Machine learning

Predicting continuous target variables

- **Regression techniques** are a category of machine learning algorithms that take labeled data and learn patterns in the data that can be used to predict a continuous output variable.
- The goal of regression is to model the **relationship** between one or multiple variables and a continuous target variable.
- In this case, the machine learning model is called a **regressor**.

- How much carbon dioxide does a household contribute to the atmosphere?
- What will the share price of a company be tomorrow?
- What is the concentration of insulin in a patient's blood?

Regression

Regression analysis is a family of statistical methods that are used to model complex numerical relationships between variables. In general, regression analysis involves three key components.

- A single numeric dependent variable, which represents the value or values that we want to predict. This variable is known as the response variable y.
- One or more independent numeric variables x that we believe we can use to predict the response variable. These variables are known as the **predictors**.
- Coefficients β, which describe the relationships between the predictors and the response variable. We don't know these values going into the analysis and use regression techniques to estimate them. The coefficients are what constitute the regression model.

$$y = f(x, \beta)$$

Simple linear regression

Linear regression is a subset of regression that assumes that the relationship between the predictor variables x and the response variable y is **linear**. In cases where we have only a single predictor variable, we can write the regression equation using the **slope-intercept format.**

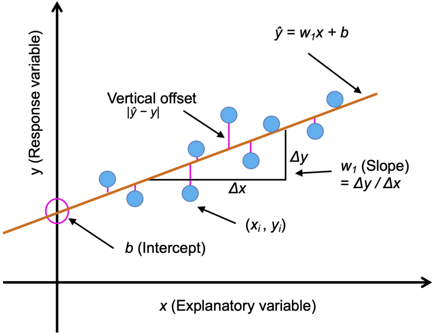
$$y = \beta_0 + \beta_1 x$$

- β_0 - intercept - is the expected value for y when $x = 0$;
- β_1 - slope - is the expected increase in y for each **unit** increase in x.

Simple linear regression

$$y = \beta_0 + \beta_1 x$$

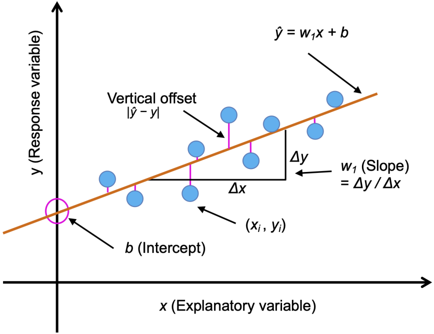
- β_0 - intercept - is the expected value for y when $x = 0$;
- β_1 - slope - is the expected increase in y for each **unit** increase in x.



Simple linear regression

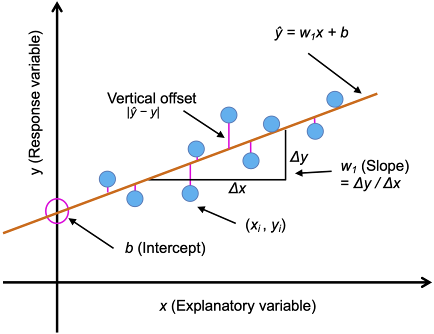
$$y = \beta_0 + \beta_1 x$$

Based on the linear equation that we defined previously, linear regression can be understood as finding the best-fitting straight line through the training examples.



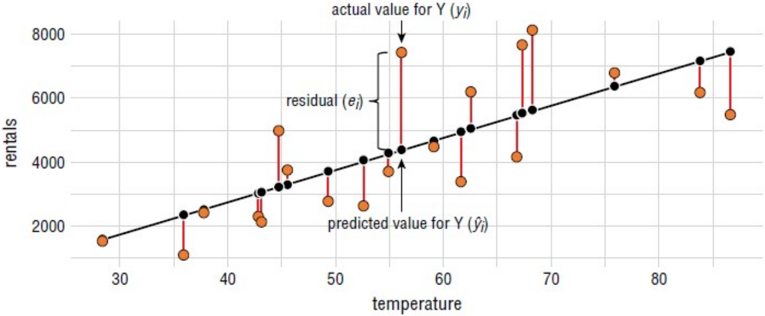
Simple linear regression

$$y = \beta_0 + \beta_1 x$$



The **best-fitting line is also called the regression line**, and the vertical lines from the regression line to the training examples are the so-called **offsets or residuals** — the errors of the prediction.

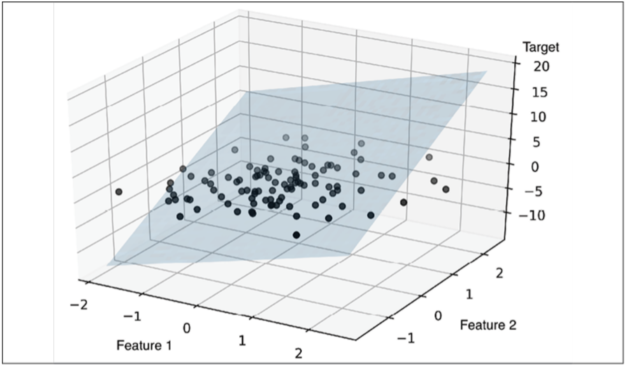
Simple linear regression



Multiple linear regression

Multiple linear regression (MLR)is a method that uses several independent variables to predict the target. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in} = \sum_{i=1}^{n} \beta_i x_i + \beta_0 = \beta^T x + \beta_0$$



Multiple linear regression

The MLR model is based on the following assumptions:

- There is a linear relationship between the dependent variables and the independent variables;
- The independent variables are not too highly correlated with each other;
- observations are selected independently and randomly from the population;
- Residuals should be normally distributed with a mean of 0 and variance.

Regression Metrics

In regression, we have

- y - the actual target values (authentic values)
- $\hat{y}$ - the predicted value (values expected)

The performance of the model is assessed by means of measuring the similarity among these sets.

Regression metrics are quantitative measures used to evaluate the fit of a regression model.

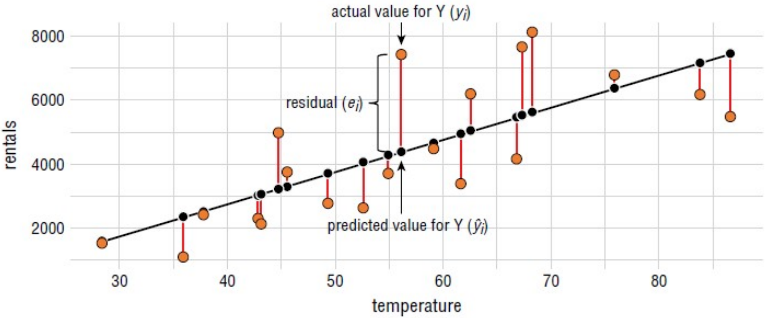
Types of Regression Metrics

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R-squared (R^2) Score
- Root Mean Squared Error (RMSE)

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

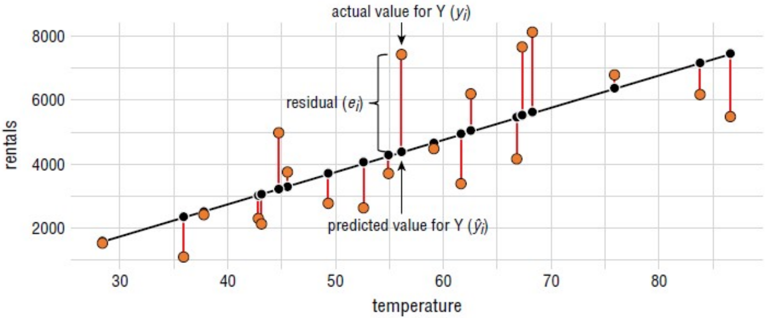
It's a measurement of the typical absolute difference of a dataset's actual values and projected values.



Mean Squared Error (MSE)

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

It's a measurement of the typical absolute difference of a dataset's actual values and projected values.



R-squared (R^2) Score

$$R2 = R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

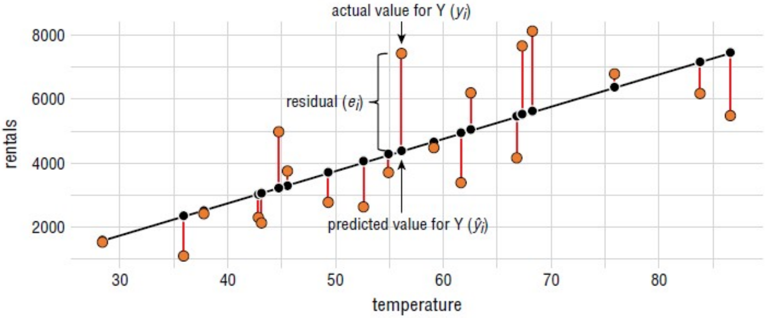
where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

- A statistical metric frequently used to assess the goodness of fit of a regression model is the R-squared (R2) score, also referred to as the coefficient of determination.
- It quantifies the percentage of the dependent variable's variation that the model's independent variables contribute to.
- R2 is a useful statistic for evaluating the overall effectiveness and explanatory power of a regression model.

Root Mean Squared Error (RMSE)

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

It is a usually used metric in regression analysis and machine learning to measure the accuracy or goodness of fit of a predictive model, especially when the predictions are continuous numerical values.



Understanding the output

Let us assume, that ours model describe the prices of cars and that we obtain the following values:

- Mean Absolute Error (MAE): 0.5332001304956553
- Mean Squared Error (MSE): 0.5558915986952444
- R-squared (R2): 0.5757877060324508
- Root Mean Squared Error (RMSE): 0.7455813830127764

So

- An MAE of 0.5332 means that, on average, the model's predictions are approximately 0.5332 away from the true car prices.
- An MSE of 0.5559 means that, on average, the squared prediction errors are approximately 0.5559.
- An R2 of 0.5758 indicates that the model can explain approximately 57.58 % of the variance in car prices.
- An RMSE of 0.7456 indicates that, on average, the model's predictions have an error of approximately 0.7456.

Quantile Quantile plots

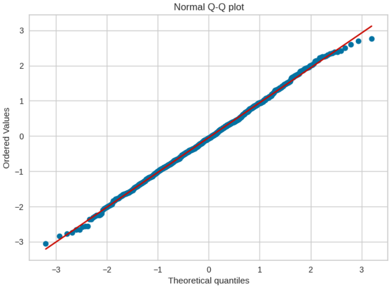
The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not.

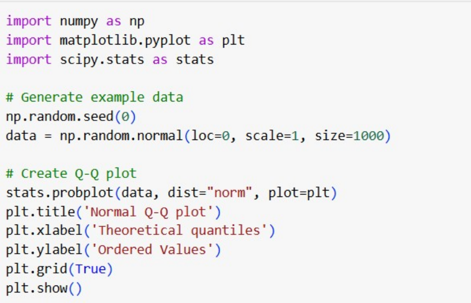
Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.

They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

The most common quantiles are: quartiles (25th, 50th, and 75th percentiles), percentiles.

Quantile Quantile plots





```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

# Generate example data
np.random.seed(0)
data = np.random.normal(loc=0, scale=1, size=1000)

# Create Q-Q plot
stats.probplot(data, dist="norm", plot=plt)
plt.title('Normal Q-Q plot')
plt.xlabel('Theoretical quantiles')
plt.ylabel('Ordered Values')
plt.grid(True)
plt.show()
```

Thank you for your attention!!!