BM5702 MAKİNE ÖĞRENMESİNE GİRİŞ

Hafta 4

Doç. Dr. Murtaza CİCİOĞLU

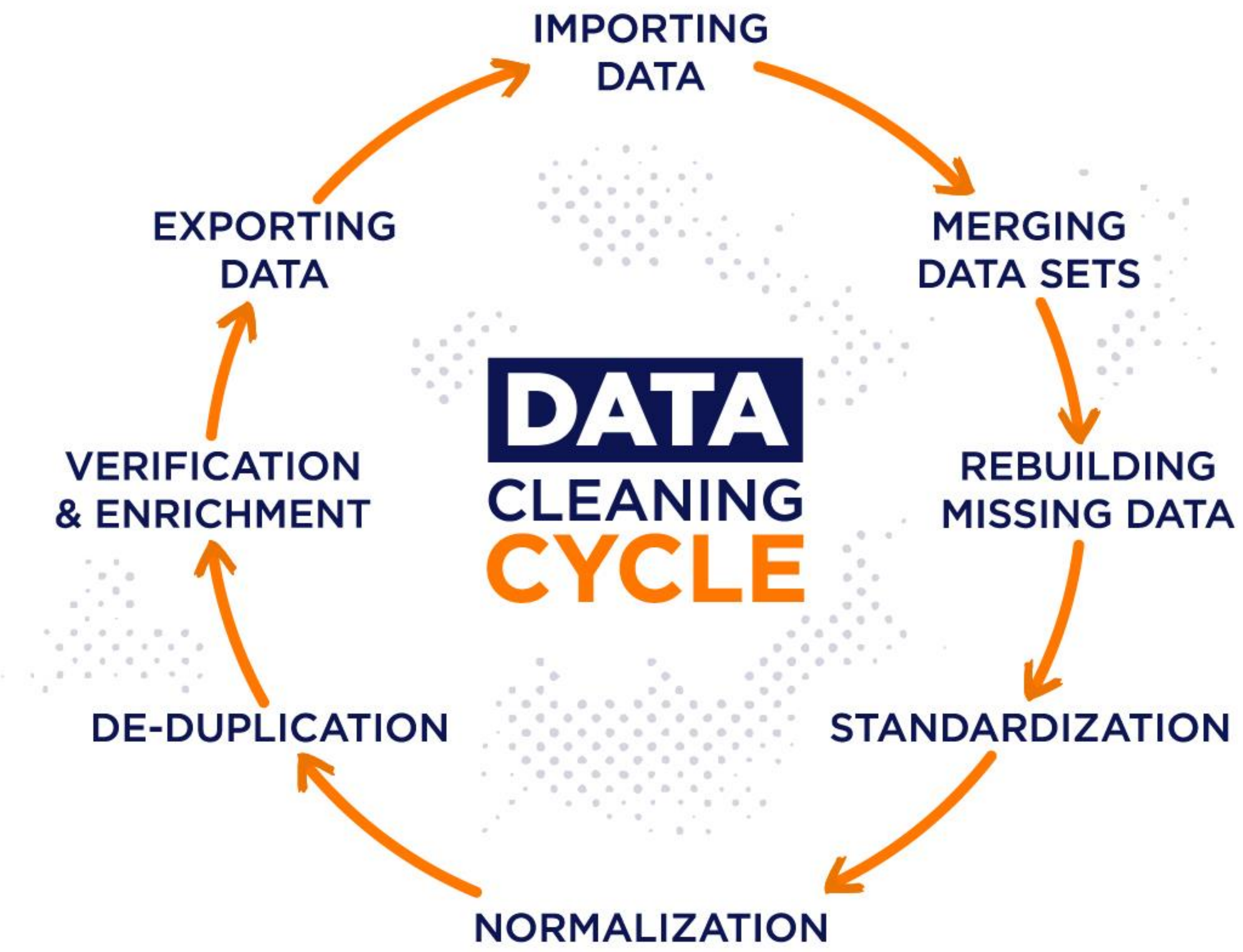
How to Preprocess Data

- Data does not always come in forms ready for analysis.
- It could, for example, be in the wrong format, incorrect or even missing.
- Industry experience has shown that data scientists can spend as much as 75% of their time preparing data before they begin their studies.
- Preparing data for analysis is called data munging or data wrangling.
- data cleaning and transforming data into the optimal formats for your database systems and analytics software.

How to Preprocess Data

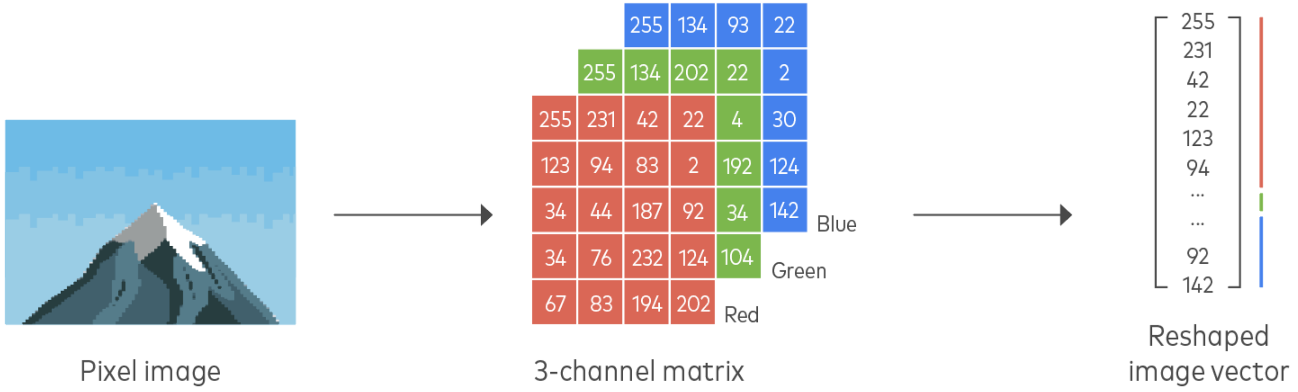
- Some common data cleaning examples are:
  - deleting observations with missing values,
  - substituting reasonable values for missing values,
  - deleting observations with bad values,
  - substituting reasonable values for bad values,
  - tossing outliers (although sometimes you'll want to keep them),
  - duplicate elimination (although sometimes duplicates are valid),
  - dealing with inconsistent data
  - and more.

How to Preprocess Data



How to Preprocess Data

- Some common data transformations include:
  - removing unnecessary data and features
  - combining related features,
  - sampling data to obtain a representative subset
  - standardizing data formats,
  - grouping data,
  - and more



Cleaning Your Data

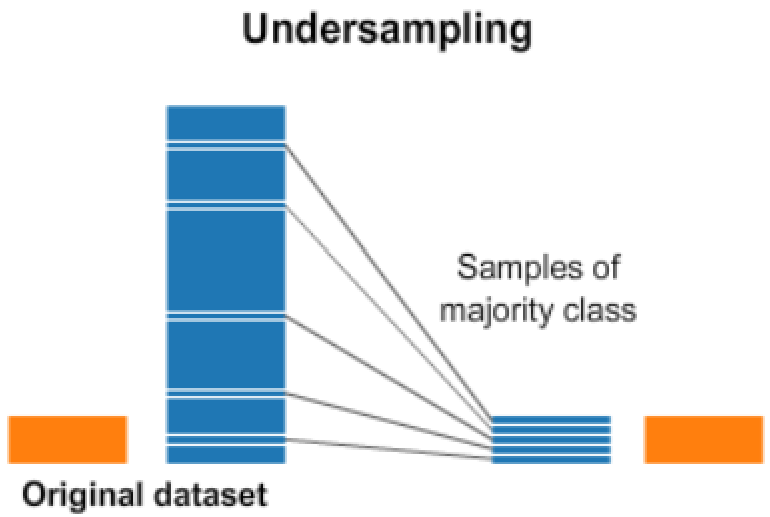
- Bad data values and missing values can significantly impact data analysis.

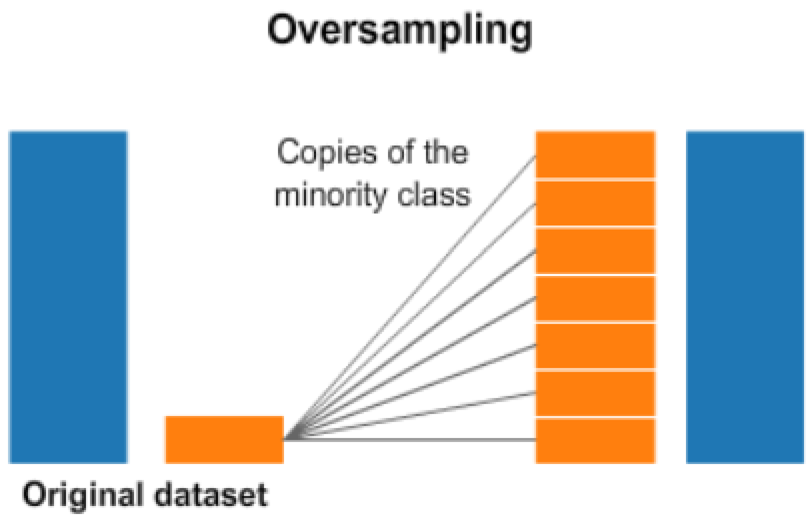
```
['Brown, Sue', 98.6, 98.4, 98.7, 0.0]
```

- The average of the first three values is 98.57
- The average is only 73.93
- Substituting reasonable values' does not mean students should feel free to change values to get the results they want.

Pre-Processing

- **Duplicate Values**
- duplicate values are removed so as to not give that particular data object an advantage or bias
- **Imbalanced Data**
- An Imbalanced dataset is one where the number of instances of a classes are significantly higher than another classes





Pre-Processing

- Missing Values
  - Eleminate missing values
  - Filling with mean, mode or median
- isnull()
- notnull()
- dropna()
- fillna()
- replace()
- interpolate()

Categorical Values

- **One-hot encoding**

id	color
1	red
2	blue
3	green
4	blue

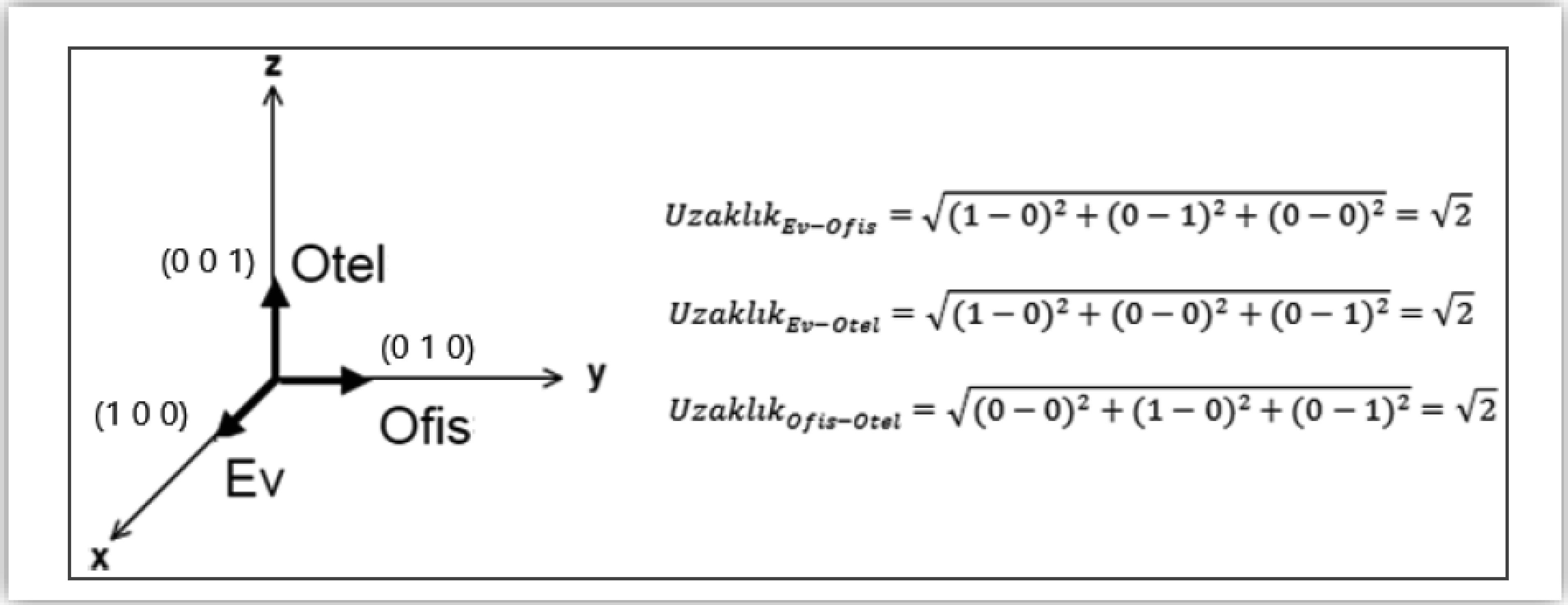
One Hot Encoding →

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

```
df_encoded = pd.get_dummies(df["color"])
df_encoded .head()
```

Categorical Values

- **One-hot encoding**




```
df_encoded = pd.get_dummies(df["color"])
df_encoded .head()
```

Categorical Values

```
df[" Team "] = df["Team"].astype('category')
df[" Team "] = df["Team "].cat.codes
```

- **Label encoding**

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

→

Label Encoded Data

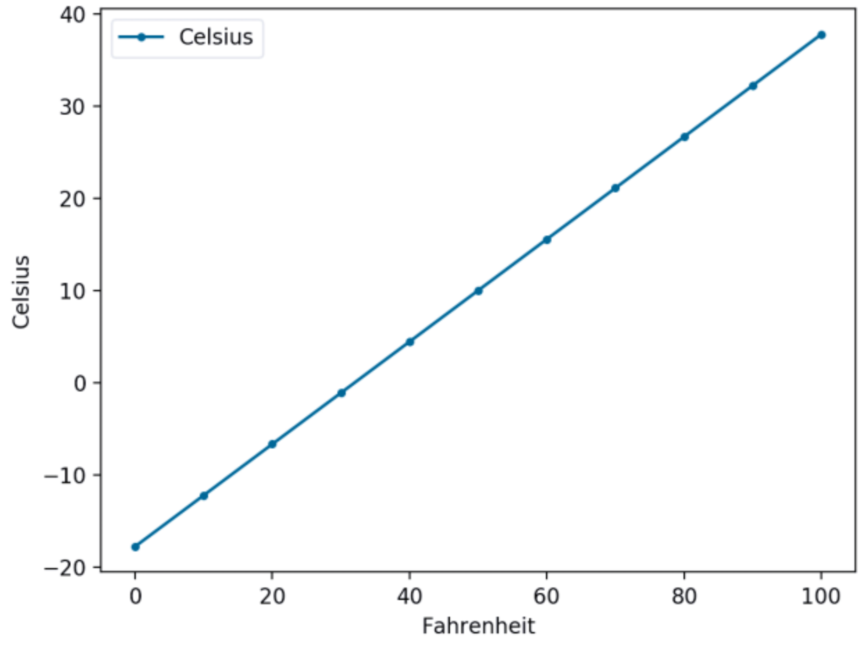
Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

Categorical Values

- One-hot encoding is appropriate when the categories do not have an intrinsic ordering or relationship with each other. This is because one-hot encoding treats each category as a separate entity with no relation to the other categories.
- One-hot encoding is also useful when the number of categories is relatively small, as the number of columns can become unwieldy for very large numbers of categories.
- Label encoding is appropriate when the categories have a natural ordering or relationship with each other, such as in the case of ordinal variables like "small," "medium," and "large."

Simple Linear Regression

- Regression searches for relationships among variables.
- independent and dependent variable,
- regression line

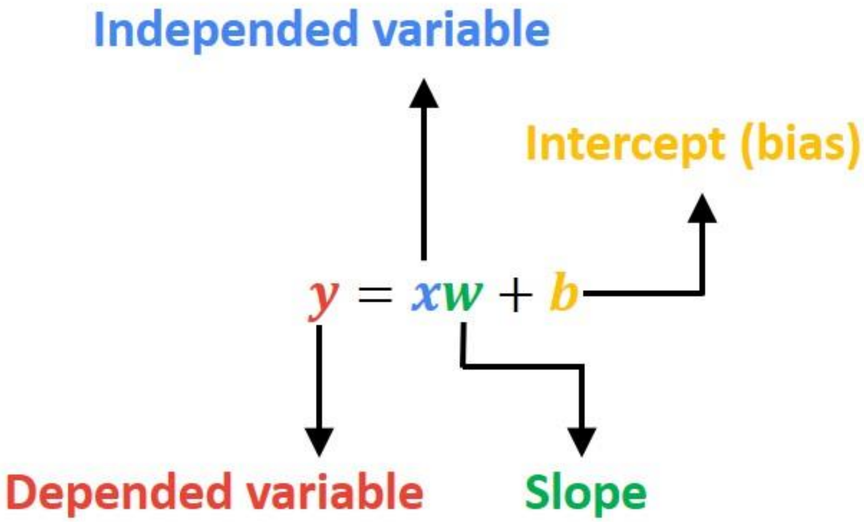


$$y = mx + b$$

where

- m is the line's **slope**,
- b is the line's **intercept** with the y-axis (at $x = 0$),
- x is the independent variable (the date in this example), and
- y is the dependent variable (the temperature in this example).

In simple linear regression, y is the *predicted value* for a given x.



Simple Linear Regression

Temel amaç, bağımlı ve bağımsız değişken arasındaki ilişkiyi ifade eden doğrusal fonksiyonu bulmaktır.

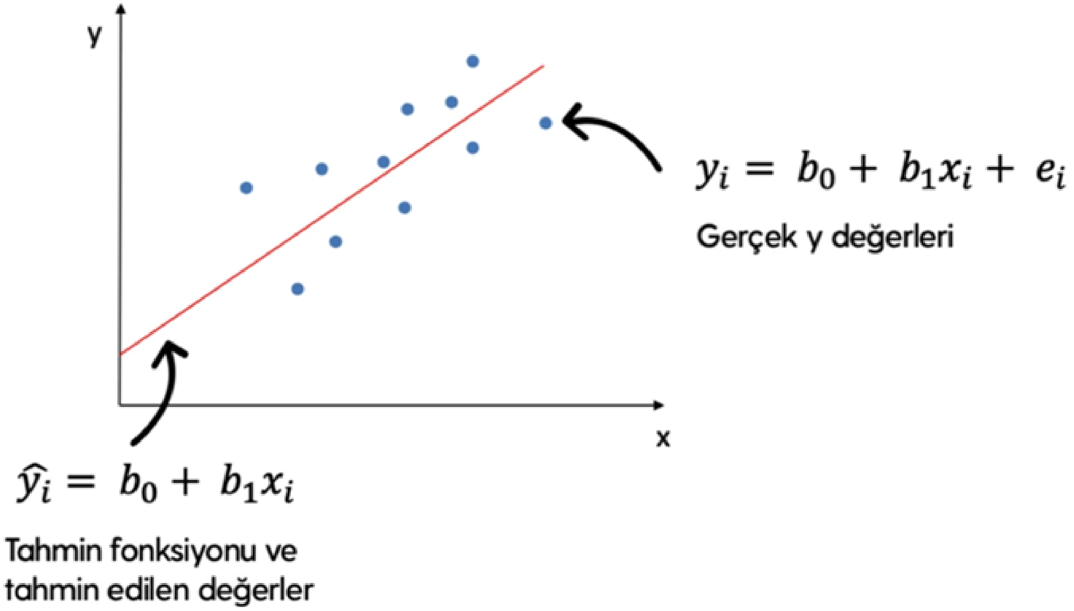
$$\boxed{\hat{y}_i} = b_0 + b_1 x_i$$

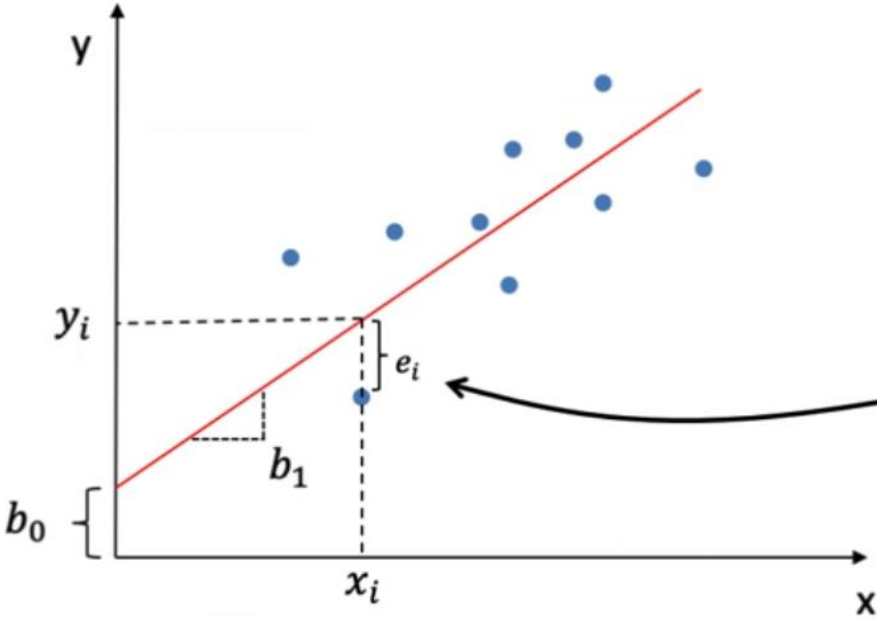
Tahmin edilen değerleri ifade eder.

Bağımsız değişken değerleri ifade eder.

Veri seti içerisinden bulunması gereken parametrelerdir.

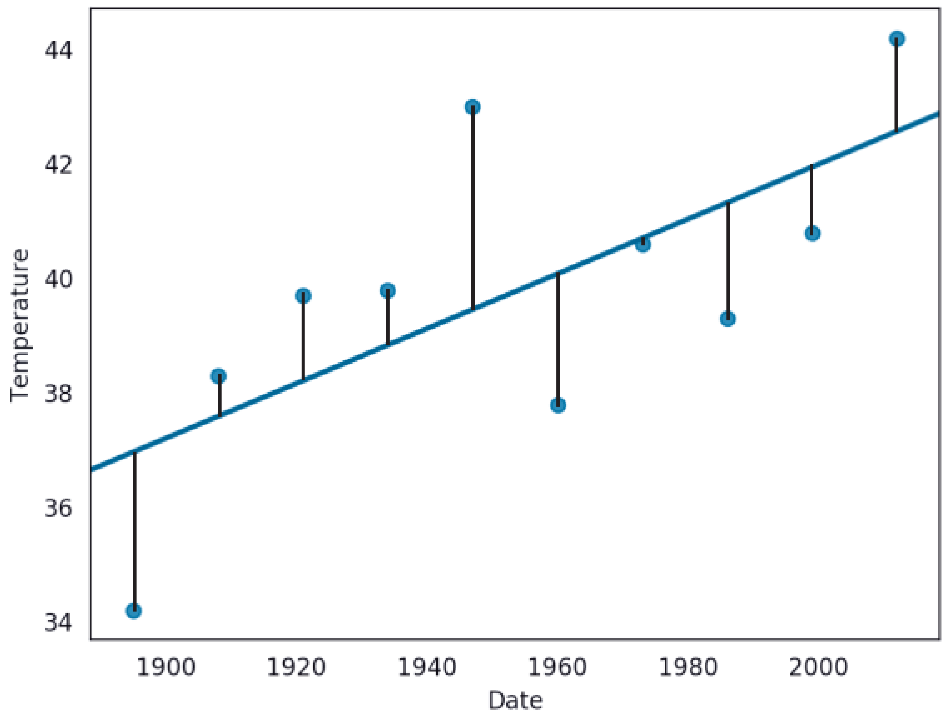
Simple Linear Regression





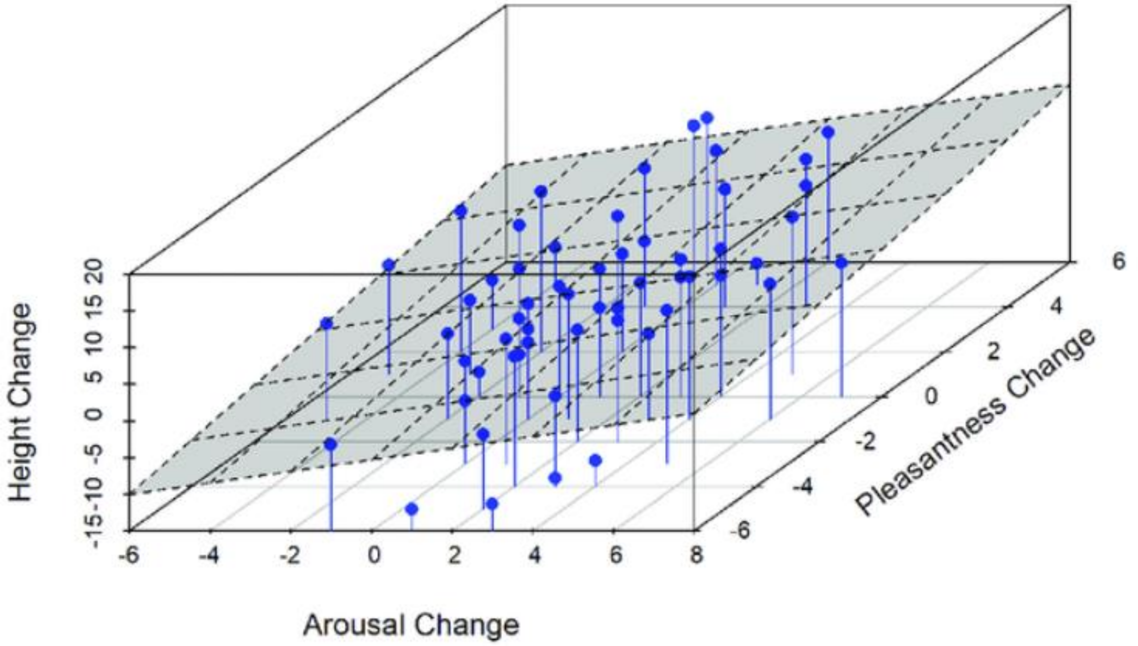
Simple Linear Regression

- When implementing linear regression of some dependent variable y on the set of independent variables $\mathbf{x} = (x_1, \ldots, x_r)$, where r is the number of predictors, you assume a linear relationship between y
- and $\mathbf{x}$: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon$.
- This equation is the regression equation. $\beta_0, \beta_1, \ldots, \beta_r$ are the regression coefficients, and ε is the random error.



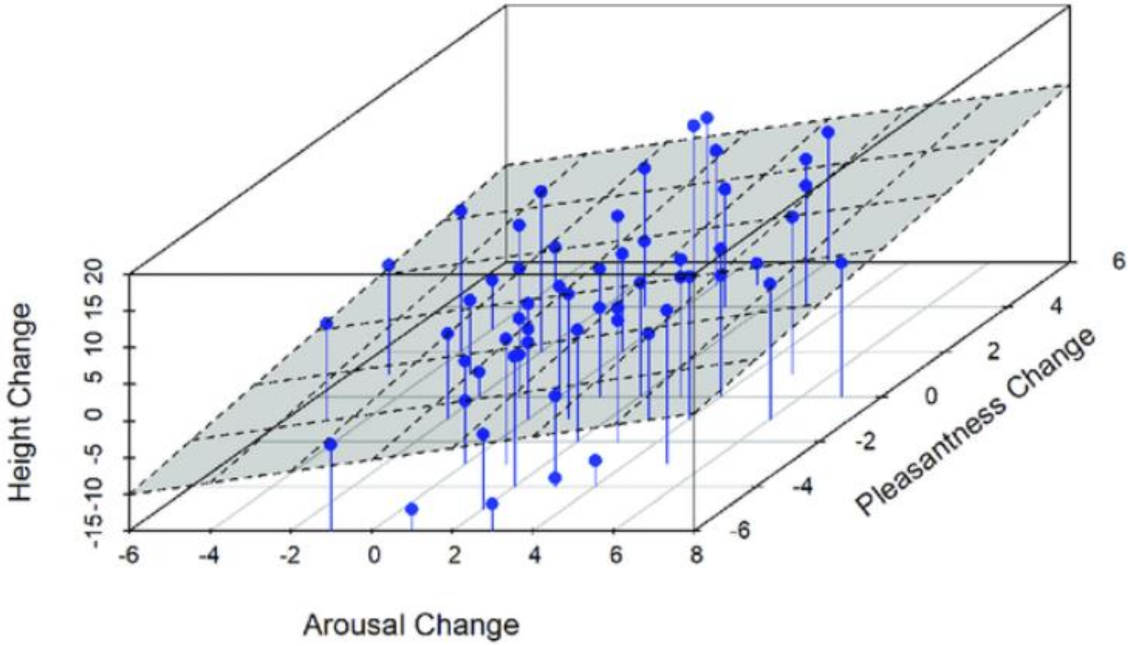
Multiple Linear Regression

- Multiple or multivariate linear regression is a case of linear regression with two or more independent variables.
- If there are just two independent variables, the estimated regression function is:
- $f(x1,x2)=b0+b1x1+b2x2$
- It represents a regression plane in a three-dimensional space. The goal of regression is to determine the values of the weights b_0, b_1, and b_2 such that this plane is as close as possible to the actual responses and yield the minimal SSR.

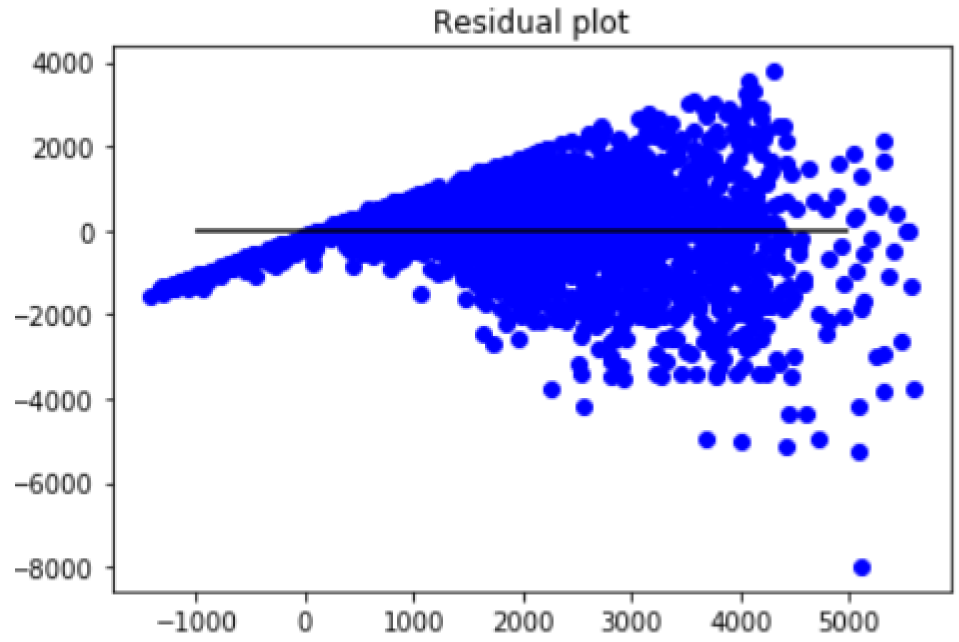


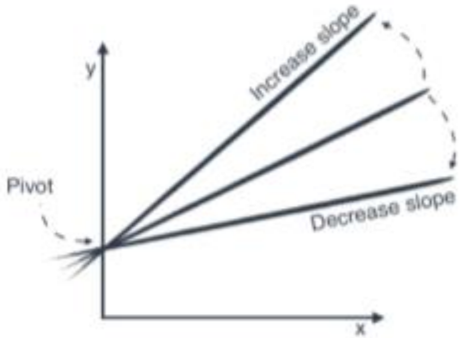
Polynomial Regression

- in addition to linear terms like b_1x_1, your regression function f can include non-linear terms such as $b_2x_1^2$, $b_3x_1^3$, or even $b_4x_1x_2$, $b_5x_1^2x_2$, and so on.
- The simplest example of polynomial regression has a single independent variable, and the estimated regression function is a polynomial of degree 2:
- $f(x)=b_o+b_1x+b_2x^2$

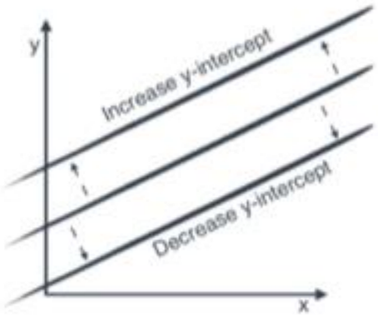


Slope & Intercept

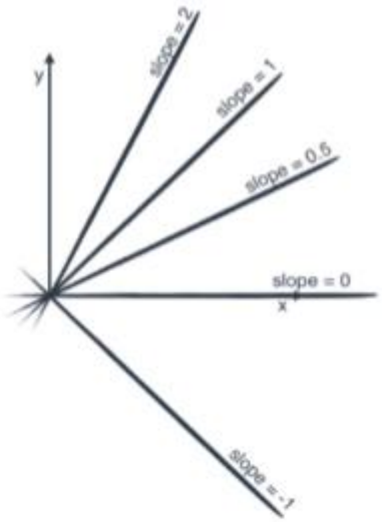


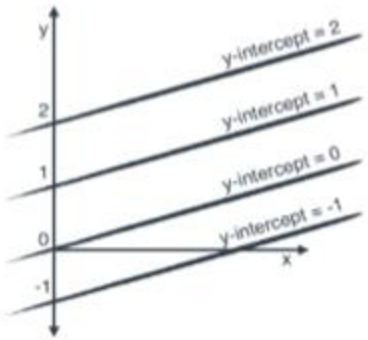


Rotate clockwise and counterclockwise

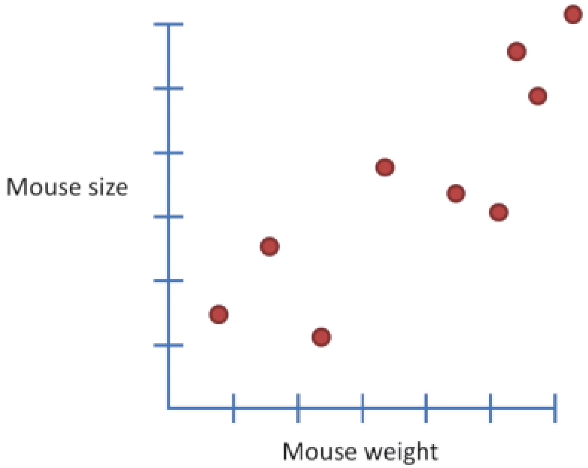


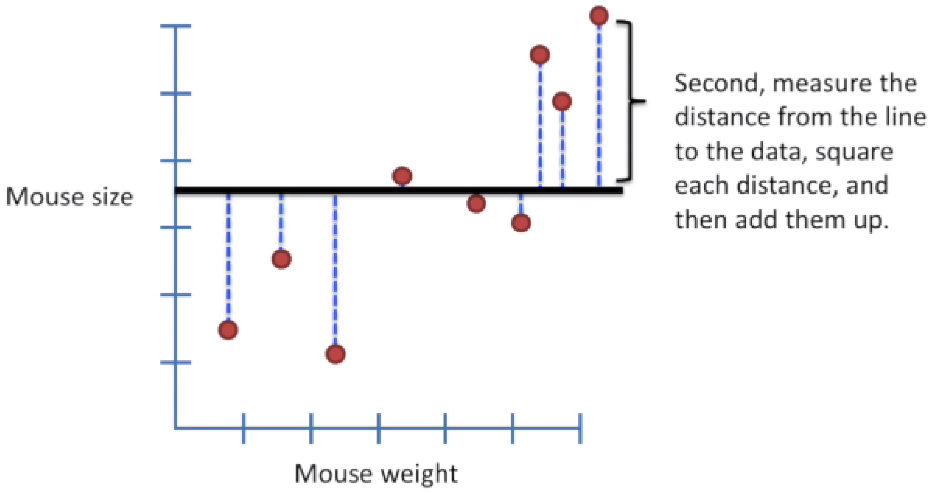
Translate up and down

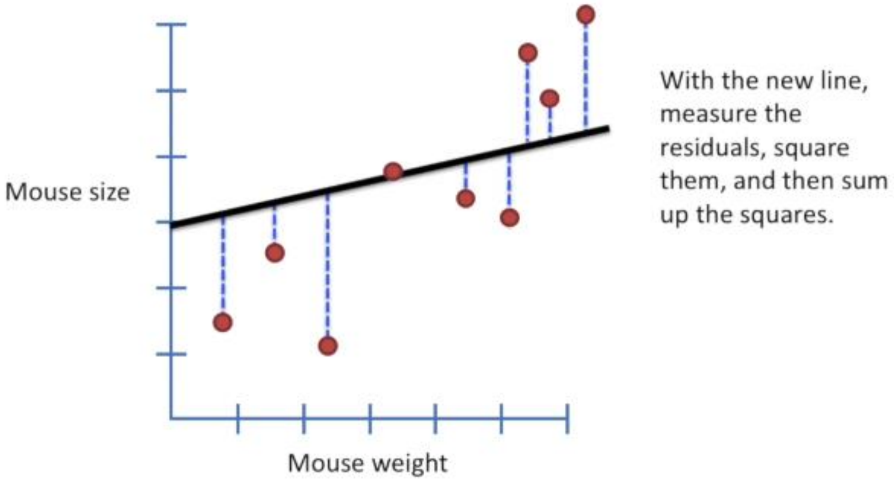


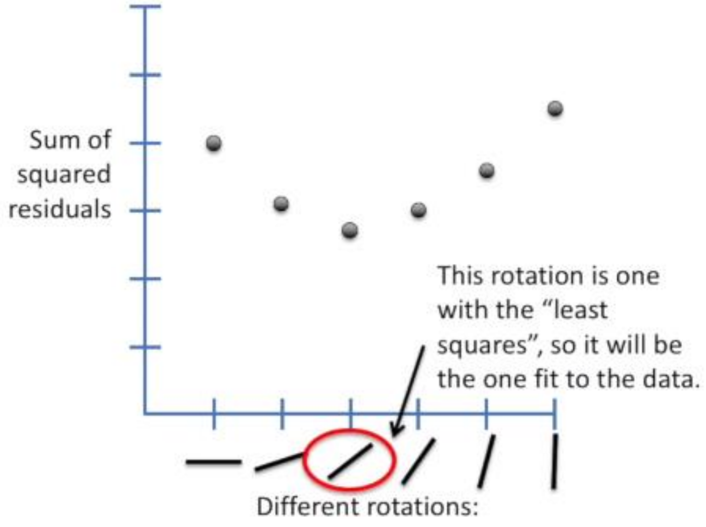


Slope & Intercept

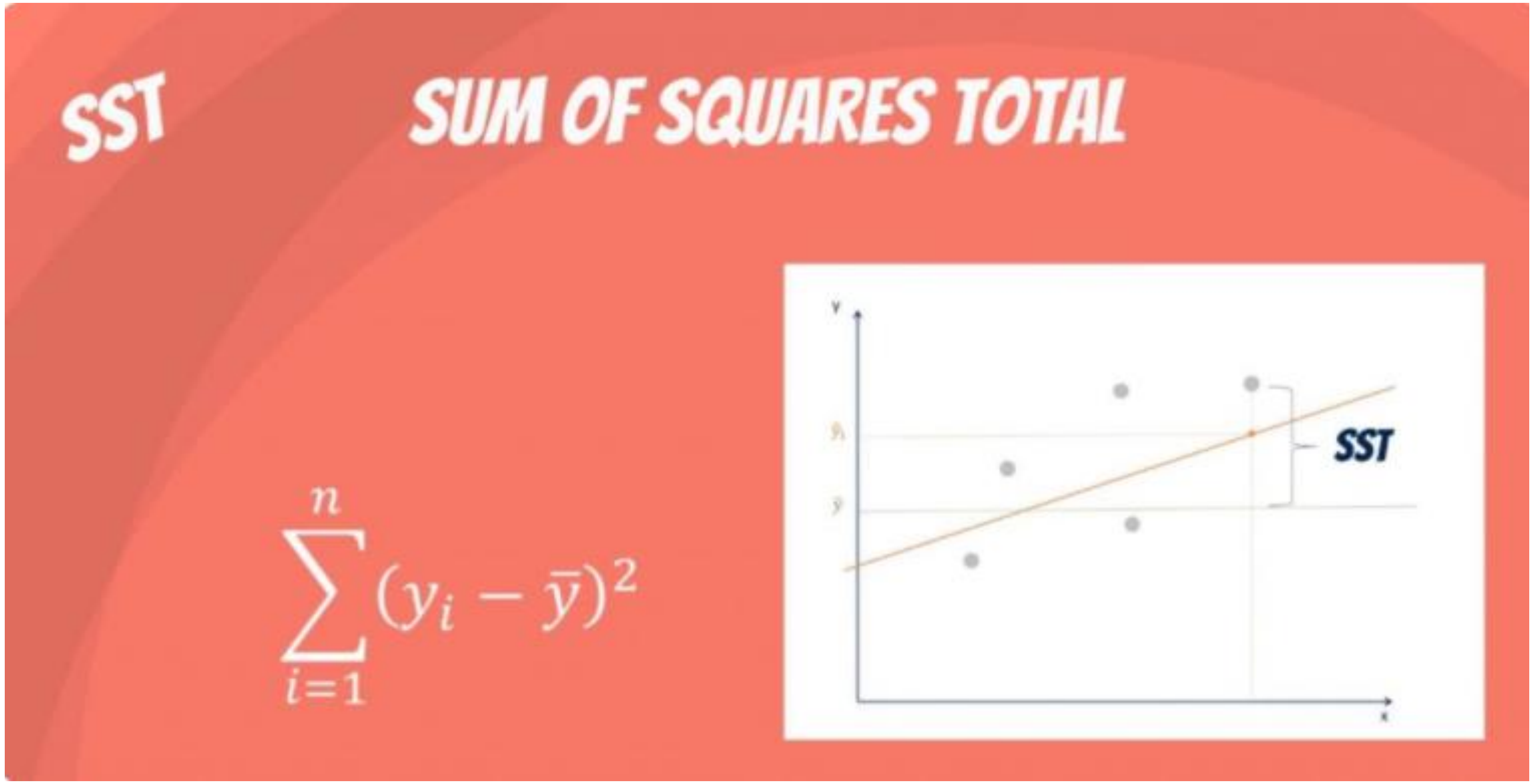




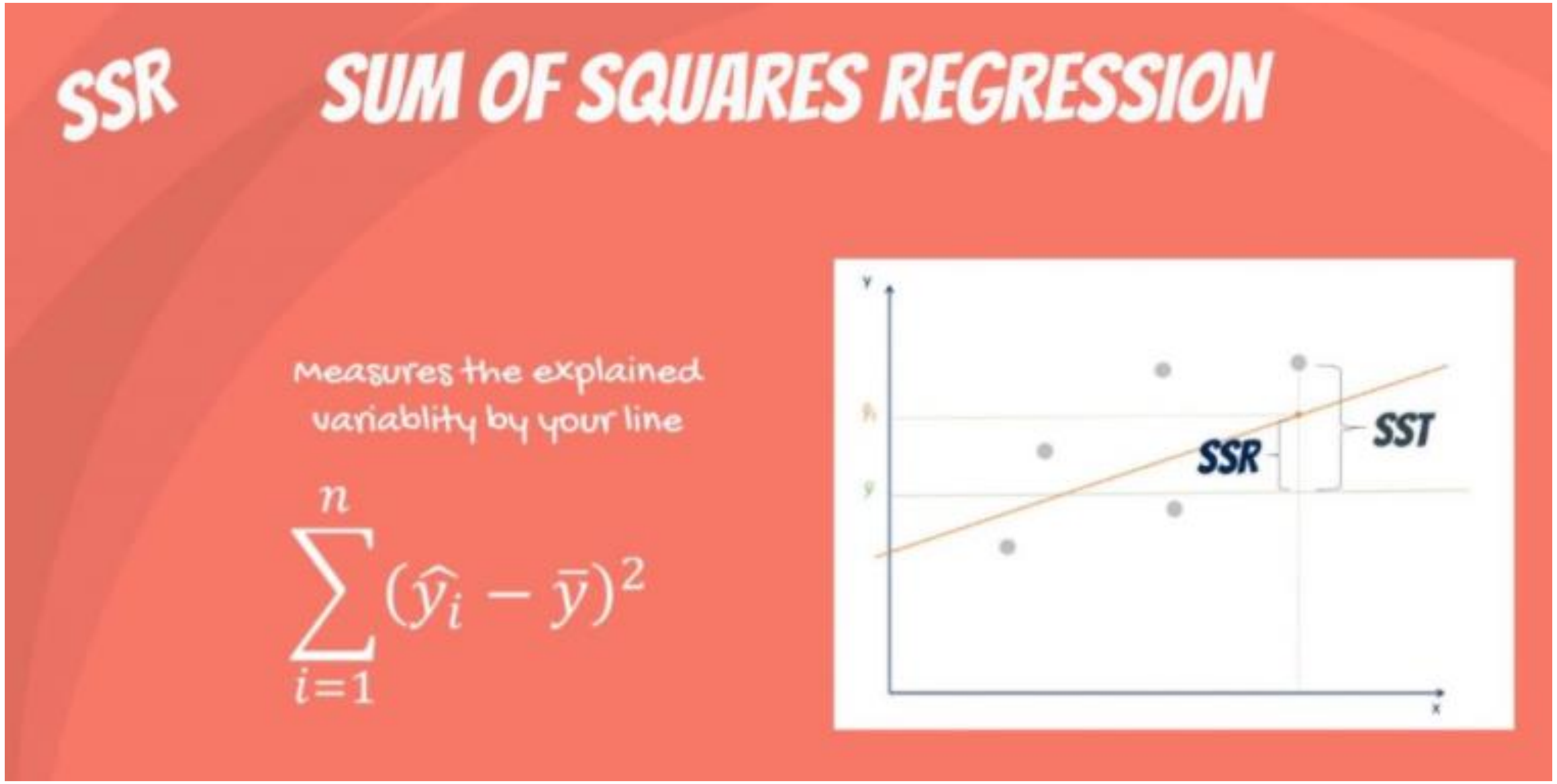




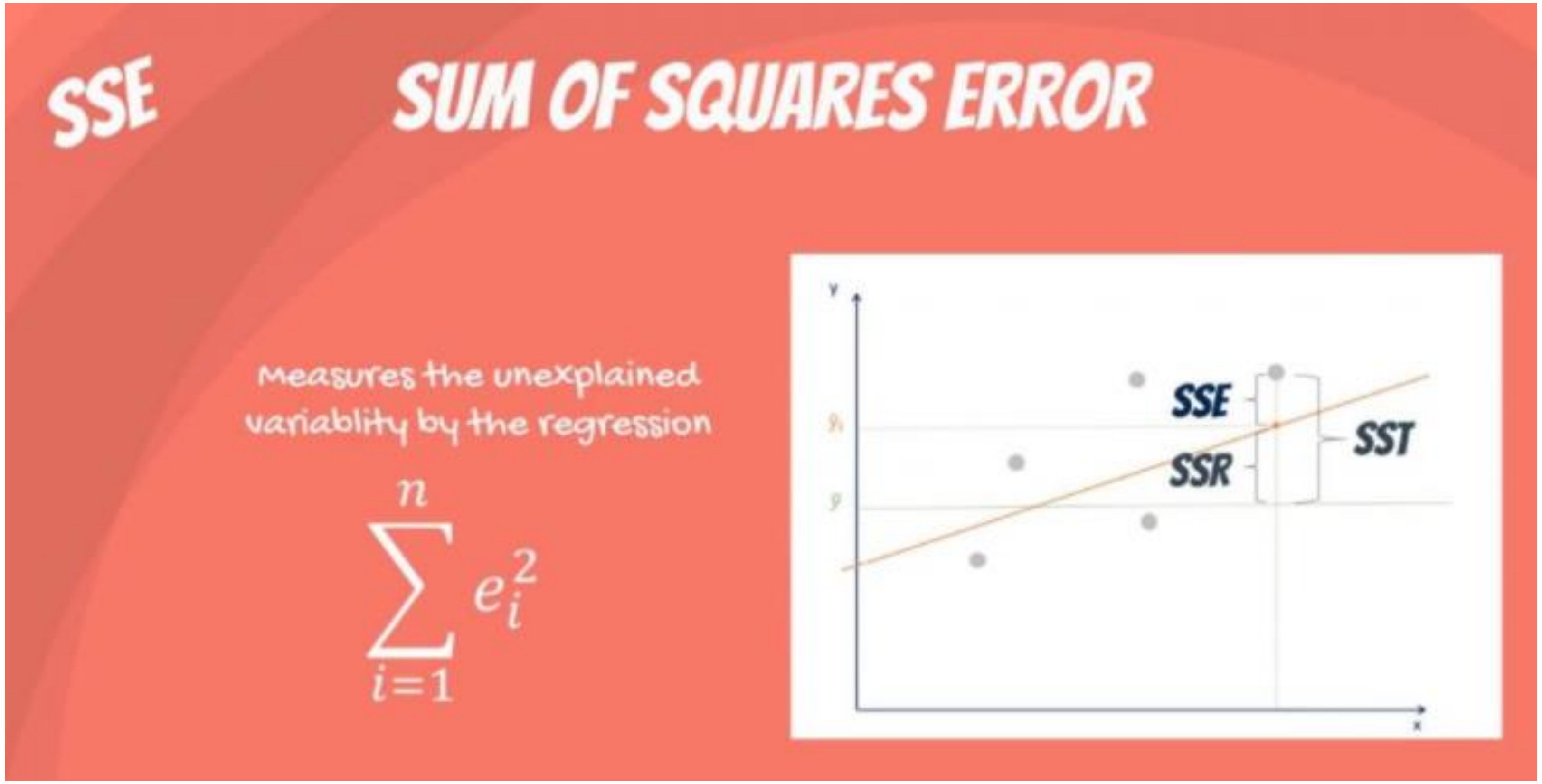
Sum of Squares Total



Sum of Squares Regression



Sum of Squares Error



Metrics

Divide by the total number of data points

Predicted output value

Actual output value

$$MAE = \boxed{\frac{1}{n}} \underbrace{\Sigma}_{\text{Sum of}} \underbrace{\left| \boxed{y} - \boxed{\widehat{y}} \right|}_{\text{The absolute value of the residual}}$$

Mean Absolute Error (MAE)

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Mean Square Error (MSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

Root Mean Square Error

Metrics

- R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$