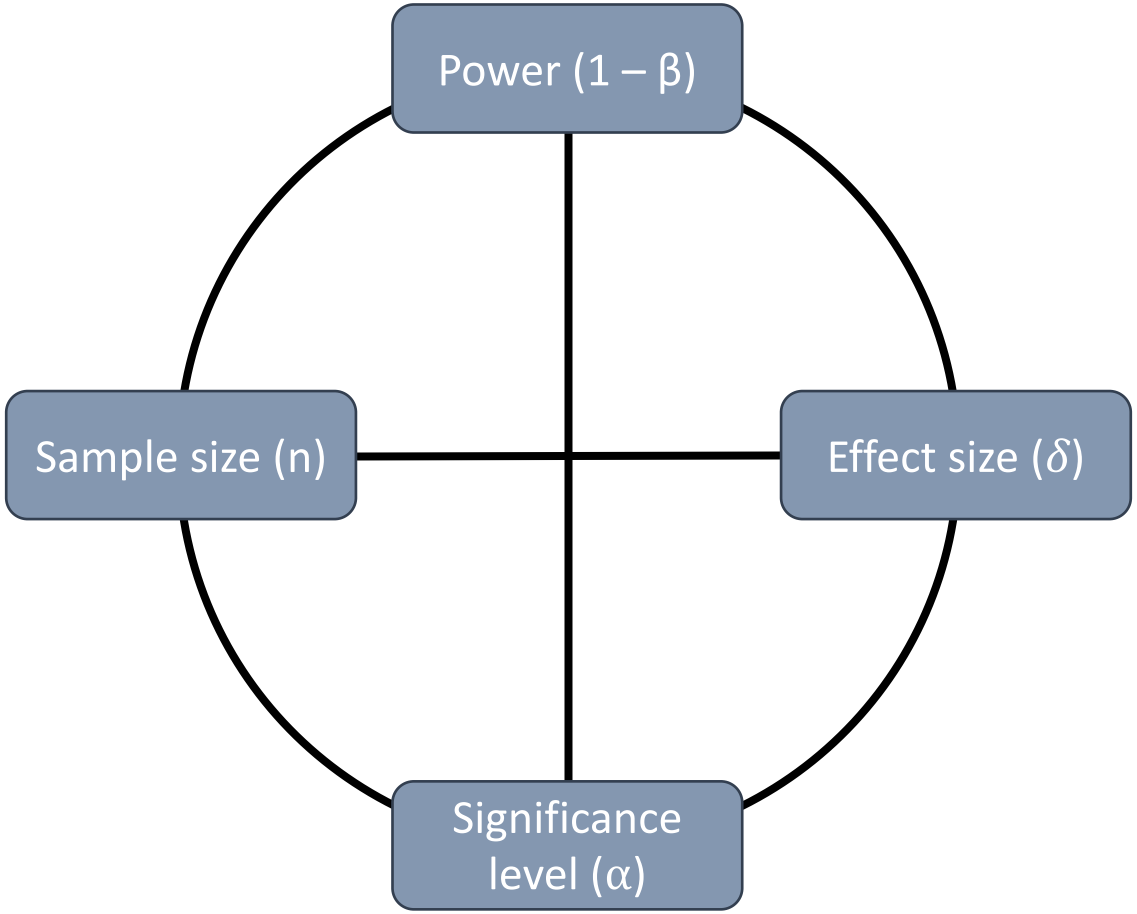
H_0	Decision	
	Fail to reject	Reject
True	Correct decision	**Type I Error** α
False	**Type II Error** **ß**	Correct decision

- **Statistical power** = $1 - \beta$
  - P(reject H_0 | H_0 is false)

Statistical Power

- Power is affected by:
  - Significance level (α)
  - Effect size (δ)
  - Sample size (n)

Power Analysis/Sample Size Calculation



- Given any three, the fourth can be estimated

Default Values

- Power = usually **0.80**, 0.90
- Significance level = usually **0.05**, 0.01, 0.001
- Effect size
  - Literature review
  - Pilot study
  - Cohen’s recommendations

Missing Data

- Missing data, or missing values, occur when no data value is stored for the variable in an observation

- **complications** in handling and analyzing the data
- **bias** resulting from differences between missing and complete data

Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002 Jun;7(2):147–77.

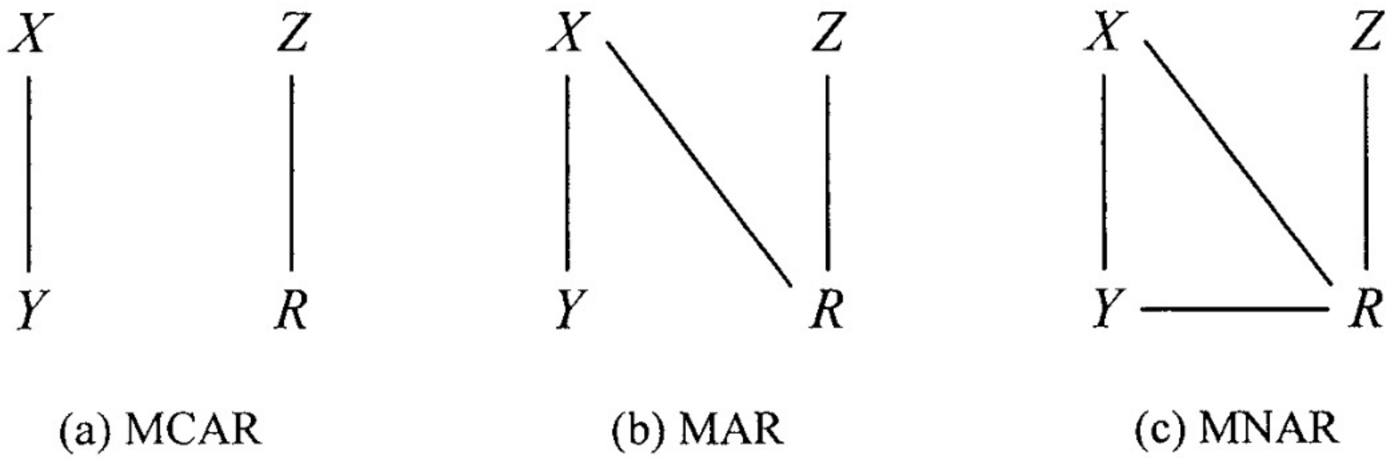


Figure 2. Graphical representations of (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR) in a univariate missing-data pattern. X represents variables that are completely observed, Y represents a variable that is partly missing, Z represents the component of the causes of missingness unrelated to X and Y, and R represents the missingness.

Missing Data - Missing Completely at Random (MCAR)

- The missingness of the data is not associated with either any variable or outcome
- There is **nothing systematic** going on that makes some data more likely to be missing than others

- e.g., Questionnaire lost, blood tube for testing a blood level broken etc.

Missing Data - Missing at Random (MAR)

- The missingness of the data is **associated with a variable**

- e.g., Supposing men are more likely to tell their weight than women, missingness in weight is MAR

Missing Data - Missing Not at Random (MNAR)

- The missingness of the data is **related with the outcome**
- e.g., In a depression study, the depression score wasn’t calculated for a participant because they committed suicide

Missing Data

- There are several strategies to cope with missing data:
    - **Try to collect the missing data** (obvious best choice)
    - **Exclude** subjects with any missing data (may reduce the power of the study)
    - **Replace** the missing data with a conservative estimate (e.g., sample mean)
    - **Estimate** the missing data from other data on the same subject (imputation)

Brief Summary

- Given any three of the following, the fourth can be determined:
  - Power
  - Significance level
  - Effect size
  - Sample size
- Determining sample size prior to starting a study is important
  - Too small of a sample size can under detect the effect of interest in your experiment
  - Too large of a sample size may lead to unnecessary wasting of resources
- There are 3 kinds of missing data:
  - MCAR: nothing systematic
  - MAR: missingness associated with a variable
  - MNAR: missingness related with the outcome

Biostatistics
Week XII – part II

Ege Ülgen, MD, PhD

22 December 2022

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 (or more) variable(s) 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 based on certain variables

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

$$\underbrace{Y_i}_{\text{Dependent variable}} = \overbrace{\beta_0}^{\text{Intercept}} + \overbrace{\beta_1}^{\text{slope}} \underbrace{X_i}_{\text{Independent variable}} + \overbrace{\varepsilon_i}^{\text{residual}} \qquad \varepsilon \sim N(0, \sigma^2)$$

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

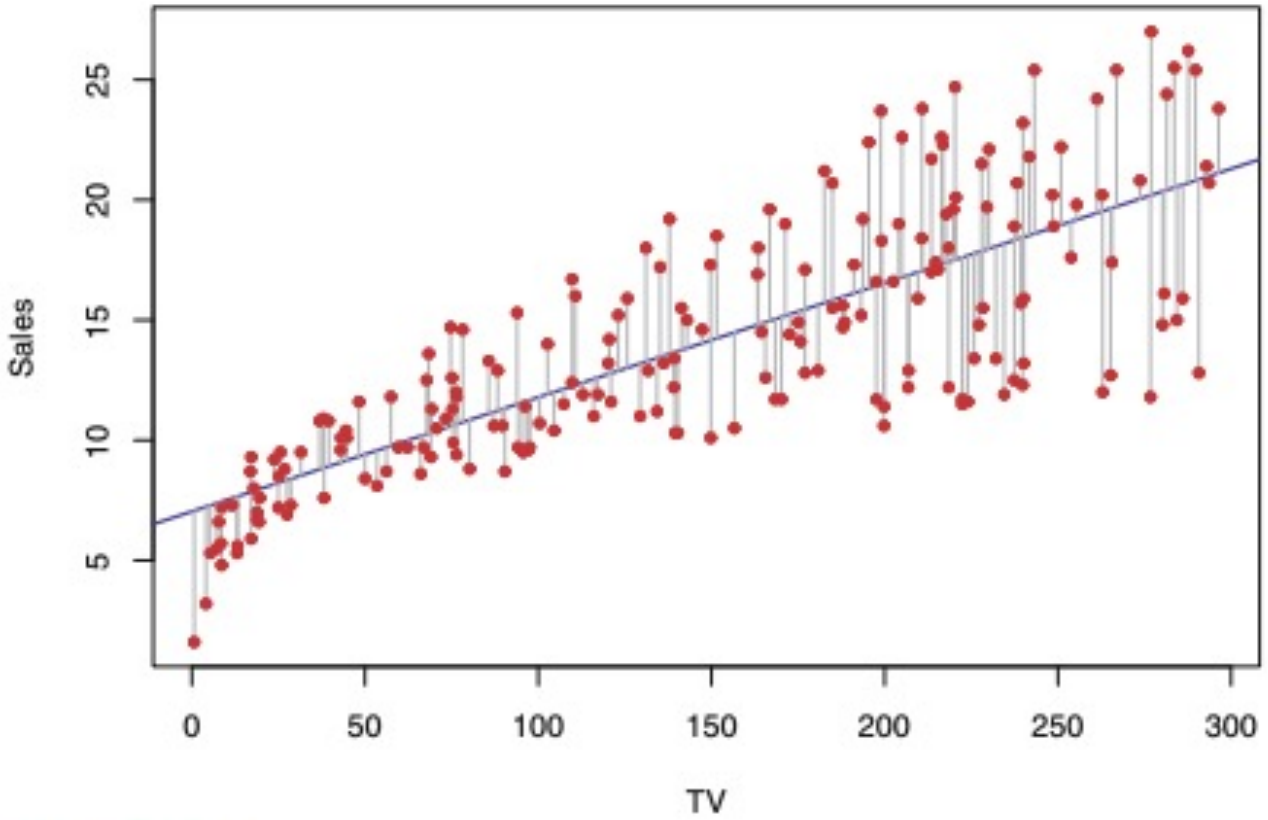


FIGURE 3.1. *For the* Advertising *data, the least squares fit for the regression of* sales *onto* TV *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures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

Error/residual = Actual value – Predicted value

Estimation of Coefficients

- $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are estimates by **least squares**
  - minimizing the sum of the squares of the residuals

$$\boldsymbol{Residual\ sum\ of\ squares(RSS)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\boldsymbol{2}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \widehat{\boldsymbol{\beta_0}} - \widehat{\boldsymbol{\beta_1}}\boldsymbol{x_i})^{\boldsymbol{2}}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

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

Estimate of σ

- Known as the **residual standard error (RSE)**

$$RSE = \sqrt{\frac{RSS}{n - 2}}$$

- Using RSE, the standard errors for $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are estimated
  - **Significance testing** by t-test $\left(\frac{\widehat{\beta_i}}{se(\widehat{\beta_i})} \sim t_{n-1}\right)$

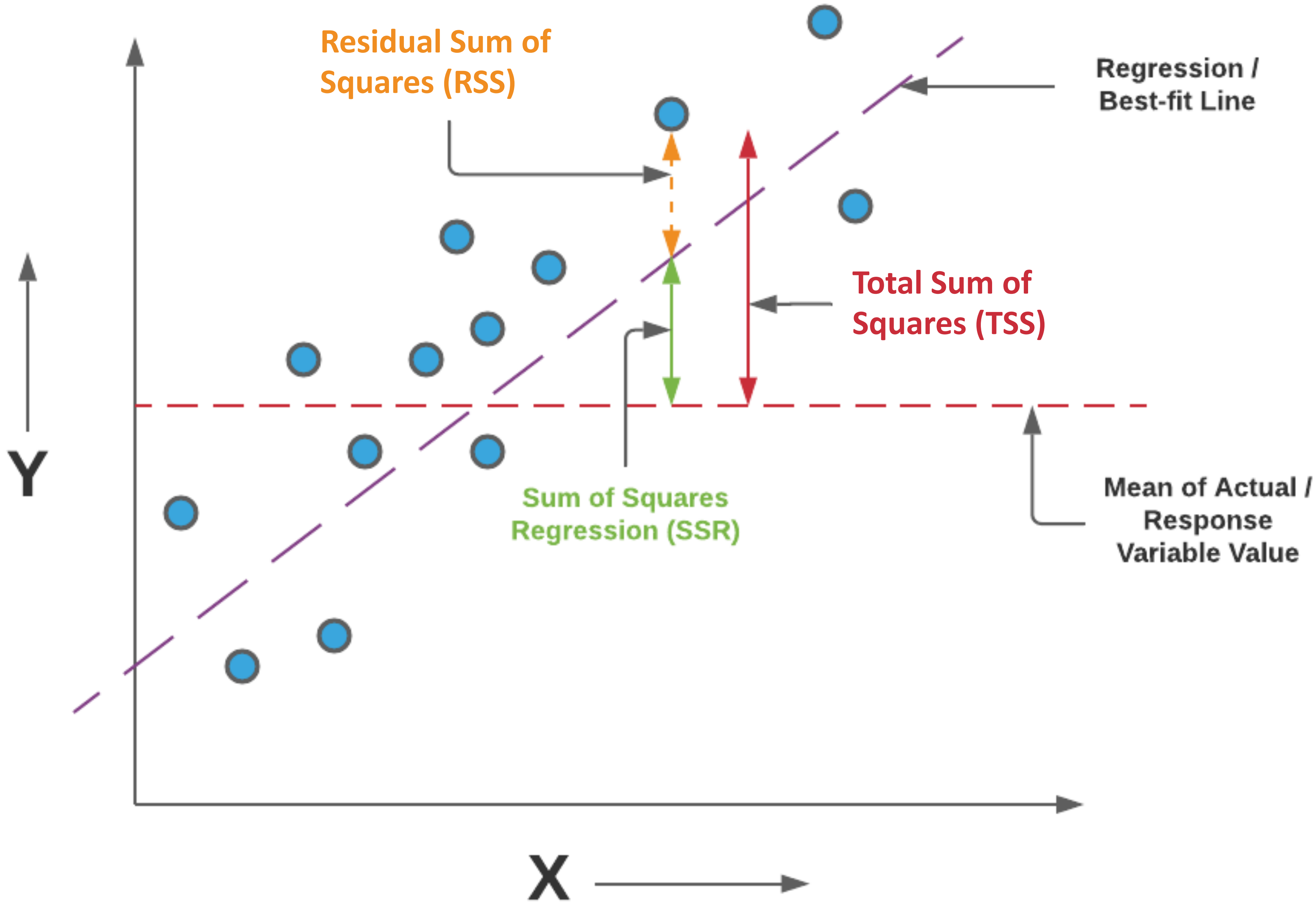
R^2 – Coefficient of Determination

- R^2 is measure of **how good is the regression** or best fit line
- It is also termed as **coefficient of determination**

$$R^2 = 1 - \frac{RSS}{TSS}$$

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

- Measures **the proportion of variation in Y that is explained by the independent variable X** in the regression model
- Greater the value of R-Squared, the better is the regression line as higher is the variance explained by the regression line



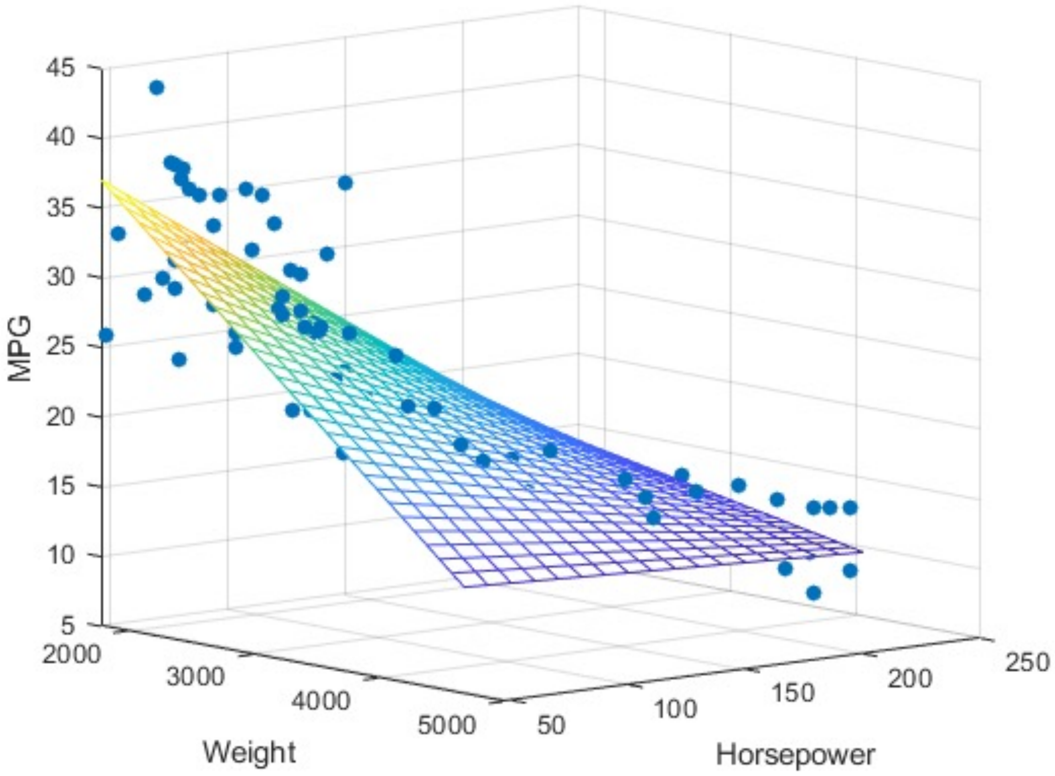
Kumar A. Linear regression explained with python examples [Internet]. Data Analytics. 2020 [cited 2021 Oct 19]. Available from: https://vitalflux.com/linear-regression-explained-python-sklearn-examples/

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



Multiple Linear Regression (cont.)

- $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_p}$ are estimates by **least squares**
- Significance testing by ANOVA
  - H_0: $\widehat{\beta_0} = \widehat{\beta_1} = \ \ldots = \widehat{\beta_p} = 0$
- $RSE = \sqrt{\frac{RSS}{n-p-1}}$
- **Adjusted R^2** is a modified version of R^2 that has been adjusted for the **number of predictors** in the model

$$R^2_{adj} = \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

$$where:$$

$$R^2 = R - squared$$

$$n \quad = number\ of\ samples/rows\ in\ the\ data\ set$$

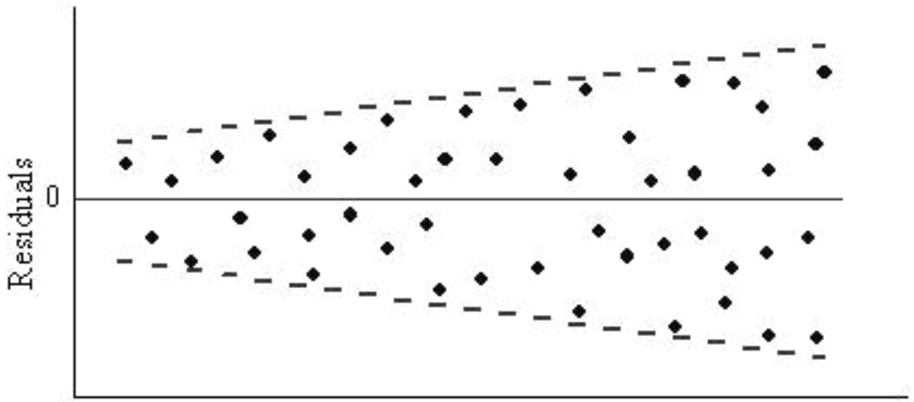
$$p \quad = number\ of\ predictors/features$$

Muralidhar KSV. Demystifying r-squared and adjusted r-squared [Internet]. Medium. 2021 [cited 2021 Dec 14]. Available from: https://towardsdatascience.com/demystifying-r-squared-and-adjusted-r-squared-52903c006a60

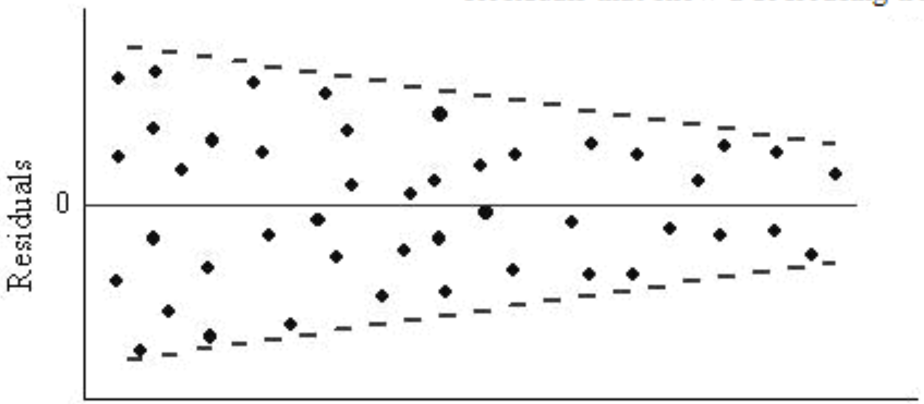
Linear Regression Assumptions

- 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**

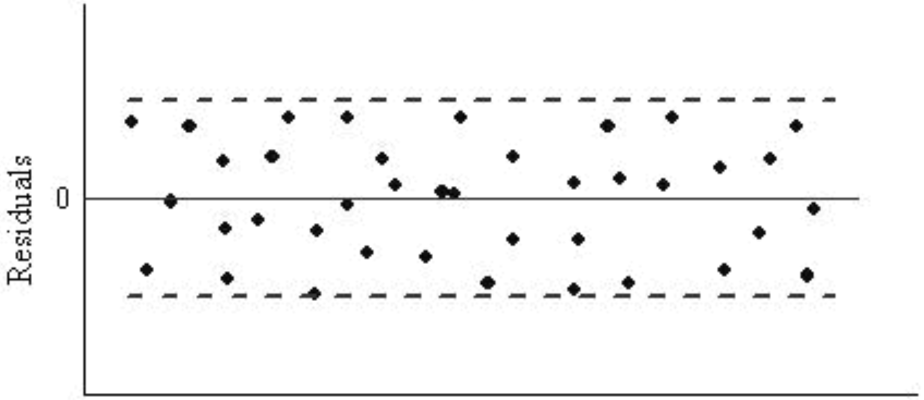
Residuals that show an increasing trend



Residuals that show a decreasing trend



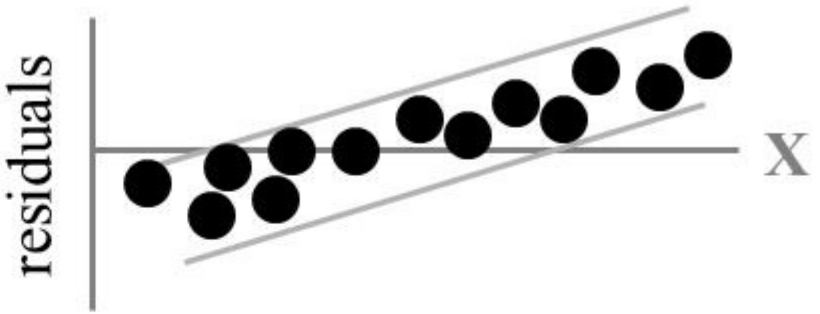
Constant variance

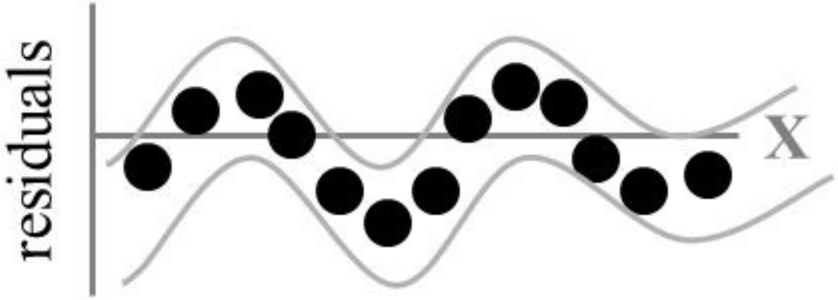


ww.originlab.com/doc/Origin-Help/Residual-Plot-Analysis

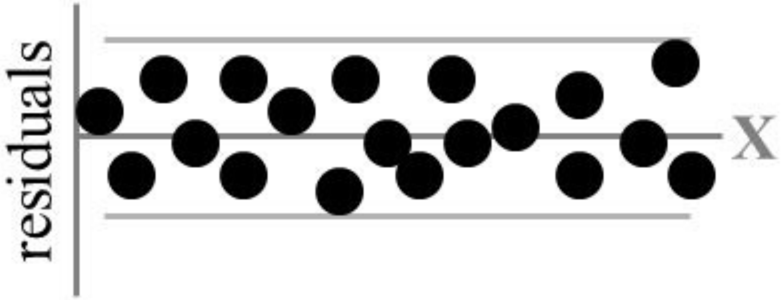
Residual Analysis for Independence

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

Example - 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

$R^2 = 0.681$, $\boxed{R^2_{adj} = 0.677}$ the proportion of the variation in the dependent variable that is predictable from the independent variable

$$Estimated\ Body\ Fat = -60.045 + 0.123 * bmi + 0.438 * abdomen + 38.468 * waist\_hip\_ratio$$

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

Example - 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

- For a person with bmi = 0, abdomen = 0, waist_hip_ratio = 0, the boy fat is estimated to be -60.045 ($p < 0.001$)
- (Keeping all other variables the same) with one unit increase in bmi, body fat increases by 0.123 (not significant since $p > 0.05$)
- With 95% confidence, it can be stated that with one unit increase in abdomen, body fat increases by 0.438 ($p < 0.001$)
- With one unit increase in waist_hip_ratio, body fat increases by 38.468 ($p < 0.001$)

Example II

- We’ll analyze the prostate cancer dataset
- The main aim of collecting this data set was to inspect the associations between **prostate-specific antigen (PSA)** and **prognostic clinical measurements** in men advanced prostate cancer
- Data were collected on 97 men who were about to undergo radical prostectomies

**PSA was transformed to logPSA for “normalization”*

Example II – Model 1

$$logPSA = 1.8 + 0.07 \; * \boldsymbol{vol} + 0.77 \; * I(\boldsymbol{invasion} = \mathbf{1})$$

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	1.8035	0.1141	15.81	<0.001
vol	0.0725	0.0133	5.43	<0.001
invasion1	0.7755	0.2541	3.05	0.003

Adjusted R-squared: 0.472

Example II – Model 2

$$logPSA = 1.67 + 0.1021 \ * \boldsymbol{vol} + 1.326 \ * I(\boldsymbol{invasion} = \mathbf{1}) - 0.056 * I(\boldsymbol{invasion} = \mathbf{1}) * \boldsymbol{vol}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.6673	0.1289	12.94	<0.001
vol	0.1021	0.0191	5.35	<0.001
invasion1	1.326	0.3588	3.7	<0.001
vol:invasion1	-0.056	0.0262	-2.13	0.0354

Adjusted R-squared: 0.491

For a patient with invasion, there is an additional -0.056 change in PSA when vol changes one unit
= For a patient with invasion, one unit change in volume results in (0.1021 – 0.056) change in PSA

Example II – Model 3

$$logPSA = 1.55 + 0.076 \ * \boldsymbol{vol} + 0.45 \ * I(\boldsymbol{Gleason} = \mathbf{7}) + 0.9 * I(\boldsymbol{Gleason} = \mathbf{8})$$

(compared to **Gleason = 6**)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5523	0.1548	10.02	< 2e-16
vol	0.0758	0.0131	5.79	9.30E-08
Gleason7	0.4521	0.1928	2.34	0.0212
Gleason8	0.9043	0.2747	3.29	0.0014

Adjusted R-squared: 0.48

Brief Summary

- 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