LAB_Multiple_Linear_Regression

December 28, 2025

1 Multiple Linear Regression

1.1 Objectives

After completing this lab you will be able to:

- Use scikit-learn to implement Multiple Linear Regression
- Create a model, train it, test it and use the model

Table of contents

```
<ol>
    <li><a href="#understanding-data">Understanding the Data</a></li>
    <li><a href="#reading_data">Reading the Data in</a></li>
    <li><a href="#multiple_regression_model">Multiple Regression Model</a></li>
    <li><a href="#prediction">Prediction</a></li>
    <li><a href="#practice">Practice</a></li>
</ol>
```

1.1.1 Importing Needed packages

[1]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import pylab as pl
```

1.1.2 Downloading Data

To download the data, we will use !wget to download it from IBM Object Storage.

[ ]:
```
!wget -O FuelConsumption.csv https://cf-courses-data.s3.us.cloud-object-storage.
↪appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/
↪Module%202/data/FuelConsumptionCo2.csv
```

Understanding the Data

1.1.3 `FuelConsumption.csv`:

We have downloaded a fuel consumption dataset, **`FuelConsumption.csv`**, which contains model-specific fuel consumption ratings and estimated carbon dioxide emissions for new light-duty vehicles for retail sale in Canada. Dataset source

- **MODELYEAR** e.g. 2014
- **MAKE** e.g. Acura
- **MODEL** e.g. ILX
- **VEHICLE CLASS** e.g. SUV
- **ENGINE SIZE** e.g. 4.7
- **CYLINDERS** e.g 6
- **TRANSMISSION** e.g. A6
- **FUELTYPE** e.g. z
- **FUEL CONSUMPTION in CITY(L/100 km)** e.g. 9.9
- **FUEL CONSUMPTION in HWY (L/100 km)** e.g. 8.9
- **FUEL CONSUMPTION COMB (L/100 km)** e.g. 9.2
- **CO2 EMISSIONS (g/km)** e.g. 182 –> low –> 0

Reading the data in

[2]:
```
df = pd.read_csv ('FuelConsumptionCo2.csv')
df.head()
```

[2]:
```
   MODELYEAR   MAKE       MODEL VEHICLECLASS  ENGINESIZE  CYLINDERS  \
0       2014  ACURA         ILX      COMPACT         2.0          4
1       2014  ACURA         ILX      COMPACT         2.4          4
2       2014  ACURA  ILX HYBRID      COMPACT         1.5          4
3       2014  ACURA     MDX 4WD  SUV - SMALL         3.5          6
4       2014  ACURA     RDX AWD  SUV - SMALL         3.5          6

  TRANSMISSION FUELTYPE  FUELCONSUMPTION_CITY  FUELCONSUMPTION_HWY  \
0          AS5        Z                   9.9                  6.7
1           M6        Z                  11.2                  7.7
2          AV7        Z                   6.0                  5.8
3          AS6        Z                  12.7                  9.1
4          AS6        Z                  12.1                  8.7

   FUELCONSUMPTION_COMB  FUELCONSUMPTION_COMB_MPG  CO2EMISSIONS
0                   8.5                        33           196
1                   9.6                        29           221
2                   5.9                        48           136
3                  11.1                        25           255
4                  10.6                        27           244
```

Let's select some features that we want to use for regression.

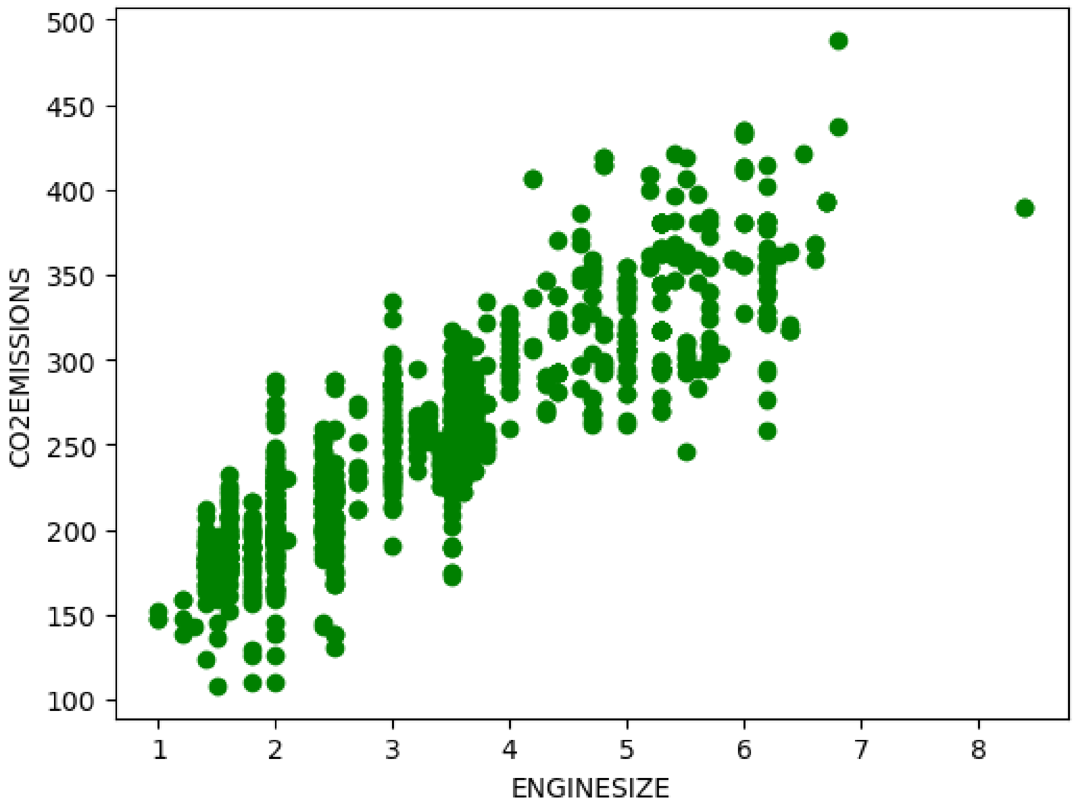
[3]:
```
cdf =␣
 ↪df[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_CITY','FUELCONSUMPTION_HWY','FUELCONSUMPTION_
cdf.head()
```

[3]:
```
   ENGINESIZE  CYLINDERS  FUELCONSUMPTION_CITY  FUELCONSUMPTION_HWY  \
0         2.0          4                   9.9                  6.7
1         2.4          4                  11.2                  7.7
2         1.5          4                   6.0                  5.8
```

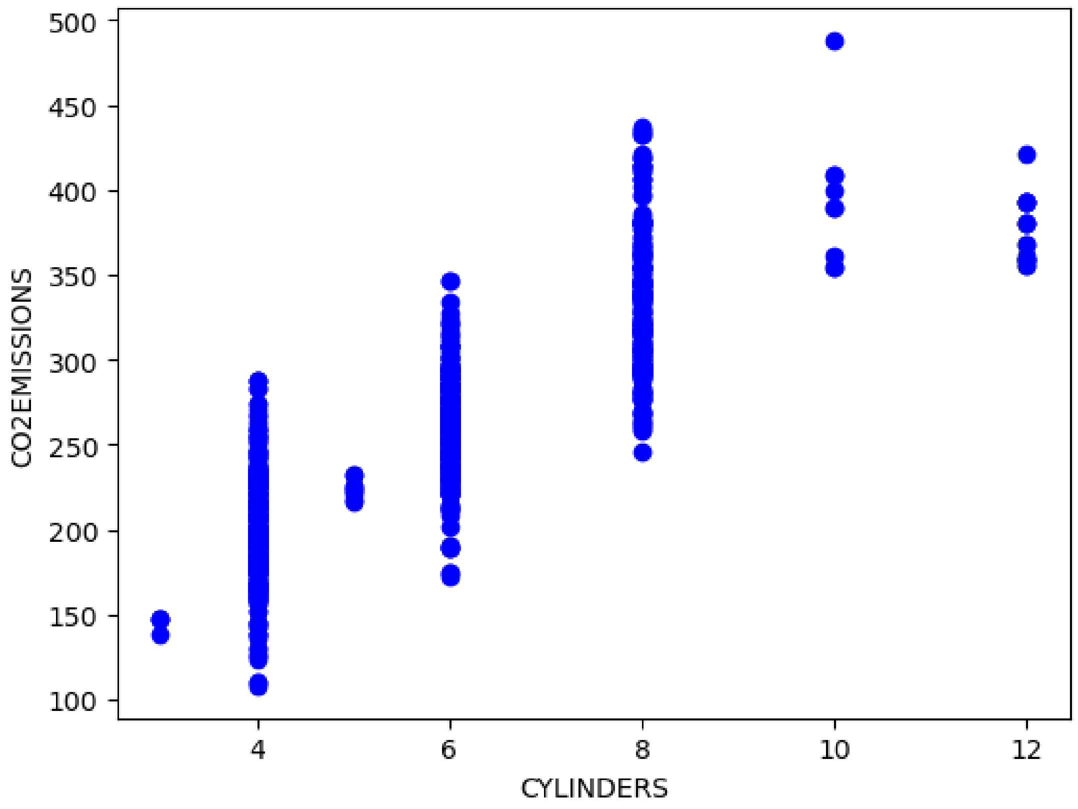
```
3          3.5          6                    12.7                     9.1
4          3.5          6                    12.1                     8.7

   FUELCONSUMPTION_COMB  CO2EMISSIONS
0                   8.5           196
1                   9.6           221
2                   5.9           136
3                  11.1           255
4                  10.6           244
```

Let's plot CO2EMISSIONS values with respect to ENGINESIZE, CYLINDERS, and FUELCONSUMPTION_COMB:

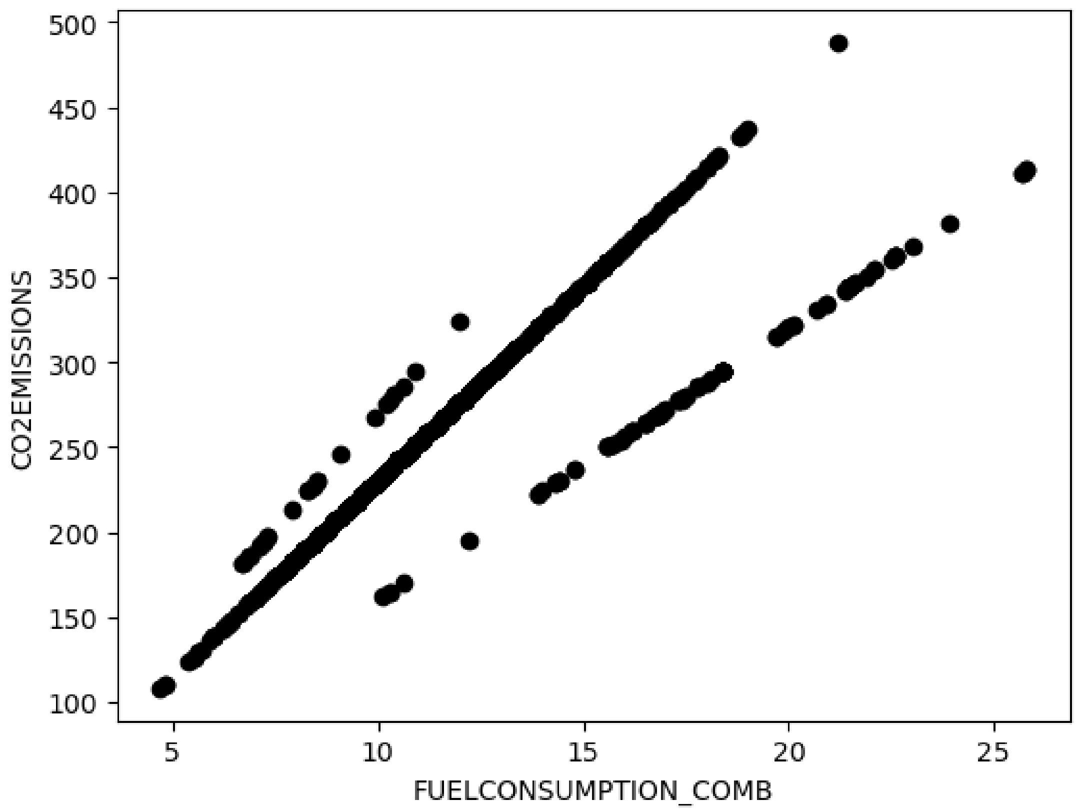
[4]:
```python
plt.scatter (cdf.ENGINESIZE, cdf.CO2EMISSIONS, color = 'green')
plt.xlabel ('ENGINESIZE')
plt.ylabel ('CO2EMISSIONS')
plt.show()
```



[5]:
```python
plt.scatter (cdf.CYLINDERS, cdf.CO2EMISSIONS, color = 'blue')
plt.xlabel ('CYLINDERS')
plt.ylabel ('CO2EMISSIONS')
plt.show()
```



[6]:
```
plt.scatter (cdf.FUELCONSUMPTION_COMB, cdf.CO2EMISSIONS, color = 'black')
plt.xlabel ('FUELCONSUMPTION_COMB')
plt.ylabel ('CO2EMISSIONS')
plt.show()
```



Creating train and test dataset Train/Test Split involves splitting the dataset into training and testing sets respectively, which are mutually exclusive. After which, you train with the training set and test with the testing set. This will provide a more accurate evaluation on out-of-sample accuracy because the testing dataset is not part of the dataset that have been used to train the model. Therefore, it gives us a better understanding of how well our model generalizes on new data.

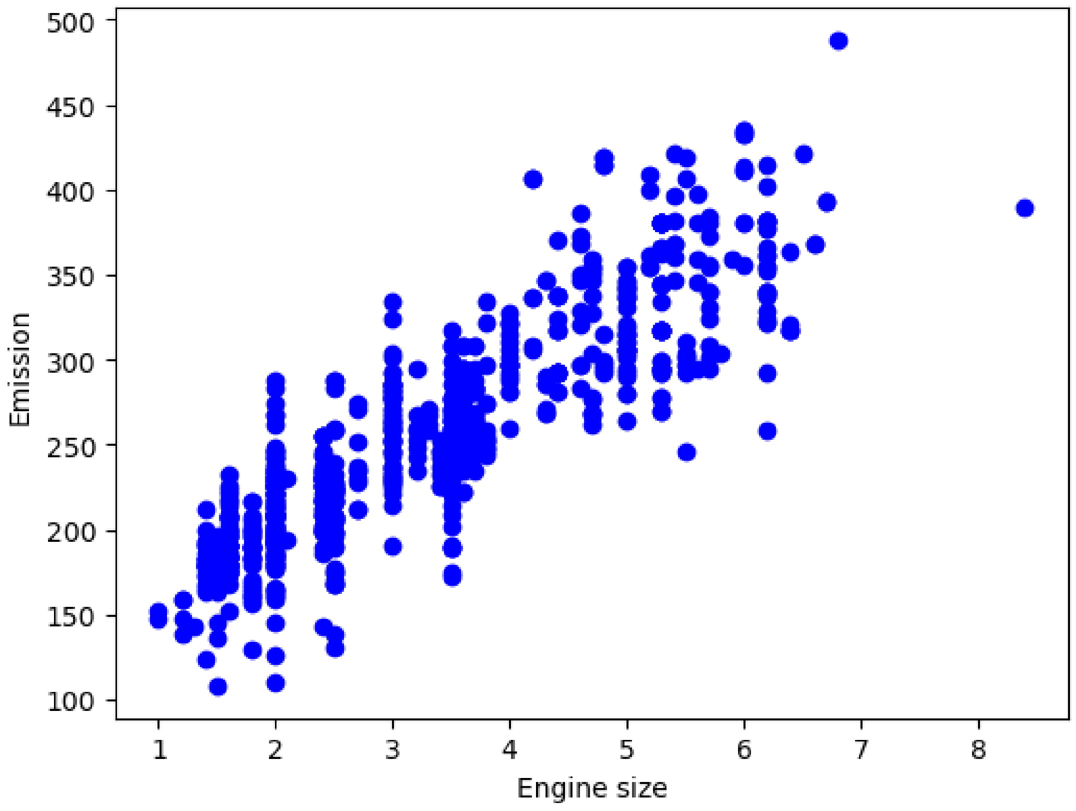
We know the outcome of each data point in the testing dataset, making it great to test with! Since this data has not been used to train the model, the model has no knowledge of the outcome of these data points. So, in essence, it is truly an out-of-sample testing.

Let's split our dataset into train and test sets. Around 80% of the entire dataset will be used for training and 20% for testing. We create a mask to select random rows using the **np.random.rand()** function:

[7]:
```
msk = np.random.rand(len(cdf)) < 0.8
train = cdf[msk]
test = cdf[~msk]
```

Train data distribution

[8]:
```
plt.scatter(train.ENGINESIZE, train.CO2EMISSIONS,  color='blue')
plt.xlabel("Engine size")
plt.ylabel("Emission")
plt.show()
```



Multiple Regression Model

In reality, there are multiple variables that impact the Co2emission. When more than one independent variable is present, the process is called multiple linear regression. An example of multiple linear regression is predicting co2emission using the features FUELCONSUMPTION_COMB, EngineSize and Cylinders of cars. The good thing here is that multiple linear regression model is the extension of the simple linear regression model.

[9]:
```
from sklearn import linear_model
regr = linear_model.LinearRegression()
x = np.asanyarray(train[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_COMB']])
y = np.asanyarray(train[['CO2EMISSIONS']])
regr.fit (x, y)
# The coefficients
print ('Coefficients: ', regr.coef_)
# The intercept
print ('Intercept: ', regr.intercept_)
```

```
Coefficients:  [[10.32446752  7.7028241   9.92272436]]
Intercept:  [62.95125069]
```

As mentioned before, **Coefficient** and **Intercept** are the parameters of the fitted line. Given that it is a multiple linear regression model with 3 parameters and that the parameters are the intercept and coefficients of the hyperplane, sklearn can estimate them from our data. Scikit-learn uses plain Ordinary Least Squares method to solve this problem.

Ordinary Least Squares (OLS) OLS is a method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by minimizing the sum of the squares of the differences between the target dependent variable and those predicted by the linear function. In other words, it tries to minimizes the sum of squared errors (SSE) or mean squared error (MSE) between the target variable (y) and our predicted output ($\hat{y}$) over all samples in the dataset.

OLS can find the best parameters using of the following methods:

- Solving the model parameters analytically using closed-form equations
- Using an optimization algorithm (Gradient Descent, Stochastic Gradient Descent, Newton's Method, etc.)

Prediction

```
[10]: y_hat= regr.predict(test[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_COMB']])
x = np.asanyarray(test[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_COMB']])
y = np.asanyarray(test[['CO2EMISSIONS']])
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))

# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
```

```
Residual sum of squares: 563.91
Variance score: 0.86
```

```
C:\Users\hesam\anaconda3\Lib\site-packages\sklearn\utils\validation.py:2732:
UserWarning: X has feature names, but LinearRegression was fitted without
feature names
  warnings.warn(
```

Explained variance regression score:
Let $\hat{y}$ be the estimated target output, y the corresponding (correct) target output, and Var be the Variance (the square of the standard deviation). Then the explained variance is estimated as follows:

explainedVariance$(y, \hat{y}) = 1 - \frac{Vary - \hat{y}}{Vary}$
The best possible score is 1.0, the lower values are worse.

Practice

Try to use a multiple linear regression with the same dataset, but this time use **FUEL CONSUMPTION in CITY** and **FUEL CONSUMPTION in HWY** instead of FUELCONSUMPTION_COMB. Does it result in better accuracy?

```
[ ]: # write your code here
```

Click here for the solution

```
regr = linear_model.LinearRegression()
x = np.asanyarray(train[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_CITY','FUELCONSUMPTION_HWY']
y = np.asanyarray(train[['CO2EMISSIONS']])
regr.fit (x, y)
print ('Coefficients: ', regr.coef_)
y_= regr.predict(test[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_CITY','FUELCONSUMPTION_HWY']]
x = np.asanyarray(test[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_CITY','FUELCONSUMPTION_HWY']
y = np.asanyarray(test[['CO2EMISSIONS']])
print("Residual sum of squares: %.2f"% np.mean((y_ - y) ** 2))
print('Variance score: %.2f' % regr.score(x, y))
```

Want to learn more?

IBM SPSS Modeler is a comprehensive analytics platform that has many machine learning algorithms. It has been designed to bring predictive intelligence to decisions made by individuals, by groups, by systems – by your enterprise as a whole. A free trial is available through this course, available here: SPSS Modeler

Also, you can use Watson Studio to run these notebooks faster with bigger datasets. Watson Studio is IBM's leading cloud solution for data scientists, built by data scientists. With Jupyter notebooks, RStudio, Apache Spark and popular libraries pre-packaged in the cloud, Watson Studio enables data scientists to collaborate on their projects without having to install anything. Join the fast-growing community of Watson Studio users today with a free account at Watson Studio

1.1.4 Thank you for completing this lab!

1.2 Author

Saeed Aghabozorgi

1.2.1 Other Contributors

Joseph Santarcangelo

1.3 Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2020-11-03	2.1	Lakshmi	Made changes in URL
2020-08-27	2.0	Lavanya	Moved lab to course repo in GitLab

##