STATISTICS

ENG 3120

2023 - 2024 Spring Semester

Assoc. Prof. Dr. Bora CANBULA

www.canbula.com

github.com/canbula/Statistics

wn45g9v

Manisa Celal Bayar University *Department of Computer Engineering*

Syllabus

STATISTICS

Instructor

Assoc. Prof. Dr.
Bora CANBULA

Phone

0 (236) 201 21 08

Email

bora.canbula@cbu.edu.tr

Office Location

Dept. of CENG

Office C233

Office Hours

4 pm – 5 pm, Mondays

Course Overview

Statistics (Teams Code: wn45g9v)

We are going to learn both the mathematical foundations and real-world application of the statistics and the probability in this course. Focus of this course will be to provide the required background for a data science / machine learning course. Python is preferred as the programming language for the applications of this course.

Required Text

Probability And Statistics for Computer Scientists, CRC Press, *Michael Baron*

Introduction to Probability and Statistics, Elsevier, *Sheldon M. Ross*

Probability and Statistics for Engineers and Scientists, Brooks/Cole, *A.J. Hayter*

Course Materials

Python 3.x (Anaconda is preferred)

Jupyter Notebook from Anaconda

Pycharm from JetBrains / Visual Studio Code from Microsoft

Week	Subject	Week	Subject
1	Definitions of Descriptive Statistics	8	Linear Regression
2	Data, Sampling, and Variation	9	Linear Regression with Matrix Algebra
3	Visualization of Data	10	Regression with High Degree Polynomials
4	Measures of Central Tendency	11	Data Linearization and Transformation
5	Measures of Variation	12	Chi-Square and Goodness-of-Fit Tests
6	Measures for Multiple Variables	13	Central Limit Theorem
7	Box Plots and Outliers	14	Probability Distributions

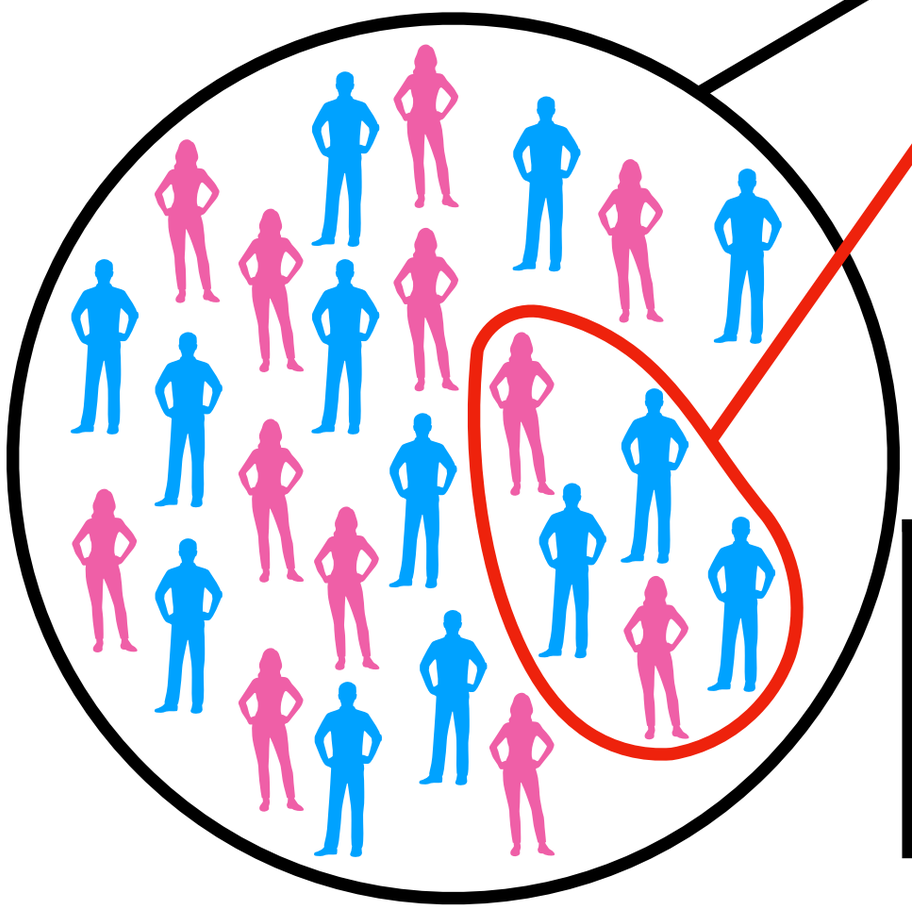
Definitions of Descriptive Statistics

STATISTICS

Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting the **data**.

Data is any kind of information.

Data are the actual values of the variable and can be categorical or numerical.



good data == good sample

a good sample must be both random and representative

Statistics

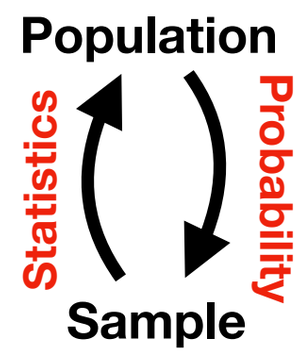
- **Descriptive Statistics** is organizing and summarizing the data.
- **Inferential Statistics** is drawing conclusions from good data.

Population is the collection of people, things, or objects under study.

Sample is a subset of the population.

Statistic is a number that represents a property of the sample.

Parameter is a characteristic of the whole population that can be estimated by a statistic.



Variable is a characteristic or measurement that can be determined for each member of a population. They can be **dependent** or **independent**.

- **Qualitative Variables** take on **Categorical** values.
- **Quantitative Variables** take on **Numerical** values.
  - **Discrete Variables** take on finite number of values such as integers.
  - **Continuous Variables** take on infinite number of values such as real numbers.

Definitions of Descriptive Statistics

STATISTICS

QUESTION

A study was conducted at our department to analyze the average GPA's of students who graduated last year. Match the key terms given below with the phrases that describes best.

A) Population **B)** Statistic **C)** Parameter **D)** Sample **E)** Variable **F)** Data

D A group of students who graduated from our department last year

X All students who attended last year

E GPA of one student who graduated from our department last year

C The average GPA of students who graduated from our department last year

A All students who graduated from our department last year

F 3.65, 2.80, 3.15, 3.90

B _____

Definitions of Descriptive Statistics

STATISTICS

QUESTION

We plan on conducting a survey to our recent graduates to determine information on their yearly salaries. We randomly select 50 recent graduates and sent them questionnaires dealing with their present jobs. Of these 50, however, only 36 were returned. Suppose that the average of the yearly salaries reported was 415000 TL.

- The population is: **Our all recent graduates**
- The sample is: **36 recent graduates who returned to questionnaire**
- The statistic is: **Yearly salary of 36 students**
- The parameter is: **Yearly salary of our all recent graduates**
- The variable is: **Yearly salary of one recent graduates**
- Would we be correct in thinking that 415000 TL was a good approximation to the average salary level for all of our graduates? **No**
- If your answer is no, can you think of any set of conditions relating to the group that returned questionnaires for which it would be a good approximation? **Suggest some questions**

Definitions of Descriptive Statistics

STATISTICS

QUESTION

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been in a malpractice lawsuit.

- The population is: **All medical doctors listed in the prof. directory**
- The sample is: **Selected 500 doctors**
- The statistic is: **The proportion of medical doctors ……. in the sample**
- The parameter is: **The proportion of medical doctors ……. in population**
- The variable is: **The number of medical doctors who have been ……**
- The data are: **Yes / No**

Definitions of Descriptive Statistics

STATISTICS

QUESTION

Determine the correct data type for the variables given below. Indicate whether quantitative data are continuous or discrete.

A) Numerical and discrete **B)** Numerical and continuous **C)** Categorical

A The number of pairs of shoes you own

C Gender

B The distance from your home to university

A The number of courses you take this semester

C The brand of your mobile phone

B Your weight

A Number of correct answers on a quiz

? Age

Sampling

Population (N) ⟶ **Sample (n)**

BEST

HARDEST

Simple Random Sampling

Every member of the population has an equal chance to be in the sample.

Stratified Sampling

The population is split into non-overlapping groups, which is called strata, then simple random sampling is applied to each group.

Cluster Sampling

The population is divided into groups (clusters), then some of the clusters are randomly selected.

Systematic Sampling

The sample is constructed with every n^{th} individual from the population.

Convenience Sampling

The sample is constructed with easily obtained members of the population.

WORST

EASIEST

Determine the type of sampling used in the following examples:

- A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.

  Stratified Sampling
- A pollster interviews all human resource personnel in five different high tech companies.

  Cluster Sampling
- A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.

  Stratified Sampling
- A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.

  Systematic Sampling
- A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.

  Simple Random Sampling
- A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

  Convenience Sampling

Sampling

Population (N) ⟶ **Sample (n)**

BEST

HARDEST

Simple Random Sampling

Every member of the population has an equal chance to be in the sample.

```
cities = [
    "Adana",
    "Adıyaman",
    "Afyonkarahisar",
    "Ağrı",
    # ...
    "Kilis",
    "Osmaniye",
    "Düzce",
]
```

```
import sys

sys.path.append(".")


from Week02 import data

print(data.cities)
```

```
import random
```

```
[
    "betavariate",
    "binomialvariate",
    "choice",
    "choices",
    "expovariate",
    "gammavariate",
    "gauss",
    "getrandbits",
    "getstate",
    "lognormvariate",
    "normalvariate",
    "paretovariate",
    "randbytes",
    "randint",
    "random",
    "randrange",
    "sample",
    "seed",
    "setstate",
    "shuffle",
    "triangular",
    "uniform",
    "vonmisesvariate",
    "weibullvariate",
]
```

```
Help on method sample i

sample(population, k, *
    Chooses k unique ra

    Returns a new list
    leaving the origina
    in selection order
    samples.  This allo
    into grand prize an

Members of the popu
population contains
selection in the sa

Repeated elements c
counts parameter.

    sample(['red',

is equivalent to:

    sample(['red',

To choose a sample
population argument
for sampling from a

    sample(range(10
```

```
import sys

sys.path.append(".")


from Week02 import data
import random


def simple_random_sampling(data, n):
    return random.sample(data, n)


sample = simple_random_sampling(data.cities, 10)
print(sample)
```

WORST

EASIEST

Sampling

Population (N) ⟶ **Sample (n)**

BEST

Stratified Sampling

The population is split into non-overlapping groups, which is called strata, then simple random sampling is applied to each group.

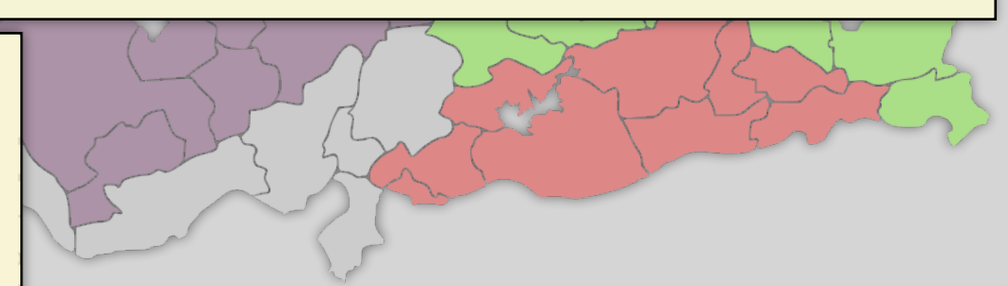
```python
regions = [
    "Marmara",
    "İç Anadolu",
    "Ege",
    "Akdeniz",
    "Karadeniz",
    "Doğu Anadolu",
    "Güneydoğu Anadolu",
]
```

```python
def stratified_sampling(data, n, strata):
    sample = []
    for key in strata:
        sample += random.sample(data[key], n)
    return sample


sample = stratified_sampling(data.cities_by_region, 3, data.regions)
print(sample)
```

Fix this!

```python
cities_by_region = {
    "Marmara": ["Edirne", "Kırklareli", "Tekirda
    "İç Anadolu": ["Aksaray", "Ankara", "Çankırı
    "Ege": ["İzmir", "Manisa", "Aydın", "Denizli
    "Akdeniz": ["Adana", "Osmaniye", "Antalya",
    "Karadeniz": ["Rize", "Trabzon", "Artvin",
    "Doğu Anadolu": ["Ağrı", "Ardahan", "Bingöl",
    "Güneydoğu Anadolu": ["Adıyaman", "Batman",
}
```



HARDEST

WORST

EASIEST

Sampling

Population (N) ⟶ Sample (n)

BEST

HARDEST

Cluster Sampling

The population is divided into groups (clusters), then some of the clusters are randomly selected.

```python
regions = [
    "Marmara",
    "İç Anadolu",
    "Ege",
    "Akdeniz",
    "Karadeniz",
    "Doğu Anadolu",
    "Güneydoğu Anadolu",
]
```

```python
def stratified_sampling(data, n, strata):
    sample = []
    for key in strata:
        sample += random.sample(data[key], n)
    return sample


sample = stratified_sampling(data.cities_by_region, 3, data.regions)
print(sample)
```

Fix this!

```python
cities_by_region = {
    "Marmara": ["Edirne", "Kırklareli", "Tekirda
    "İç Anadolu": ["Aksaray", "Ankara", "Çankırı
    "Ege": ["İzmir", "Manisa", "Aydın", "Denizli
    "Akdeniz": ["Adana", "Osmaniye", "Antalya",
    "Karadeniz": ["Rize", "Trabzon", "Artvin",
    "Doğu Anadolu": ["Ağrı", "Ardahan", "Bingöl
    "Güneydoğu Anadolu": ["Adıyaman", "Batman",
}
```

```python
def cluster_sampling(data, n, clusters):
    picked_clusters = random.sample(clusters, n)
    sample = []
    for cluster in picked_clusters:
        sample += data[cluster]
    return sample
```

Duplicates?

WORST

EASIEST

Sampling

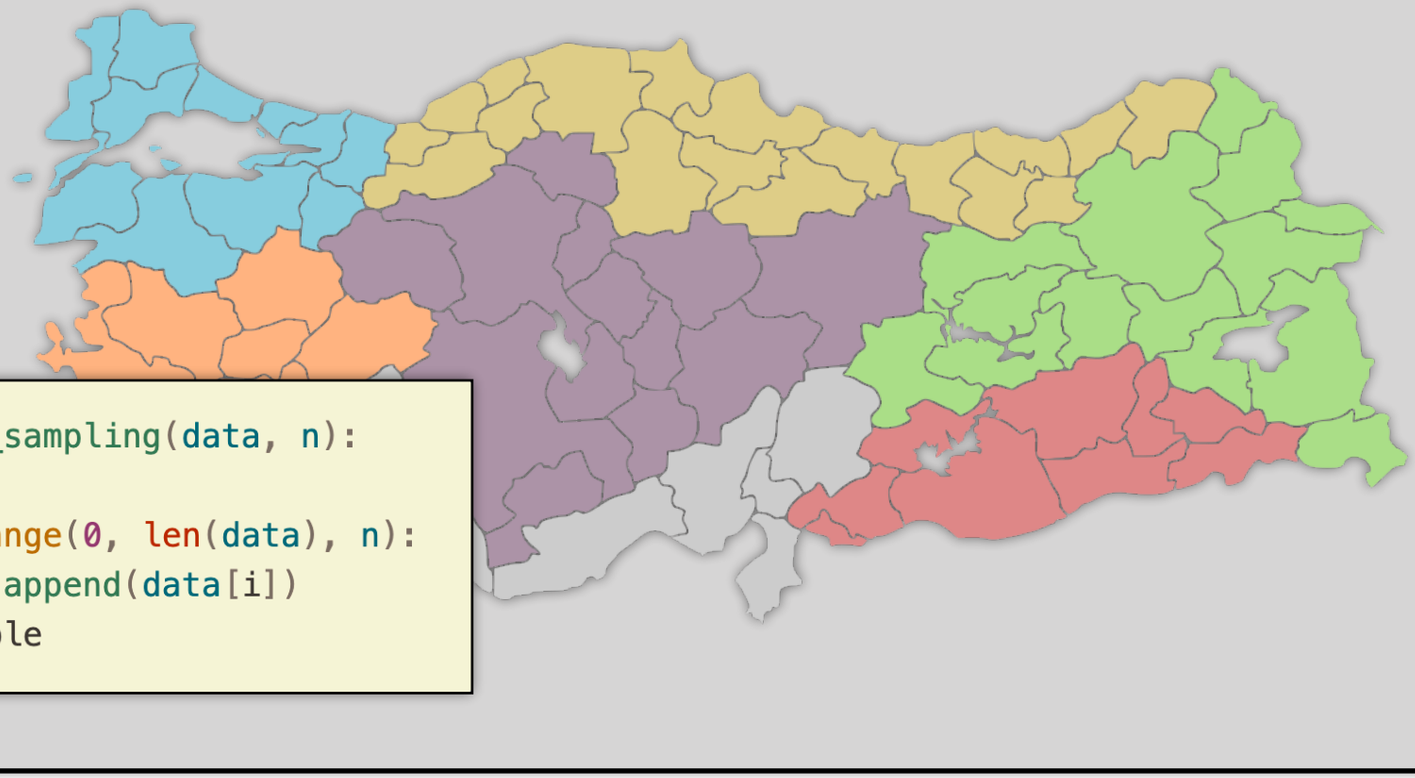
Population (N) ⟶ **Sample (n)**

BEST

HARDEST

Systematic Sampling

The sample is constructed with every n^{th} individual from the population.



```python
def systematic_sampling(data, n):
    sample = []
    for i in range(0, len(data), n):
        sample.append(data[i])
    return sample
```

WORST

EASIEST

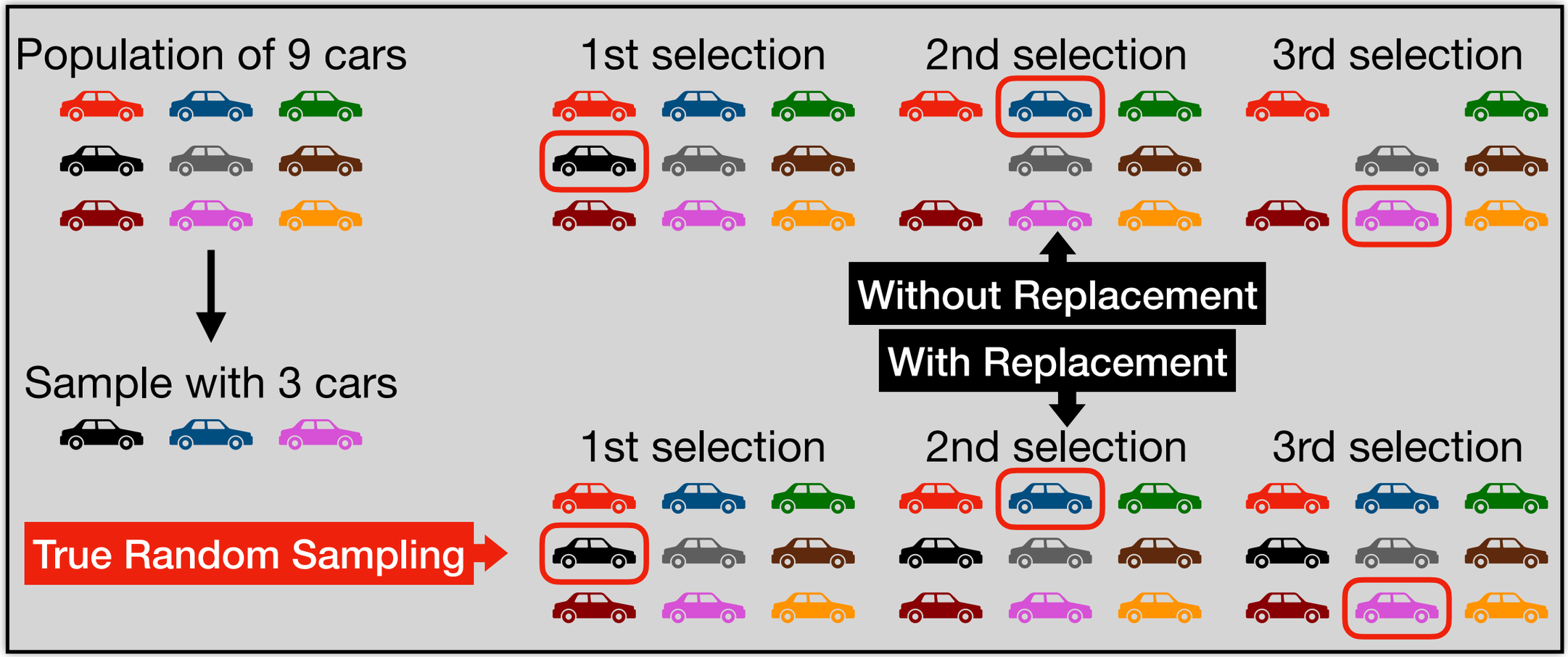
Sampling

Population (N) ⟶ **Sample (n)**

BEST

Simple Random Sampling

Every member of the population has an equal chance to be in the sample.



WORST

HARDEST

EASIEST

IMPLEMENT THIS FEATURE AS A HOMEWORK

Weighted Simple Random Sampling w/ Replacement Support

Name of Your File

Week03/weighted_firstname_lastname.py

Name of the Function

weighted_srs

Input Parameters

- data [list]: population
- n [int]: sample size
- weights [list]: weights for members of population
- with_replacement [bool]: flag for true random sampling

Other Rules

- Using maximum 10 lines of codes is allowed
- Using any modules, other than random, is forbidden

Levels of Measurement

The way a set of data is measured is called its **level of measurement**. The correct statistical procedures that can be used with a data set is specified with this level.

	Difference	Order	Similar Intervals	Natural Zero
Less Information				
Categorical Data — **Nominal**: Nominal level represents the categories that **cannot** be put in any order	✓	✗	✗	✗
Categorical Data — **Ordinal**: Ordinal level represents the categories that **can** be put in a order	✓	✓	✗	✗
Numerical Data — **Interval**: Interval level has a definite ordering, and **distances between values are equal and meaningful**	✓	✓	✓	✗
Numerical Data — **Ratio**: Ratio level provides the most information: order, fixed scale, and also a **natural zero**	✓	✓	✓	✓
More Information				

Levels of Measurement

STATISTICS

Determine the type of measure scale used in the following examples:

- Letter Grades: AA, BA, BB, CB, CC, …

  Ordinal Scale

- The number of students in a classroom

  Ratio Scale

- The dates 1997, 2004, 2020, …

  Interval Scale

- Political outlook: extreme left, left-of-center, right-of-center, extreme right

  Nominal Scale

- Turkish Republic identification number

  Nominal Scale

Measures of Central Tendency

STATISTICS

Measures of central tendency are used to determine the center of a distribution of data. It is used to find a single score that is the most representative of an entire data set.

Mean is simply the arithmetic **average** of the data observations.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Median is the value in the **middle** of the ordered data points.

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & n : \text{odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n : \text{even} \end{cases}$$

Mode is the value with the **highest frequency** of the data set.

QUESTION

If a constant c is added to each x_i in a sample, yielding $y_i = x_i + c$ how do the sample mean and median of the y_is relate to the mean and median of the x_is?

If each x_i is multiplied by a constant c, yielding $y_i = cx_i$, answer the question again.

HW

Week05/shifted_firstname_lastname.py

Function **`shifted`** calculates the difference between the mean and median in percentage.

Measures of the Location of the Data

STATISTICS

Quantiles are points that divide a data set or a distribution into intervals with equal distribution. They are cut-off points at which certain percentages of the data fall below them.

Quantiles are derived from **order statistics**, which are simply the values in a sorted data set.

x_1: first order statistic, x_2: second order statistic, …, x_n: n-th order statistic

Common types of quantiles are: Quartiles, Percentiles, Deciles, Quintiles

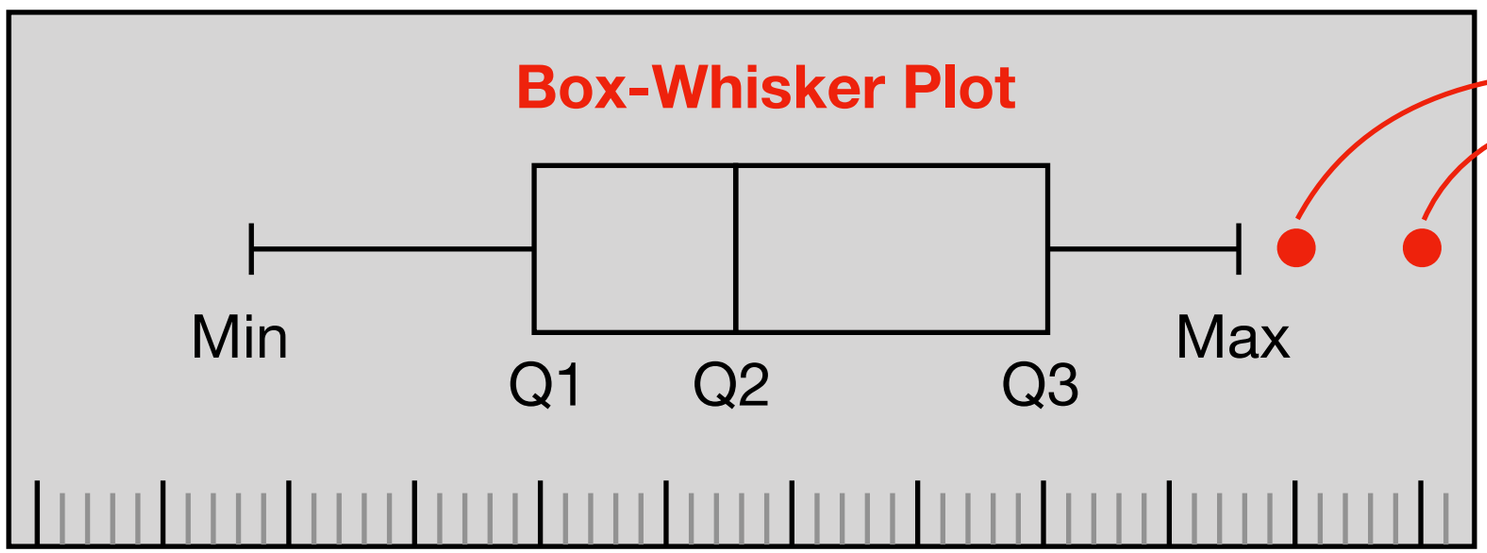
Quartiles divide order statistics into four equal parts.

- **Q1** is the first quartile (lower quartile), the value below which 25% of the data falls. It is the median of the lower half of the data set.
- **Q2** is the second quartile and essentially the median of the data set, dividing data into two equal halves. 50% of the data lies below this point.
- **Q3** is the third quartile (upper quartile), it is the value below which 75% of the data falls. It is the median of the upper half of the data set.

Quartiles
$n/4$
Percentiles
$n/100$
Deciles
$n/10$
Quintiles
$n/5$

Five-Point Summary

- Minimum
- Q1
- Q2
- Q3
- Maximum



Outliers

Data points that differ significantly compared to other observations.

$< Q1 - 1.5 \times \text{IQR}$

$> Q3 + 1.5 \times \text{IQR}$

IQR (Interquartile Range) is the range between the first quartile (Q1) and the third quartile (Q3) of order statistics. Essentially, it covers the middle 50% of the data.

$$\text{IQR} = \text{Q3} - \text{Q1}$$

QUESTION

The following data are the number of pages in 40 books on a shelf. Construct a box plot.

136; 140; 178; 190; 205; 215; 217; 218; 232; 234; 240; 255; 270; 275; 290; 301; 303; 315; 317; 318; 326; 333; 343; 349; 360; 369; 377; 388; 391; 392; 398; 400; 402; 405; 408; 422; 429; 450; 475; 512

Q

Collect the heights of the students in the class within 3 or 4 samples. Construct box plots for each sample. Compare the box plots.

Measures of Variation

STATISTICS

Measures of variation are used to determine the spread or variability of the data. These measures are crucial in understanding the diversity and distribution of data within a data set.

Range is the difference between the minimum and the maximum values in the data set. The range is sensitive to outliers and may not represent the typical spread of the data.

Variance measures how far a data set is spread out.
It is the average value of the squares of the distances between the values and the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \longrightarrow \text{Bessel's Correction}$$

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Standard Deviation is an easy fix to situation that variance can be too large in many cases, also it much more meaningful than variance as it is in the same unit with data points.

$$\sigma = \sqrt{\sigma^2} \qquad s = \sqrt{s^2}$$

Coefficient of Variation is a standardized measure of dispersion. While it doesn't have a unit of measurement, it is universal and perfect for comparisons of different data sets.

$$c_v = \frac{\sigma}{\mu} \qquad \hat{c}_v = \frac{s}{\bar{x}}$$

Calculate the measures of variation for your collected data of the heights of the students.

Measures of Relationship Between Variables

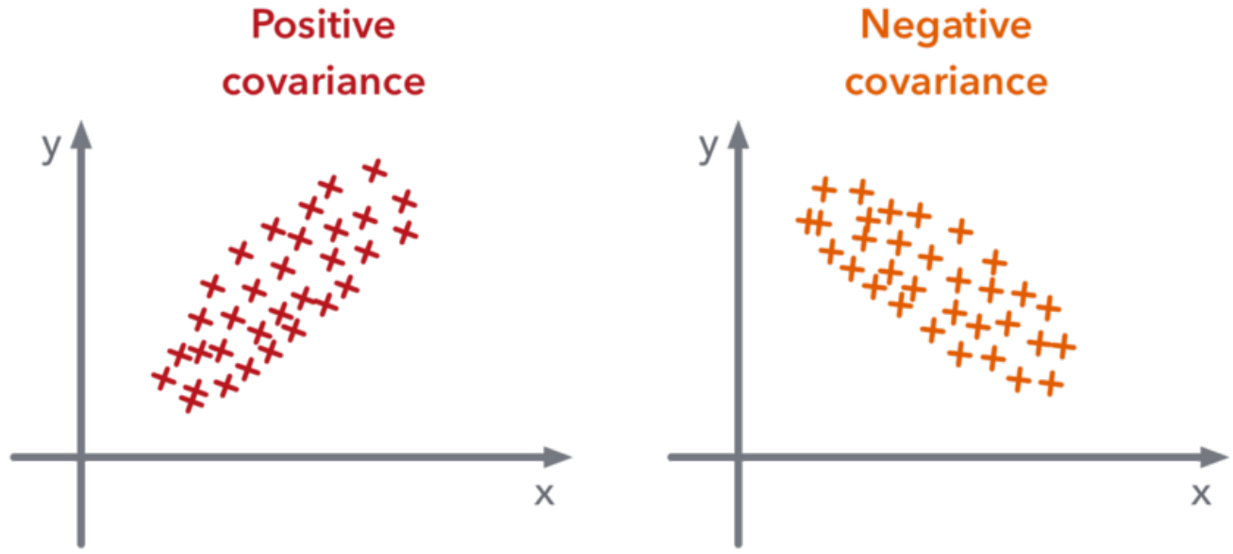
STATISTICS

Measures of relationship between variables are used to determine the dependency of one variable to another. They quantify the strength and direction of the relationship.

Covariance is used to describe how much two random variables vary together. It is similar to variance, but where variance is about one variable, covariance is about two variables.

$$\sigma_{xy} = \frac{\sum_{i=1}^{N} (x_i - \mu_x) \cdot (y_i - \mu_y)}{N}$$

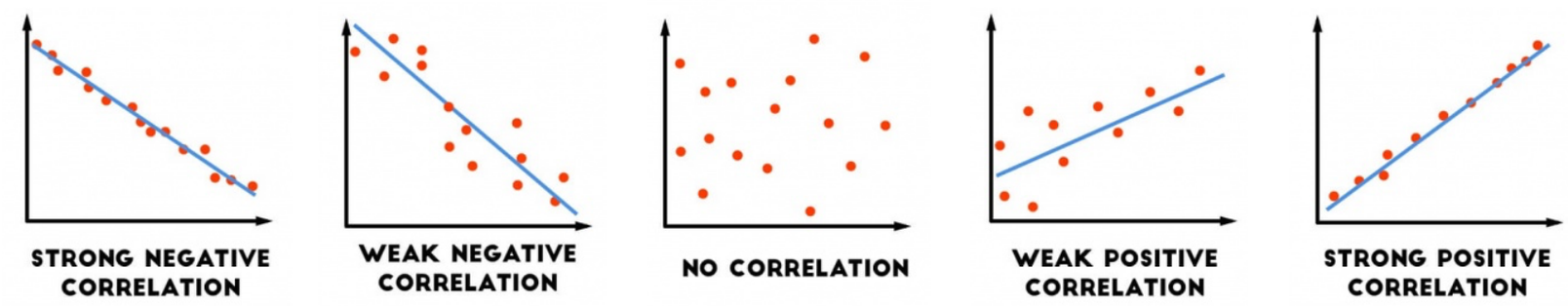
$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n - 1}$$



Correlation Coefficient is used to describe how strong a relationship exists between two variables. Pearson Correlation Coefficient is the most common correlation metric, which measures the linear relationship between two interval or ratio level variables.

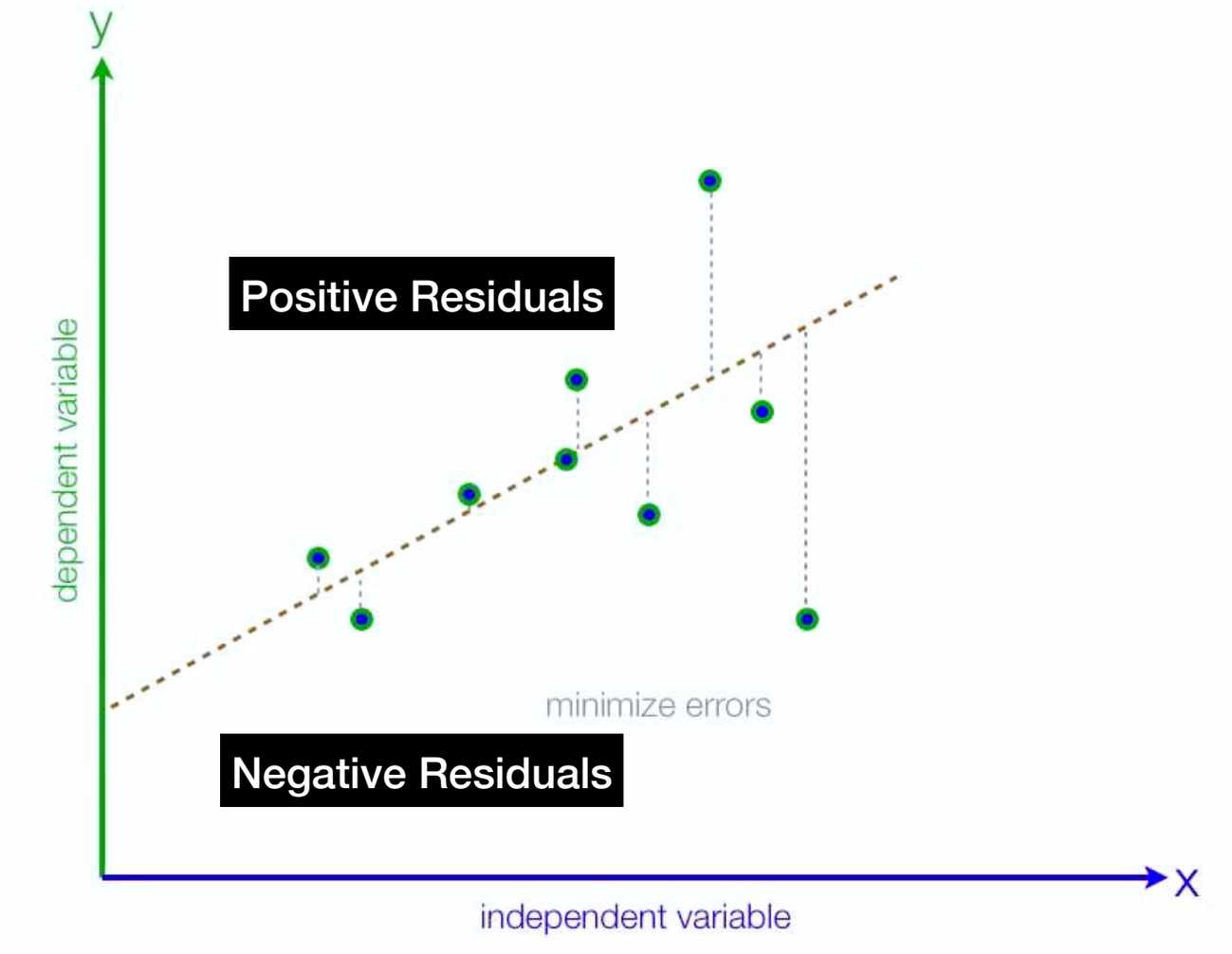
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$



Linear Regression

STATISTICS



Ordinary Least Squares Regression Method

Regression Line is the line with the smallest sum of squared residuals.

$$\hat{y} = a + bx$$

Predicted value of y

Intercept $a = \bar{y} - b\bar{x}$

Regression coefficient $b = r_{xy}\frac{s_y}{s_x}$

$$b = r_{xy}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x s_y}\frac{s_y}{s_x} = \frac{s_{xy}}{{s_x}^2}$$

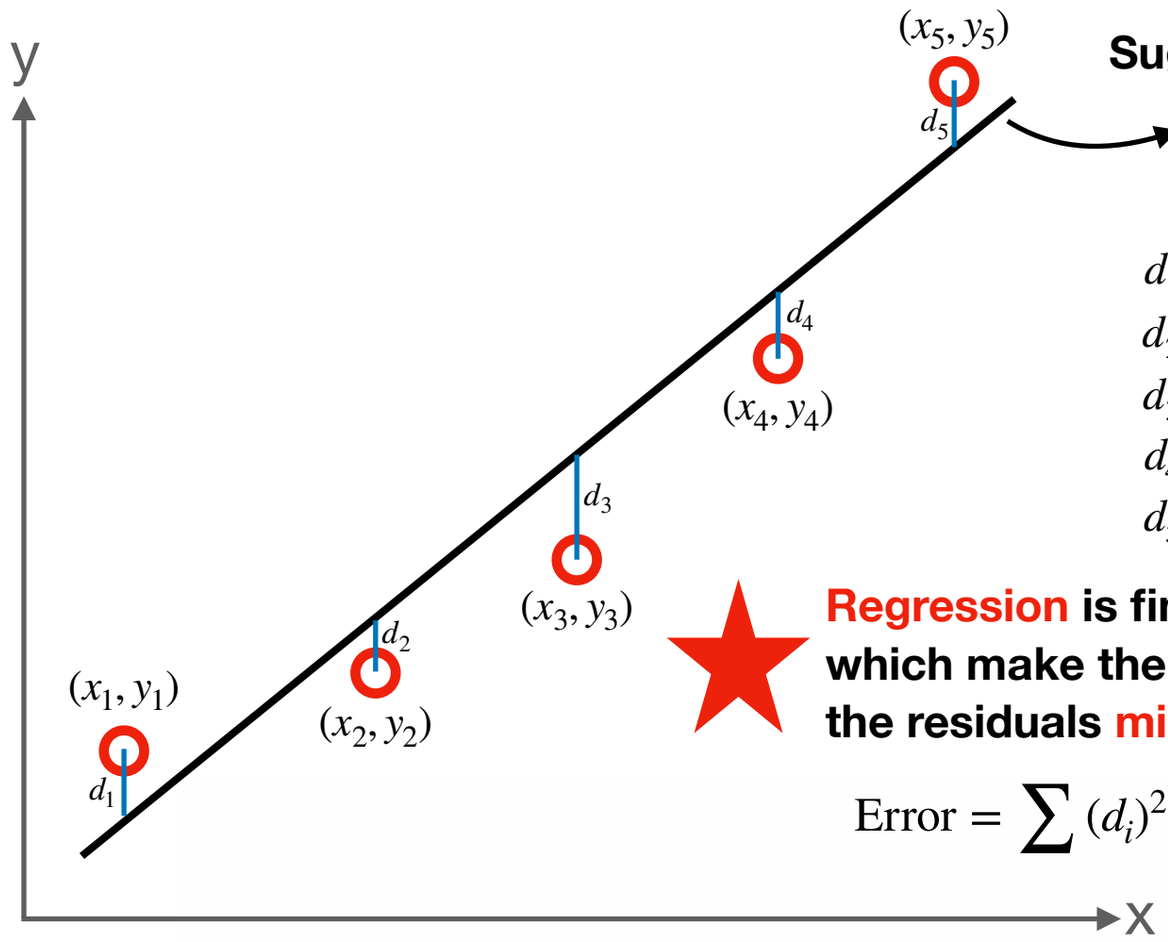
$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{n-1}$$

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

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Linear Regression

STATISTICS



Suggested Equation

$$\hat{y} = b_0 + b_1 x$$

Residuals

$$d_1 = y_1 - (b_0 + b_1 x_1)$$
$$d_2 = y_2 - (b_0 + b_1 x_2)$$
$$d_3 = y_3 - (b_0 + b_1 x_3)$$
$$d_4 = y_4 - (b_0 + b_1 x_4)$$
$$d_5 = y_5 - (b_0 + b_1 x_5)$$

Regression is finding the coefficients, which make the sum of the squares of the residuals minimum.

$$\text{Error} = \sum (d_i)^2 = \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_i) \right]^2$$

$$\frac{\partial \text{Error}}{\partial b_0} = -2 \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_i) \right] = 0$$

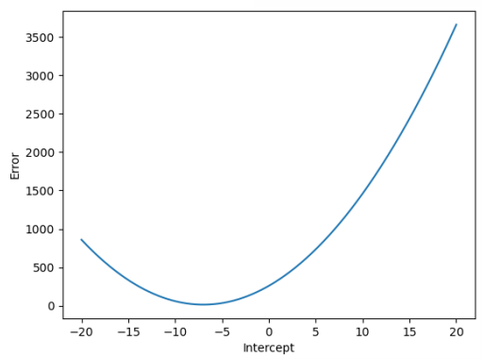
$$\frac{\partial \text{Error}}{\partial b_1} = -2 \sum_{i=1}^{n} x_i \left[y_i - (b_0 + b_1 x_i) \right] = 0$$

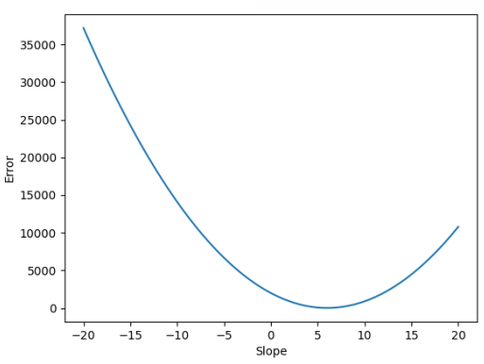
$$b_0 n + b_1 \sum x_i = \sum y_i$$

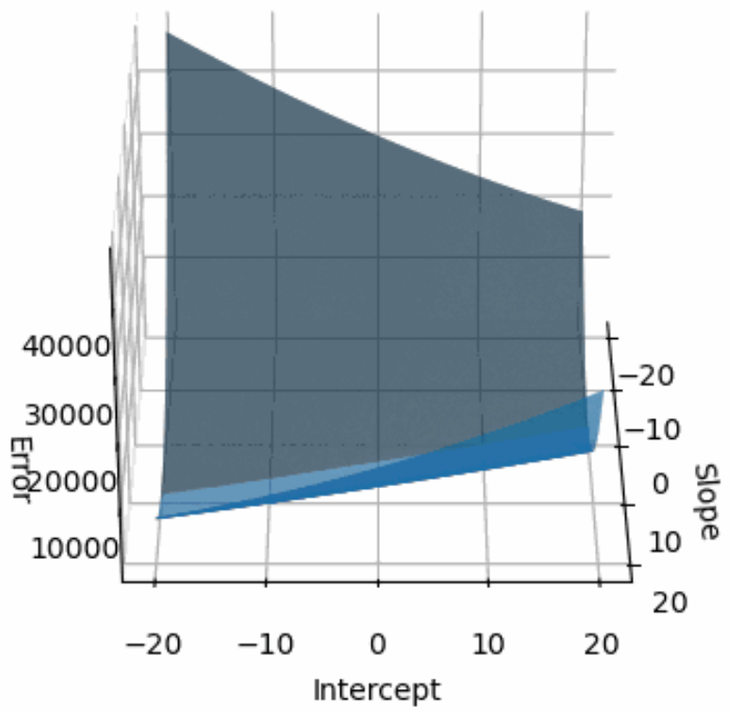
$$b_0 \sum x_i + b_1 \sum x_i^2 = \sum (x_i y_i)$$

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum (x_i y_i) \end{bmatrix}$$

$$\begin{bmatrix} \sum x_i^0 & \sum x_i^1 \\ \sum x_i^1 & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum (x_i^0 y_i) \\ \sum (x_i^1 y_i) \end{bmatrix}$$







$$\begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 55 \\ 225 \end{bmatrix}$$

$$-3 \;/\; 5b_0 + 15b_1 = 55$$
$$15b_0 + 55b_1 = 225$$

$$-15b_0 - 45b_1 = -165$$
$$+ \quad 15b_0 + 55b_1 = 225$$
$$10b_1 = 60$$

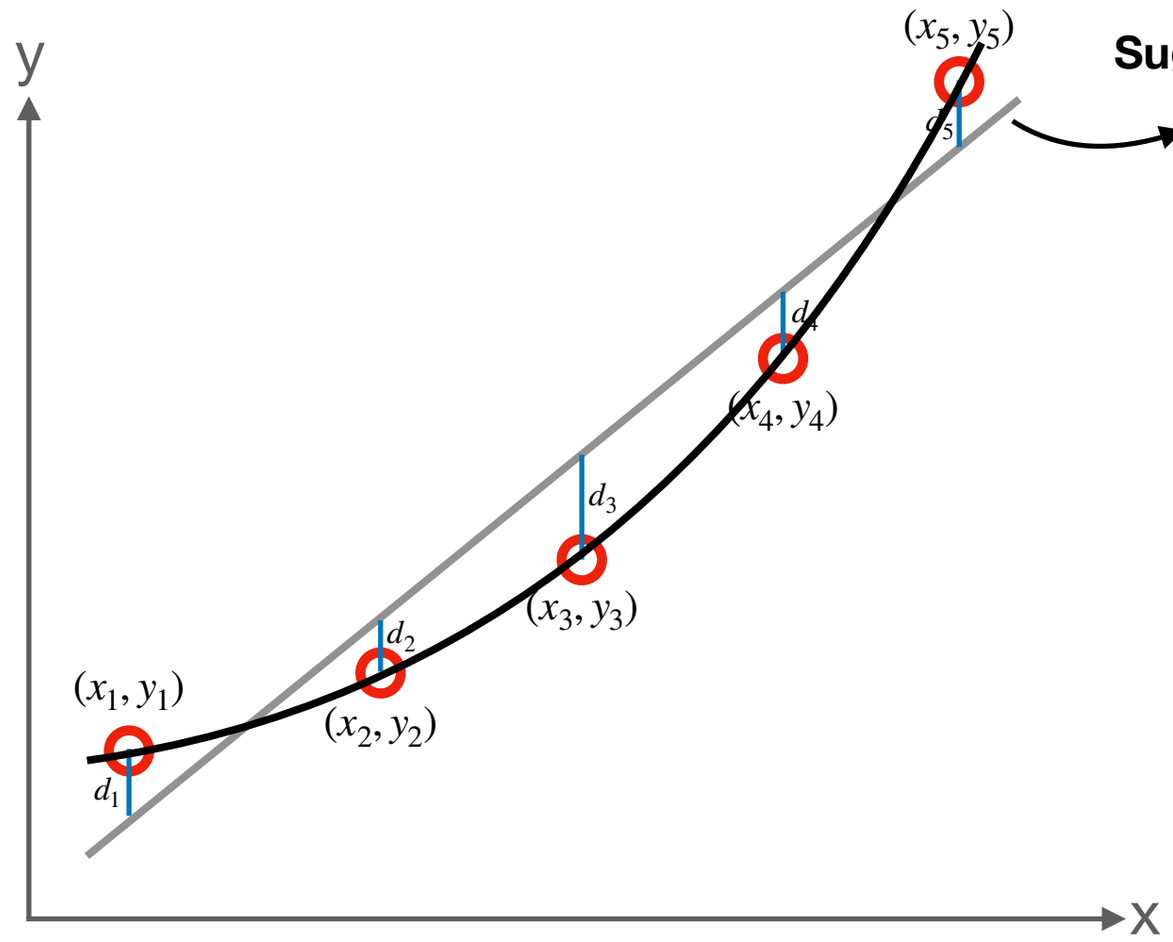
$$b_1 = 6$$
$$b_0 = -7$$

$$\hat{y} = -7 + 6x$$

x	y	x^2	x y
1	1	1	1
2	4	4	8
3	9	9	27
4	16	16	64
5	25	25	125
15	**55**	**55**	**225**

Linear Regression

STATISTICS



Suggested Equation

~~$\hat{y} = b_0 + b_1 x$~~

$$\hat{y} = b_0 + b_1 x + b_2 x^2$$

nth Order Polynomial

$$\hat{y} = \sum_{k=0}^{j} b_k x^k$$

$$\text{Error} = \sum (d_i)^2 = \sum_{i=1}^{n} \left[y_i - \left(\sum_{k=0}^{j} b_k x^k \right) \right]^2$$

$$\frac{\partial \text{Error}}{\partial b_j} = -2 \sum_{i=1}^{n} x_i \left[y_i - \left(\sum_{k=0}^{j} b_k x^k \right) \right] x^j = 0$$

x	y	x^2	x^3	x^4	x y	x^2 y
1	1	1	1	1	1	1
2	4	4	8	16	8	16
3	9	9	27	81	27	81
4	16	16	64	256	64	256
5	25	25	125	625	125	625
15	**55**	**55**	**225**	**979**	**225**	**979**

$$\begin{bmatrix} \sum x_i^0 & \sum x_i^1 & \sum x_i^2 & \cdots & \sum x_i^j \\ \sum x_i^1 & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{j+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \cdots & \sum x_i^{j+2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum x_i^j & \sum x_i^{j+1} & \sum x_i^{j+2} & \cdots & \sum x_i^{j+j} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_j \end{bmatrix} = \begin{bmatrix} \sum (x_i^0 y_i) \\ \sum (x_i^1 y_i) \\ \sum (x_i^2 y_i) \\ \vdots \\ \sum (x_i^j y_i) \end{bmatrix}$$

$$\begin{bmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 55 \\ 225 \\ 979 \end{bmatrix}$$

How to solve for higher orders?

Gaussian Elimination

A method to find the unknowns of a system of linear equations

$$\begin{bmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 55 \\ 225 \\ 979 \end{bmatrix}$$

Start with an Augmented Matrix

$$\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 15 & 55 & 225 & 225 \\ 55 & 225 & 979 & 979 \end{array}\right)$$

The aim is to convert our matrix to an upper triangular matrix

$$\left(\begin{array}{ccc|c} ? & ? & ? & ? \\ 0 & ? & ? & ? \\ 0 & 0 & ? & ? \end{array}\right)$$

Allowed Operations:

- **Swap rows**
- **Add one row to another**
- **Multiply every factor of one row with a constant**

Back Substitution:

- **Starting from the bottom row, solve for each variable**

We want to make this number zero

$$\begin{matrix} R_1 \rightarrow \\ R_2 \rightarrow \\ R_3 \rightarrow \end{matrix}\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 15 & 55 & 225 & 225 \\ 55 & 225 & 979 & 979 \end{array}\right) \qquad R_2' = R_2 - \frac{15}{5} R_1$$

$$\left(\begin{array}{ccc|c} \underline{5} & 15 & 55 & 55 \\ \underline{15} & 55 & 225 & 225 \\ 55 & 225 & 979 & 979 \end{array}\right)$$

If you go downwards flip the numbers vertically

$$R_2' = (15,55,225\,|\,225) - \frac{15}{5}(5,15,55\,|\,55) = (15,55,225\,|\,225) - (15,45,165\,|\,165) = (0,10,60\,|\,60)$$

$$\begin{matrix} R_1 \rightarrow \\ R_2 \rightarrow \\ R_3 \rightarrow \end{matrix}\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 0 & 10 & 60 & 60 \\ 55 & 225 & 979 & 979 \end{array}\right) \qquad R_3' = R_3 - \frac{55}{5} R_1$$

$$\left(\begin{array}{ccc|c} \underline{5} & 15 & 55 & 55 \\ \underline{0} & 10 & 60 & 60 \\ 55 & 225 & 979 & 979 \end{array}\right)$$

If you go downwards flip the numbers vertically

$$R_3' = (55,225,979\,|\,979) - \frac{55}{5}(5,15,55\,|\,55) = (55,225,979\,|\,979) - (55,165,605\,|\,605) = (0,60,374\,|\,374)$$

$$\begin{matrix} R_1 \rightarrow \\ R_2 \rightarrow \\ R_3 \rightarrow \end{matrix}\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 0 & 10 & 60 & 60 \\ 0 & 60 & 374 & 374 \end{array}\right) \qquad R_3' = R_3 - \frac{60}{10} R_2$$

$$\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 0 & \underline{10} & 60 & 60 \\ 0 & \underline{60} & 374 & 374 \end{array}\right)$$

Still going downwards so flip the numbers vertically

$$R_3' = (0,60,374\,|\,374) - \frac{60}{10}(0,10,60\,|\,60) = (0,60,374\,|\,374) - (0,60,360\,|\,360) = (0,0,14\,|\,14)$$

$$\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 0 & 10 & 60 & 60 \\ 0 & 0 & 14 & 14 \end{array}\right)$$

We have an upper triangular matrix now

$$\begin{bmatrix} 5 & 15 & 55 \\ 0 & 10 & 60 \\ 0 & 0 & 14 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 55 \\ 60 \\ 14 \end{bmatrix}$$

$$5b_0 + 15b_1 + 55b_2 = 55 \quad \leftarrow \text{Back substitution} \rightarrow b_0 = 0$$

$$10b_1 + 60b_2 = 60 \quad \leftarrow \text{Back substitution} \rightarrow b_1 = 0$$

$$14b_2 = 14 \rightarrow b_2 = 1$$

Gauss-Jordan Elimination

Gauss-Jordan Elimination

A method to find the unknowns of a system of linear equations

$$\begin{bmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 55 \\ 225 \\ 979 \end{bmatrix}$$

Start with an Augmented Matrix

$$\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 15 & 55 & 225 & 225 \\ 55 & 225 & 979 & 979 \end{array}\right)$$

The aim is to convert our matrix to an identity matrix

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & ? \\ 0 & 1 & 0 & ? \\ 0 & 0 & 1 & ? \end{array}\right)$$

Allowed Operations:

- **Swap rows**
- **Add one row to another**
- **Multiply every factor of one row with a constant**

Normalization:

- **Divide the row with its leading element to make it 1**

Before back substitution

$$\begin{matrix} R_1 \to \\ R_2 \to \\ R_3 \to \end{matrix}\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 0 & 10 & 60 & 60 \\ 0 & 0 & 14 & 14 \end{array}\right) \qquad R_2' = R_2 - \frac{60}{14} R_3$$

If you go upwards don't flip the numbers

$$\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 0 & 10 & 60 & 60 \\ 0 & 0 & 14 & 14 \end{array}\right)$$

$$R_2' = (0,10,60\,|\,60) - \frac{60}{14}(0,0,14\,|\,14) = (0,10,60\,|\,60) - (0,0,60\,|\,60) = (0,10,0\,|\,0)$$

$$\begin{matrix} R_1 \to \\ R_2 \to \\ R_3 \to \end{matrix}\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 14 & 14 \end{array}\right) \qquad R_1' = R_1 - \frac{55}{14} R_3$$

Going upwards don't flip

$$\left(\begin{array}{ccc|c} 5 & 15 & 55 & \\ 0 & 10 & 0 & \\ 0 & 0 & 14 & 14 \end{array}\right)$$

$$R_1' = (5,15,55\,|\,55) - \frac{55}{14}(0,0,14\,|\,14) = (5,15,55\,|\,55) - (0,0,55\,|\,55) = (5,15,0\,|\,0)$$

$$\begin{matrix} R_1 \to \\ R_2 \to \\ R_3 \to \end{matrix}\left(\begin{array}{ccc|c} 5 & 15 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 14 & 14 \end{array}\right) \qquad R_1' = R_1 - \frac{15}{10} R_2$$

Going upwards don't flip

$$\left(\begin{array}{ccc|c} 5 & 15 & & \\ 0 & 10 & & \\ 0 & 0 & 14 & 14 \end{array}\right)$$

$$R_1' = (5,15,0\,|\,0) - \frac{15}{10}(0,10,0\,|\,0) = (5,15,0\,|\,0) - (0,15,0\,|\,0) = (5,0,0\,|\,0)$$

Normalization

$$\begin{matrix} R_1 \to \\ R_2 \to \\ R_3 \to \end{matrix}\left(\begin{array}{ccc|c} 5 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 14 & 14 \end{array}\right) \Rightarrow \begin{matrix} R_1' = \frac{R_1}{5} \\ R_2' = \frac{R_2}{10} \\ R_3' = \frac{R_3}{14} \end{matrix} \Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \Rightarrow \begin{matrix} b_0 = 0 \\ b_1 = 0 \\ b_2 = 1 \end{matrix}$$

Gauss-Jordan Elimination

A method to find the unknowns of a system of linear equations

$$\begin{bmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 55 \\ 225 \\ 979 \end{bmatrix}$$

Start with an Augmented Matrix

$$\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 15 & 55 & 225 & 225 \\ 55 & 225 & 979 & 979 \end{array}\right)$$

The aim is to convert our matrix to an identity matrix

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & ? \\ 0 & 1 & 0 & ? \\ 0 & 0 & 1 & ? \end{array}\right)$$

Allowed Operations:

- Swap rows
- Add one row to another
- Multiply every factor of one row with a constant

Normalization:

- Divide the row with its leading element to make it 1

Before back substitution

$$\begin{matrix} R_1 \to \\ R_2 \to \\ R_3 \to \end{matrix}\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 0 & 10 & 60 & 60 \\ 0 & 0 & 14 & 14 \end{array}\right)$$

$$R_2' = R_2 - \frac{60}{14} R_3$$

$$\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 0 & 10 & 60 & 60 \\ 0 & 0 & 14 & 14 \end{array}\right)$$

If you go upwards don't flip the numbers

$$R_2' = (0,10,60\,|\,60) - \frac{60}{14}(0,0,14\,|\,14) = (0,10,60\,|\,60) - (0,0,60\,|\,60) = (0,10,0\,|\,0)$$

$$\begin{matrix} R_1 \to \\ R_2 \to \\ R_3 \to \end{matrix}\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 14 & 14 \end{array}\right)$$

$$R_1' = R_1 - \frac{55}{14} R_3$$

$$\left(\begin{array}{ccc|c} 5 & 15 & 55 & 55 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 14 & 14 \end{array}\right)$$

Going upwards don't flip

$$R_1' = (5,15,55\,|\,55) - \frac{55}{14}(0,0,14\,|\,14) = (5,15,55\,|\,55) - (0,0,55\,|\,55) = (5,15,0\,|\,0)$$

$$\begin{matrix} R_1 \to \\ R_2 \to \\ R_3 \to \end{matrix}\left(\begin{array}{ccc|c} 5 & 15 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 14 & 14 \end{array}\right)$$

$$R_1' = (5,15,0\,|\,0) -$$

ng upwards
't flip

14

$$\begin{matrix} R_1 \to \\ R_2 \to \\ R_3 \to \end{matrix}\left(\begin{array}{ccc|c} 5 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 14 & 14 \end{array}\right)$$

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_0 \end{bmatrix} = \begin{bmatrix} 0 \end{bmatrix}$$

$b_0 = 0$

$= 0$

$= 1$

Q An experiment was conducted on a new model of a particular make of automobile to determine the stopping distance at various speeds. The following data were recorded.

Speed, v (km/hr)	35	50	65	80	95	110
Stopping Distance, d (m)	16	26	41	62	88	119

(a) Fit a multiple regression curve of the form $\mu_{D|v} = \beta_0 + \beta_1 v + \beta_2 v^2$.

(b) Estimate the stopping distance when the car is traveling at 70 kilometers per hour.

Q The following data are given:

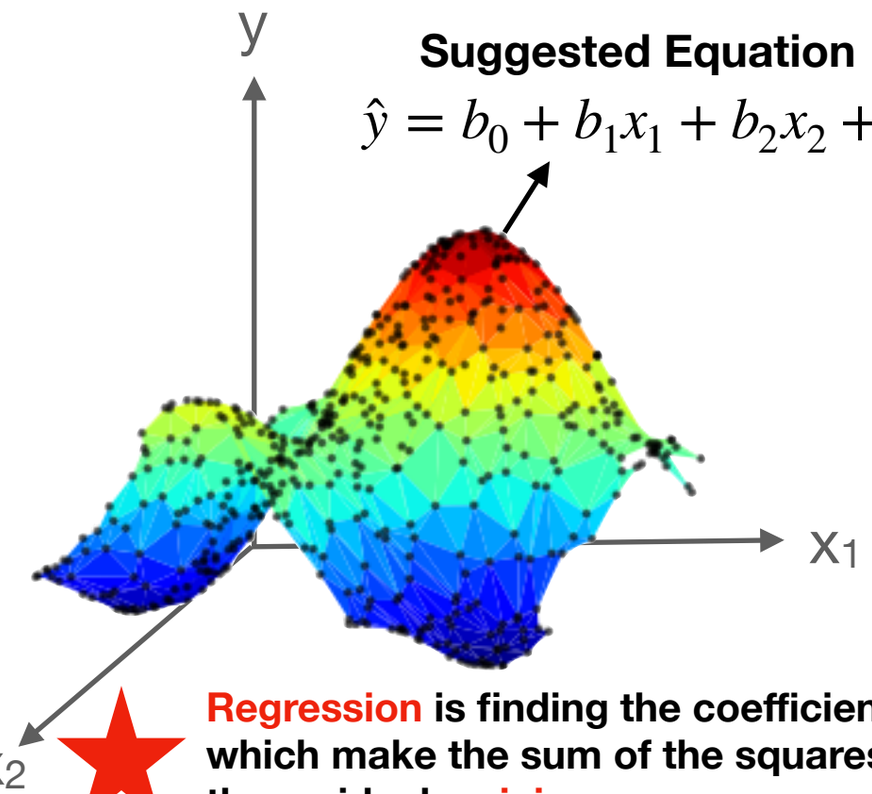
x	0	1	2	3	4	5	6
y	1	4	5	3	2	3	4

(a) Fit the cubic model $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

(b) Predict Y when $x = 2$.

Multiple Linear Regression

STATISTICS



Suggested Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \epsilon$$

$$\begin{aligned} y_1 &= b_0 + b_1 x_{11} + b_2 x_{12} + \epsilon_1 \\ y_2 &= b_0 + b_1 x_{21} + b_2 x_{22} + \epsilon_2 \\ y_3 &= b_0 + b_1 x_{31} + b_2 x_{32} + \epsilon_3 \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ y_n &= b_0 + b_1 x_{n1} + b_2 x_{n2} + \epsilon_n \end{aligned}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\vec{y} = X\vec{b} + \vec{\epsilon}$$

$$\vec{\epsilon} = \vec{y} - X\vec{b}$$

Regression is finding the coefficients, which make the sum of the squares of the residuals minimum.

SSE (Sum of Squared Errors) is a measure of unexplained variation

$$\vec{\epsilon}^T\vec{\epsilon} = (\vec{y} - X\vec{b})^T(\vec{y} - X\vec{b})$$

$$\frac{\partial \text{SSE}}{\partial \vec{b}} = 0 \quad \frac{\partial \vec{\epsilon}^T\vec{\epsilon}}{\partial \vec{b}} = -2X^T\vec{y} + 2X^TX\vec{b} = 0$$

$$X^TX\vec{b} = X^T\vec{y}$$

Normal Equation

$$X^TX = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} = \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 \end{bmatrix}$$

Augmented Matrix

$$\left(\begin{array}{ccc|c} n & \sum x_{i1} & \sum x_{i2} & \sum y_i \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \sum x_{i1}y_i \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 & \sum x_{i2}y_i \end{array}\right)$$

$$X^T\vec{y} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \sum x_{i2}y_i \end{bmatrix}$$

Fit Quality

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

SSR (Regression Sum of Squares) measure of explained variation

SST (Sum of Squares Total) measure of total variation in y

$R^2 > 0.5$ **Regression is better**

$R^2 \leq 0.5$ **Mean is better**

Multiple Linear Regression Exercises STATISTICS

Data

#1173852269
HONDA EURO CİVİC 1,6 KLİMALI
250,000 TL
Year: 1999
Km: 254,000
Color: Green
İlan Tarihi: 19 May 2024
District / Quarter: Merkez / Muradiye

#1168469220
DEĞİŞENSİZ TEMİZ BAKIMLI KAPUT TAVAN BAGAJ BOYASIZ
457,500 TL
Year: 1999
Km: 280,000
Color: Silver Gray
İlan Tarihi: 19 May 2024
District / Quarter: Merkez / Mesir Mah.

#1173783541
OTOMATİK VİTES Sahibinden 1.6 Euro Civic 163 binde orjinal
365,000 TL
Year: 1996
Km: 163,000
Color: Maroon
İlan Tarihi: 18 May 2024
District / Quarter: Merkez / Akgedik Mh.

#1173777797
2021 Honda Civic Eco Elegans Fabrikasyon LPG'li
1,300,000 TL
Year: 2021
Km: 43,500
Color: White
İlan Tarihi: 18 May 2024
District / Quarter: Merkez / Topçuasım Mh.

#1161525478
Honda civic beyaz inci 777 değişensiz
370,000 TL
Year: 1998
Km: 346,000
Color: White
İlan Tarihi: 18 May 2024
District / Quarter: Merkez / Barbaros Mah.

#1173628834
ACİL EV ALACAĞIM İÇİN SATILIK HONDA FD6
610,000 TL
Year: 2007
Km: 261,000
Color: Beige
İlan Tarihi: 17 May 2024
District / Quarter: Merkez / Güzelyurt Mh.

#1173627183
DOKTORDAN satılık HASARSIZ,MASRAFSIZ
1,200,000 TL
Year: 2018
Km: 110,000
Color: White
İlan Tarihi: 17 May 2024
District / Quarter: Merkez / Uncubozköy Mh.

#1170180952
Değişensiz civic
495,000 TL
Year: 2000
Km: 298,000
Color: Navy Blue
İlan Tarihi: 17 May 2024
District / Quarter: Merkez / Topçuasım Mh.

Cleaned Data

1999	254000	250000
1996	163000	365000
1998	346000	370000
2018	110000	1200000
1999	280000	457500
2021	43500	1300000
2007	261000	610000
2000	298000	495000

Model

$$\text{Price} = b_0 + b_1\sqrt{2024 - \text{Year}} + b_2\frac{\text{KM}}{100000}$$

Q What are the values of the regression coefficients?

$$b_0 = 1832091.85$$
$$b_1 = -281667.55$$
$$b_2 = -8365.33$$

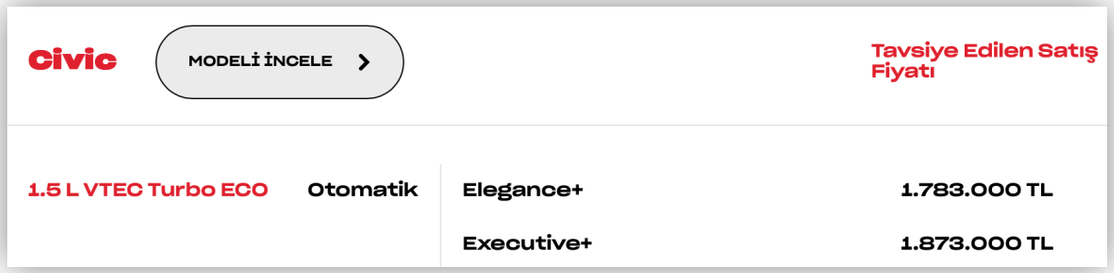
Q What is the predictive power of this model?

$$R^2 = 0.96$$

Q Which is more predictive?

Model or **Mean**

Q How much should a brand new Honda Civic cost?



Multiple Linear Regression Exercises STATISTICS

Data

	**Toshiba 65UA3E63DT** 20278193 LED / Android TV Karşılaştır	**20.495,00 TL** 2 site, 3 fiyat	65 İnç	Ultra HD (4K)	24 W
	**Toshiba 55UA3E63DT** 20278192 LED / Android TV Karşılaştır	**14.699,00 TL** 2 site, 3 fiyat	55 İnç	Ultra HD (4K)	20 W
	**Toshiba 65UL3363DT** 20278134 LED Karşılaştır	**19.999,00 TL** 1 site, 1 fiyat	65 İnç	Ultra HD (4K)	24 W
	**Toshiba 50UA3E63DT** 20278191 LED / Android TV Karşılaştır	**12.999,00 TL** 2 site, 2 fiyat	50 İnç	Ultra HD (4K)	20 W
	**Toshiba 55UL3363DT** 20278135 LED Karşılaştır	**13.899,00 TL** 1 site, 1 fiyat	55 İnç	Ultra HD (4K)	20 W
	**Toshiba 43UL3363DT** 20278132 LED Karşılaştır	**10.999,00 TL** 1 site, 1 fiyat	43 İnç	Ultra HD (4K)	16 W
	**Toshiba 43LA2363DT** 20278404 LED / Android TV Karşılaştır	**10.799,00 TL** 3 site, 3 fiyat	43 İnç	Full HD (FHD)	16 W

Cleaned Data

65	55	65	50	55	43	43
20495	14699	19999	12999	13899	10999	10799

Model A

$$\text{Price} = b_0 \cdot \text{Size}^{b_1}$$

Model B

$$\text{Price} = b_0 \cdot e^{b_1 \cdot \text{Size}}$$

Q Find the regression coefficients of these two models.

Q Compare the predictive power of these two models.

❤ **The end… see you in another course!**