Simple Linear Regression Model

According to the United Nations Environmental Program (UNEP) Sustainable Buildings and Climate Initiative, construction trade contributes as much as 30% to all global greenhouse gas emissions and consumes up to 40% of all energy used worldwide. Climate change is currently having a powerful impact on how buildings are designed and constructed.

Predicting numeric outcomes with some accuracy measure is an important facet of machine learning and data science. For this part, we will use a case study to understand linear regression and its associated cousins. We will learn about the assumptions behind linear regression, multiple linear regression, partial least squares and penalizations. We'll also focus on strategies for measuring regression performance and implementations.

In this module, we will develop a multivariate multiple regression model to study the effect of eight input variables on two output variables, which are the heating load and the cooling load, of residential buildings. The data provided is from the energy analysis data of 768 different building shapes. The features provided are the relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area and glazing area distribution.

Data Source for content: Energy efficiency dataset

Data Quiz: Appliances energy prediction dataset

Simple Linear Regression

The simple linear regression model.

A simple linear regression model estimates the relationship between two quantitative variables where one is referred to as the independent variable and the other the dependent variable. The independent variable (X) is used to predict and also called the predictor while the predicted variable is referred to as the response variable (Y) (e.g. finding the relationship between the amount of CO_2 gas emitted and the number of trees cut down). The value of Y can be obtained from X by finding the line of best fit (regression line) with minimum error for the data points on a scatter plot for both variables. A simple linear regression can be represented as:

$$y = \theta_0 x + \theta_1$$

where

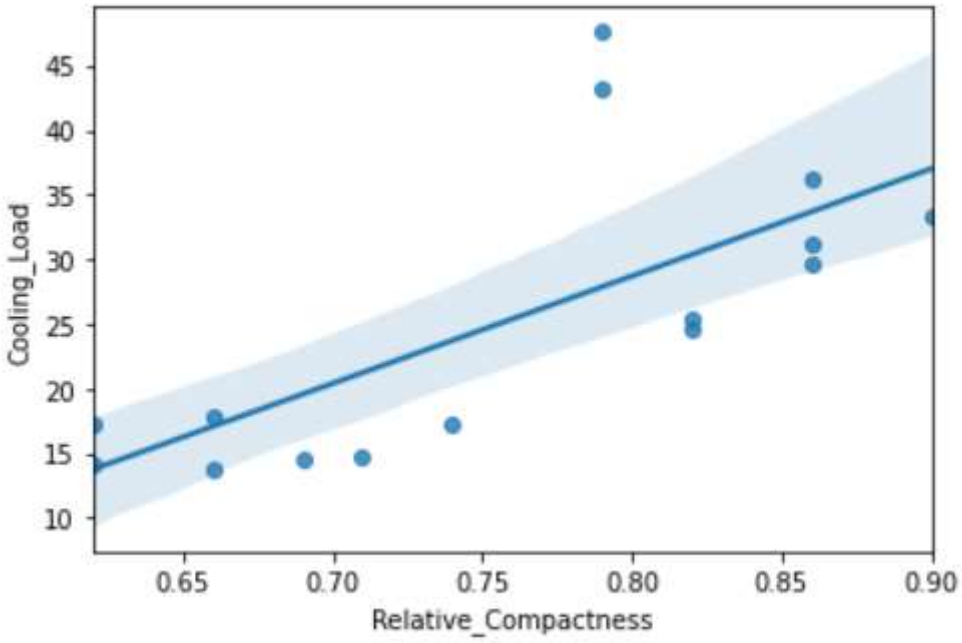
x *is the independent variable*

θ_1 *is the intercept &* θ_0 *the slope' of the line of best fit*

θ_1 *&* θ_0 *are known as regression coefficien*

The UCI Machine Learning Repository: Energy efficiency Data Set is used in this module for better understanding of the concepts(download HERE). We select a sample of the dataset and use the relative compactness column as the predictor and the heating load column the response variable.

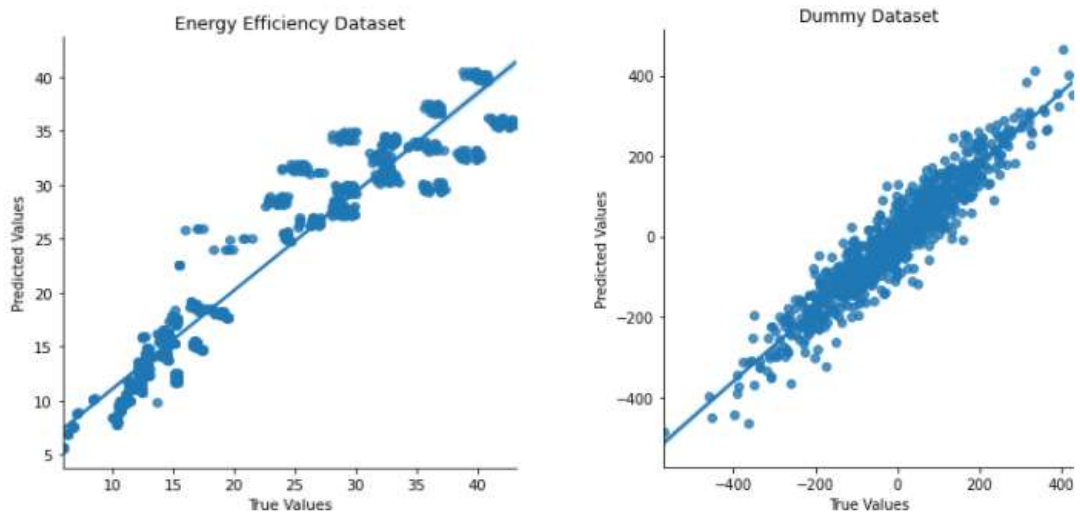
```
import pandas as pd
import numpy as np
import seaborn as sns

df = pd.read_excel('ENB2012_data.xlsx')

#rename columns
column_names = {'X1':'Relative_Compactness', 'X2': 'Surface_Area',
'X3': 'Wall_Area', 'X4': 'Roof_Area', 'X5': 'Overall_Height',
'X6': 'Orientation', 'X7': 'Glazing_Area',
'X8': 'Glazing_Area_Distribution',
'Y1': 'Heating_Load', 'Y2': 'Cooling_Load'}

df = df.rename(columns=column_names)

#select a sample of the dataset
simple_linear_reg_df = df[['Relative_Compactness', 'Cooling_Load']].sample(15, random_state=2)

#regression plot
sns.regplot(x="Relative_Compactness", y="Cooling_Load",
data=simple_linear_reg_df)
```



-Collinearity and Assumptions for Linear Regression

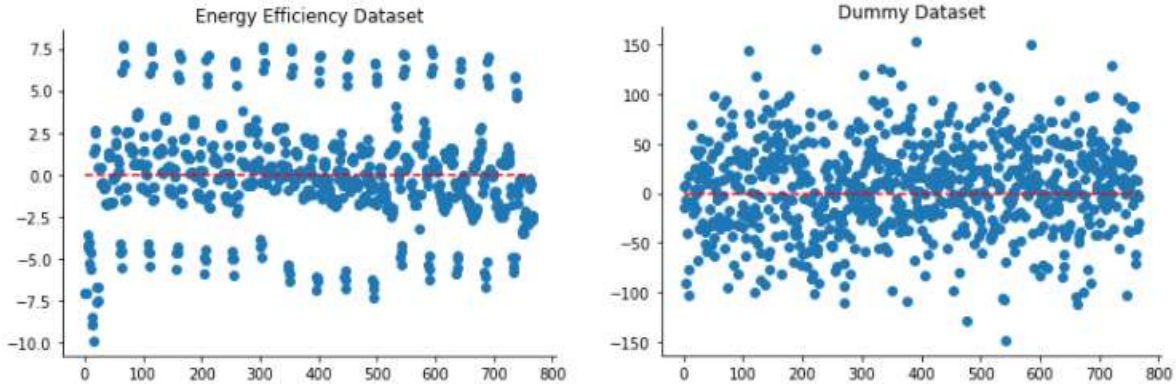
For better understanding, we explain the assumptions made by linear regression by comparing results on our energy efficiency dataset and a dummy linear dataset generated to have similar shape (same number of rows and column) as the energy efficiency dataset. Some assumptions made by linear regression models about the data are:

- Linearity: the relationship between the variables is linear such that a straight line is the line of best fit.



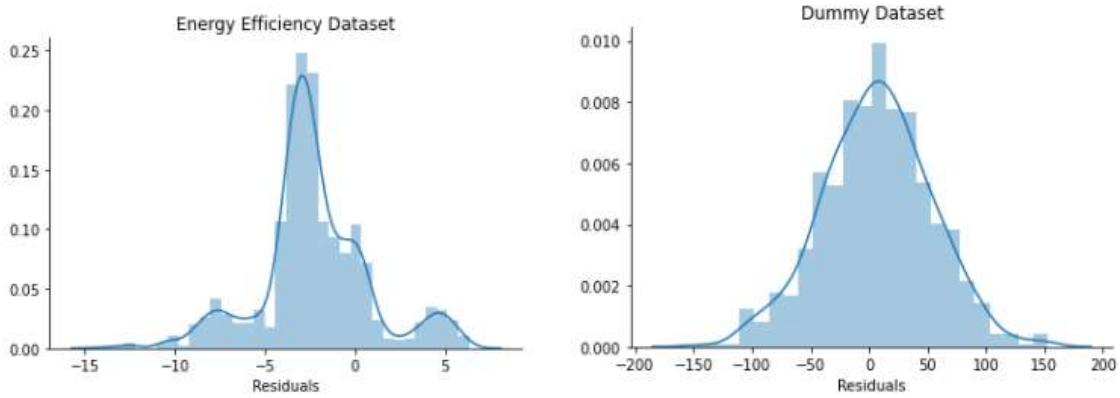
From the regression plots above, we can see that the residuals of the dummy data are spread across the regression line as they should be to meet the linearity assumption unlike the residuals of the energy efficiency dataset which are a bit farther from the regression line.

- Homoscedasticity: the residuals or prediction errors are of equal or constant variance.



The variance of the residuals for the dummy dataset appear to be uniform as opposed to the energy efficiency dataset which violates this assumption.

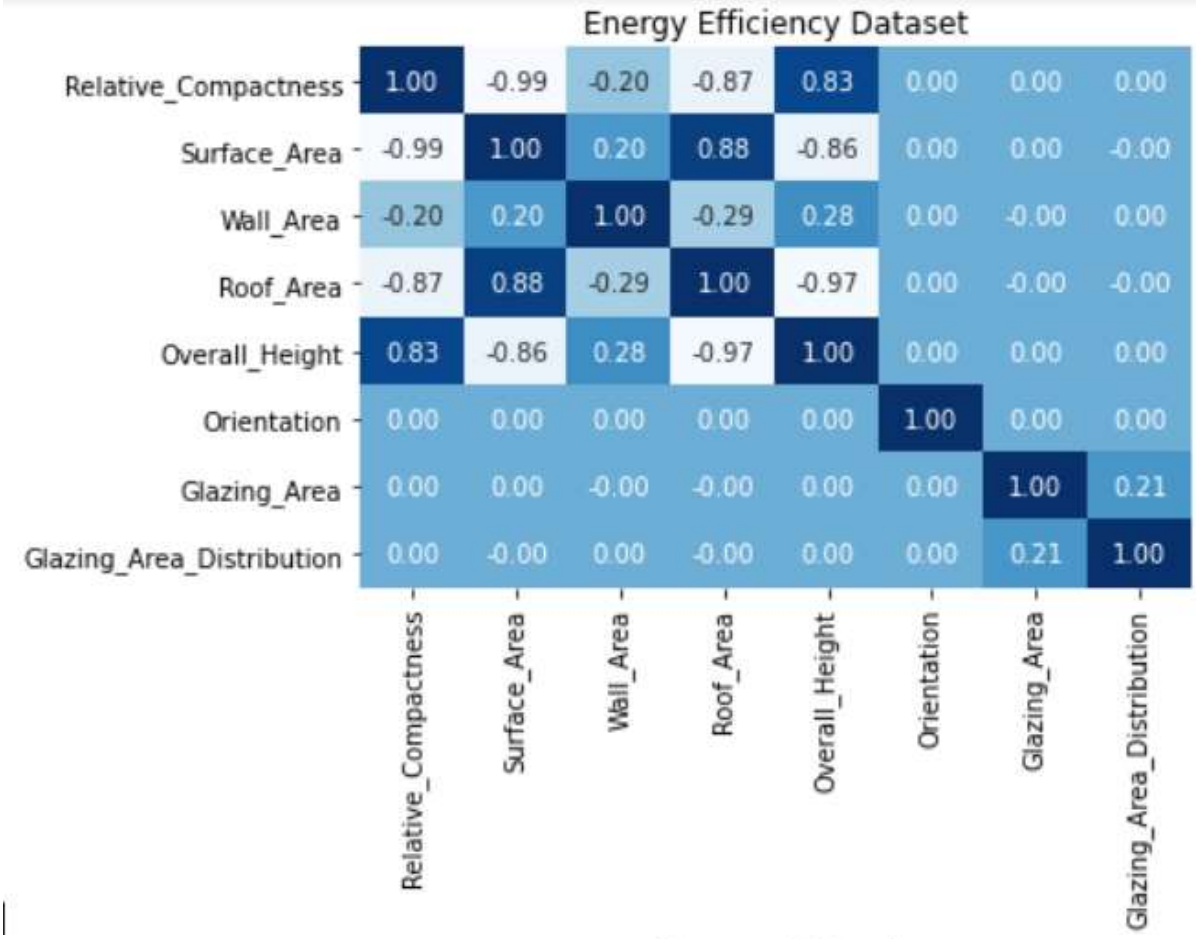
- Normality: the residuals are of a normal distribution

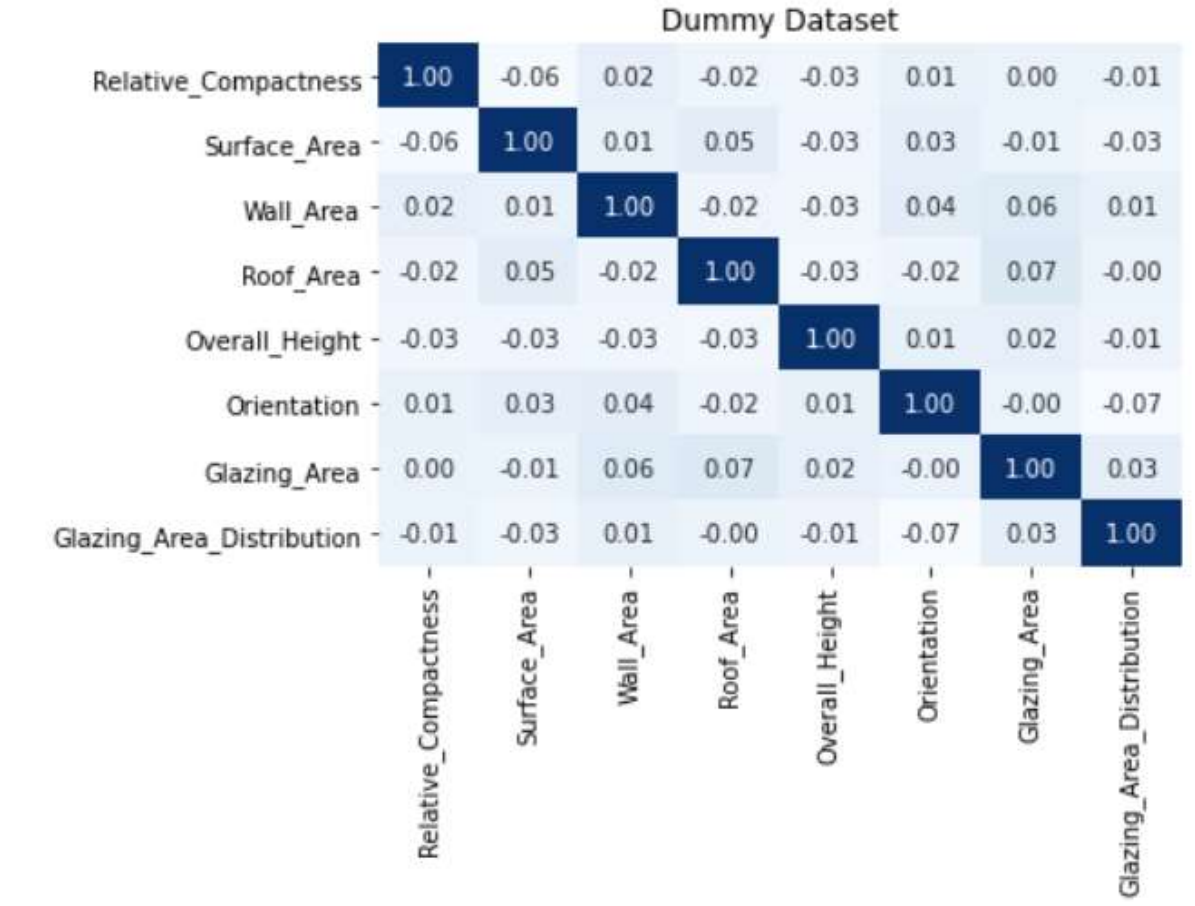


The energy efficiency dataset flouts this assumption as the residuals are clearly not normally distributed while the dummy dataset has normally distributed residuals with the mean and median at 0.

- Independence of the observations

In multiple linear regression where there are more predictors, it is assumed that these variables are independent of each other without any strong correlation between them.


Energy Efficiency Dataset
Relative_Compactness
Surface_Area
Wall_Area
Roof_Area
Overall_Height
Orientation
Glazing_Area
Glazing_Area_Distribution
1.00 -0.99 -0.20 -0.87 0.83 0.00 0.00 0.00
-0.99 1.00 0.20 0.88 -0.86 0.00 0.00 -0.00
-0.20 0.20 1.00 -0.29 0.28 0.00 -0.00 0.00
-0.87 0.88 -0.29 1.00 -0.97 0.00 -0.00 -0.00
0.83 -0.86 0.28 -0.97 1.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00
0.00 0.00 -0.00 -0.00 0.00 0.00 1.00 0.21
0.00 -0.00 0.00 -0.00 0.00 0.00 0.21 1.00


Dummy Dataset
Relative_Compactness
Surface_Area
Wall_Area
Roof_Area
Overall_Height
Orientation
Glazing_Area
Glazing_Area_Distribution
1.00 -0.06 0.02 -0.02 -0.03 0.01 0.00 -0.01
-0.06 1.00 0.01 0.05 -0.03 0.03 -0.01 -0.03
0.02 0.01 1.00 -0.02 -0.03 0.04 0.06 0.01
-0.02 0.05 -0.02 1.00 -0.03 -0.02 0.07 -0.00
-0.03 -0.03 -0.03 -0.03 1.00 0.01 0.02 -0.01
0.01 0.03 0.04 -0.02 0.01 1.00 -0.00 -0.07
0.00 -0.01 0.06 0.07 0.02 -0.00 1.00 0.03
-0.01 -0.03 0.01 -0.00 -0.01 -0.07 0.03 1.00

The energy efficiency dataset shows a strong correlation between relative compactness and surface area, relative compactness and overall height, surface area and roof area while the variables in the dummy dataset are seen to be independent of each other.

Overall, before inferences are drawn from a linear regression model, all the assumptions discussed above must have been met.

- Residual sum of squares and minimizing the cost function

A cost function is a measure of the performance of a model i.e. how far or close the predicted values are to the real values. The objective is to minimise the cost function in order for the model to continuously learn to obtain better results. In linear regression, the cost function can be defined as the sum of squared errors in a training set. The squares of the residuals are taken to penalise errors farther from the line of best fit more than those closer to the line and obtain the best parameter values.

- Gradient descent and coordinate descent algorithm

Gradient descent is an optimization algorithm that minimizes a cost function by specifying the direction to move towards to obtain a local or global minima. This is done by initially starting with random values then iteratively updating the values until the minimum cost is obtained. A learning rate is usually chosen to determine the step size to be taken for each iteration. It is important to carefully select this parameter because, if a small step is chosen, it will take a long time to converge to the minimum cost while if too large, it can result in an overshoot surpassing the location of the minimum cost.