ORDINARY LEAST SQUARES REGRESSOR

Authors: G. H. Zerwes, *Universidade Federal do Espirito Santo, Vitória, Brazil*

Research Supervisor: L. Homri, *Arts et Métiers ParisTech, Metz, France*

1 INTRODUCTION

The ordinary least squares is a regression algorithm that aims to find the best-fitting line that passes through the data points. It can be used to understand the linear relationship between the variables and make predictions for new values.

Suppose you have a set of points X, and they're values Y. The algorithms find the values of coefficients a and b for the best line that passes through X and gives values of Y.

$$Y = a \cdot X + b \tag{1}$$

The above equation can be best visualized by the image 1, in which the data points and the line passing through them can be seen graphically.

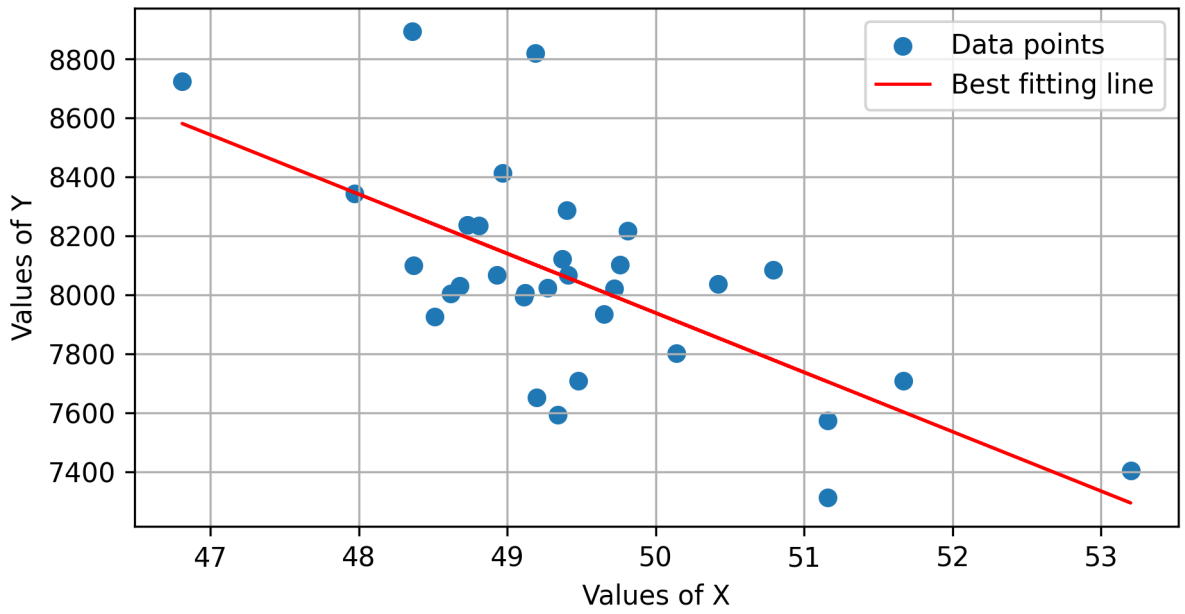

Figure 1: Example of best line for the data

2 Pseudo-code

To find the required parameters, the ordinary least squares algorithms can be used as the following pseudo-code:

```
#Step 1: Define the required training values X and their labels, Y. Ideally, the data
    should be pre-treated, that is to say, it should not contain any missing values, and
    be normalized.

X = [x_value 1, x_value 2, x_value 3, ...]
Y = [y_value 1, y_value 2, y_value 3, ...]

#Step 2: Fit the model with the data, so that it can learn the value of the coefficients
    a and b.

model = LeastSquares.fit(X,Y)

#Step 3: Evaluate the quality of the model. This can be achieved by the R score.
```

```
 R  score = model.score(X,Y)

#Step 4: If the R score is high enough (the closer to 1, the better) the model can be
    used to predict new data points.

X_new = [x_value]
Y_predicted = model.predict(X_new)
```

3 Evaluating the model and other recommendations

After you've trained your model, you should evaluate its performance to see if the parameters were well chosen.

A common metric to evaluate is the R^2 score, which represents how well the line fits the data. However, this metric must be evaluated on the testing dataset, and should not be the only metric considered, as it can also be indicative of over-fitting if the value is too close to 1.

Another metric that is commonly used is the Mean Squared Error (MSE), which represents the average error of the predictions, and penalizes heavily the larger errors of the model.

References

Bishop, C. M. (2006). *Pattern Recognition And Machine Learning*. Number 758. Springer.