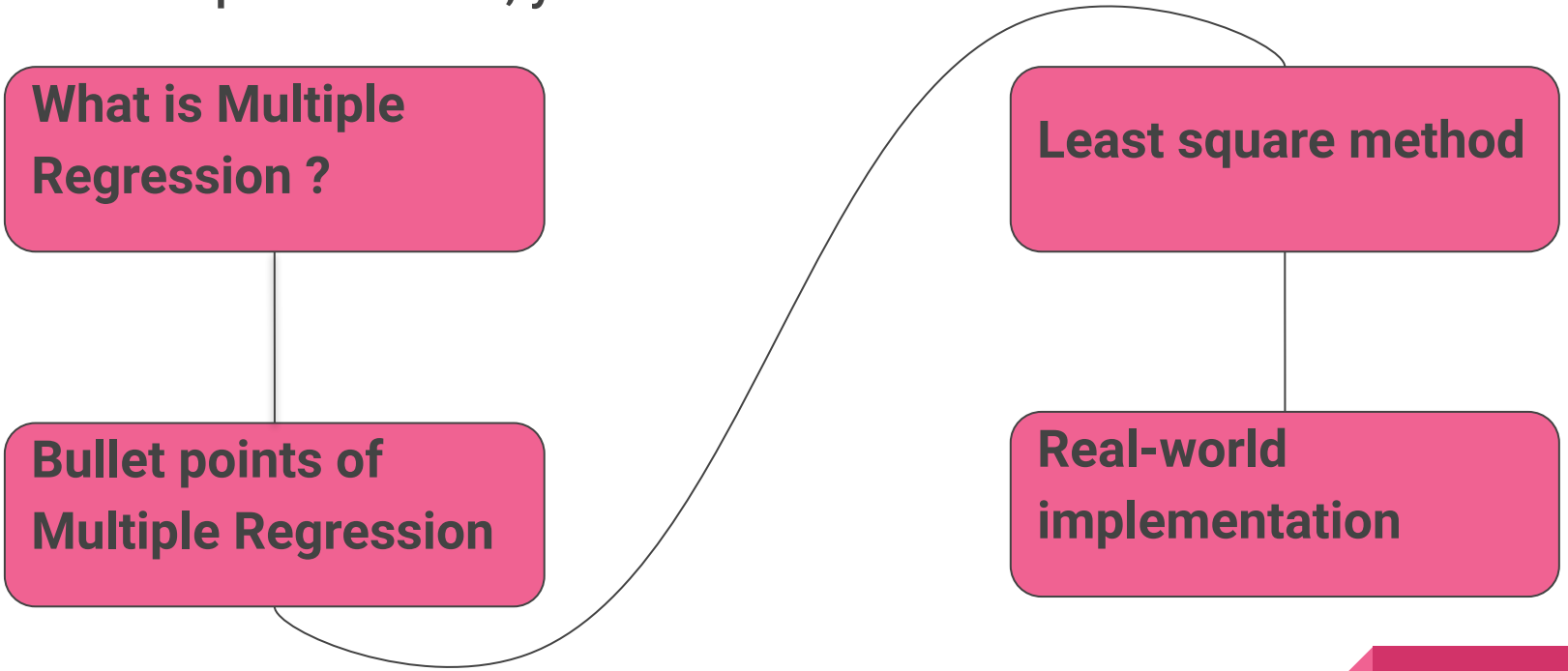
Multiple Regression

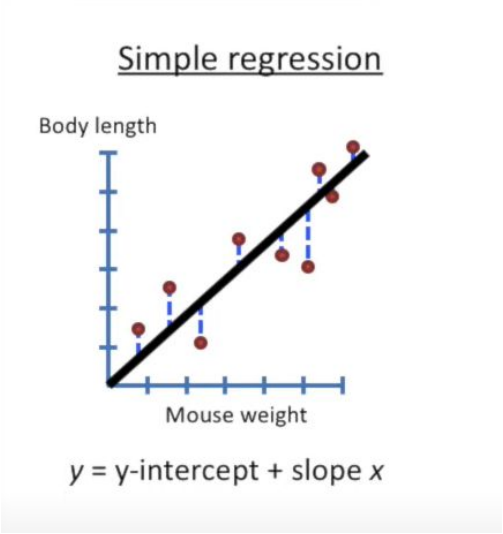
By Apex team

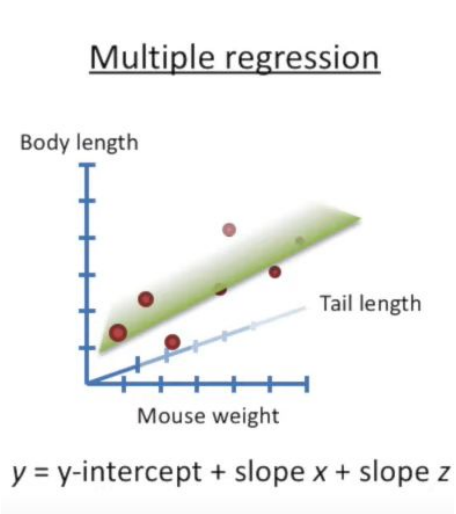
After our presentation, you will know:



What is Multiple Regression ❓

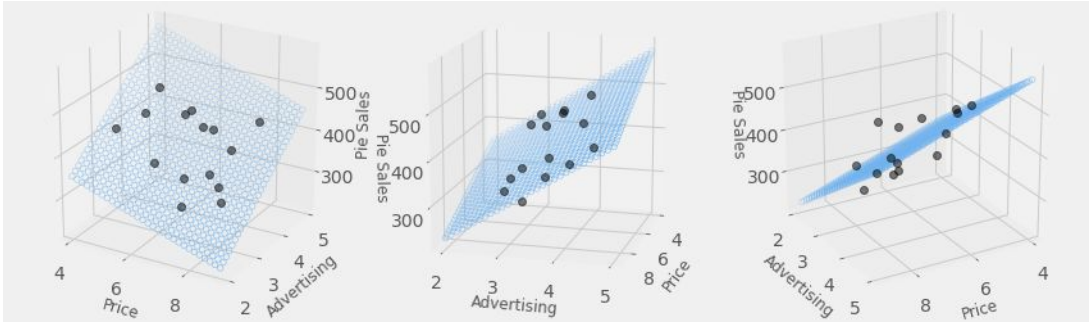
Many people struggle to understand this term , but it is actually very simple to get . So the simple **definition of Multiple Regression** is a Simple Regression, but consists more than one slope variable(predictor variable).





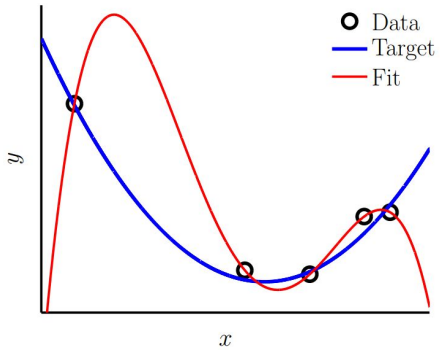
Advantages 👍

- You can have as many independent variables predicting your dependent variables as you want.
- Can allow you to predict your dependent variable with great accuracy
- You can see if 2 or more independent variables are redundant in predicting your dependent variable
- Can allow you to predict your dependent variable with great precision and with fewest numbers of independent variables.

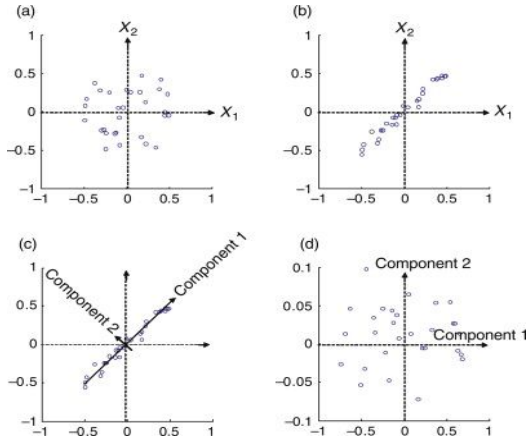


Disadvantages 👎

Overfitting - is a modeling error that occurs when a function or model is too closely fit the training set and getting a drastic difference of fitting in test set. Overfitting the model generally takes the form of making an overly complex model to explain model behaviour in data under study



Multicollinearity - if there's a lot of shared variance or overlap between 2 or more independent variables and when all of them were entered into multiple regression only one would come out significant and the rest would be insignificant. "Winner" independent variable predict the same variance other independent variables can and predict some variance others cannot.



Least Square Method 

$b = (X^T X)^{-1} X^T y$

The least square method is the process of finding the best-fitting curve or line of best fit for a set of data points by reducing the sum of the squares of the offsets (residual part) of the points from the curve.

There are two basic categories of least-squares problems:

- Ordinary or linear least squares
- Nonlinear least squares

Ordinary Least Squares:Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression)

LEAST SQUARE FORMULA

Let us assume that the given points of data are (x1, y1), (x2, y2), (x3, y3), …, (xn, yn)in which all x's are independent variables, while all y's are dependent ones. Also, suppose that f(x) is the fitting curve and d represents error or deviation from each given point.Now, we can write:

d1 = y1 − f(x1) d2 = y2 − f(x2) d3 = y3 − f(x3) ….. dn = yn − f(xn)

The least-squares explain that the curve that best fits is represented by the property that the sum of squares of all the deviations from given values must be minimum, i.e:

$$S = \sum_{i=1}^{n} d_i^{\,2}$$

$$S = \sum_{i=1}^{n} [y_i - f_{x_i}]^2$$

$$S = d_1^{\,2} + d_2^{\,2} + d_3^{\,2} + \cdots + d_n^{\,2}$$

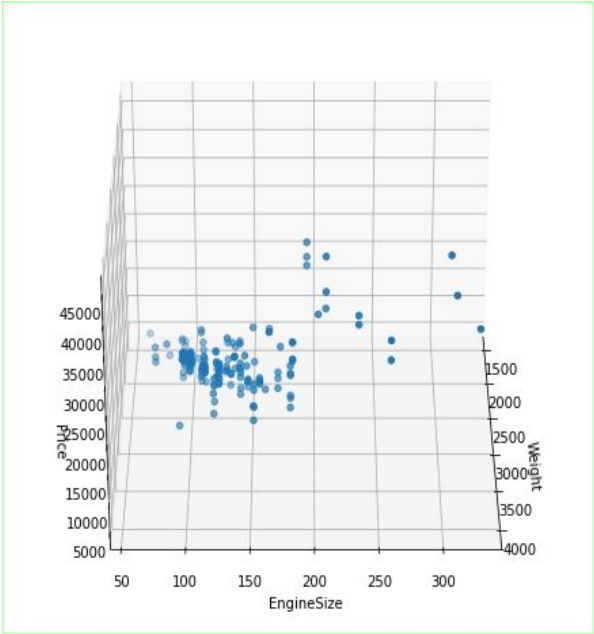
S = Sum Minimum Quantity

Predicting Car Price

First 5 rows of dataset

Weight(x1)	Enginesize(x2)	Price
2548	130	13495.0
2823	152	16500.0
2337	109	13950.0
2824	136	17450.0

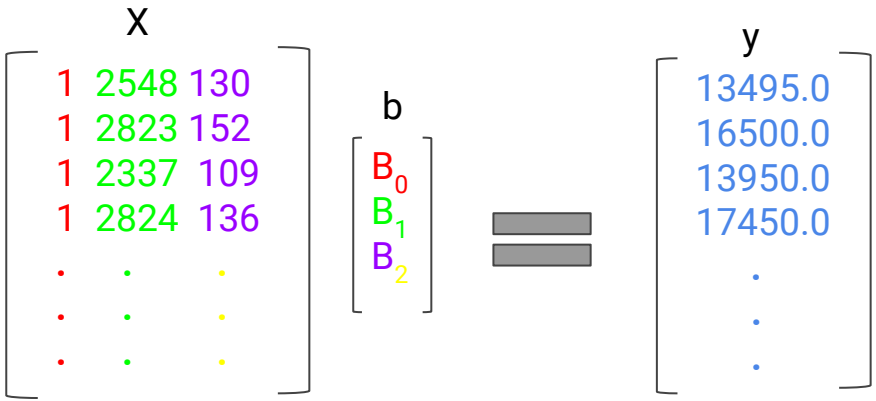
Actual shape: **205x3**



Car Price = Intercept + coef_1*Car Weight + coef_2*Engine Size

Variable to find: ***Intercept, coef_1, coef_2***

$$Y = B_0 + B_1 * X_1 + B_2 * X_2$$



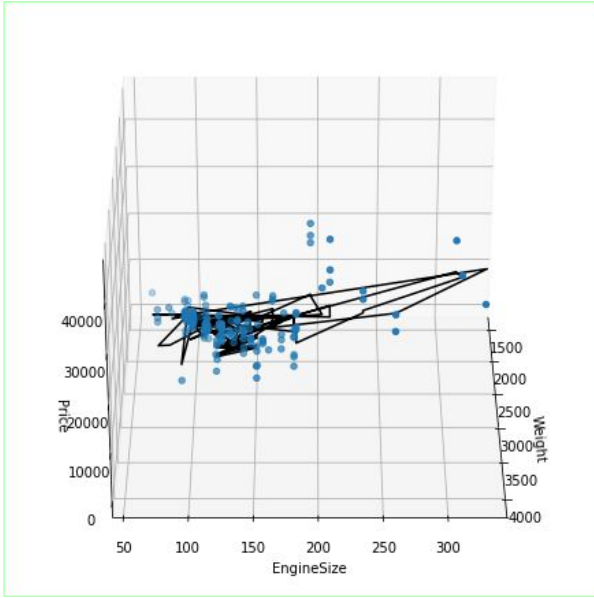
$$b = (X^T X)^{-1} X^T y$$

	curbweight	enginesize	price	B0
0	2548	130	13495.0	1
1	2548	130	16500.0	1
2	2823	152	16500.0	1
3	2337	109	13950.0	1
4	2824	136	17450.0	1
...	...	...	...	...
200	2952	141	16845.0	1
201	3049	141	19045.0	1
202	3012	173	21485.0	1
203	3217	145	22470.0	1
204	3062	141	22625.0	1

Proof of formula in Numpy Python 🐍

```
X = np.array(data.loc[:, ['B0','curbweight', 'enginesize']])
y = np.array(data.iloc[:, 2])
inverse = np.linalg.inv(((X.T).dot(X)))
b = list(inverse.dot(X.T).dot(y))
print("Intercept: ", b[0])
print("Curbweight: ", b[1])
print("EngineSize: ", b[2])
```

```
Intercept:  -14145.808249261576
Curbweight:  5.0921305561061905
EngineSize:  113.54147405172175
```



$$b = (X^T X)^{-1} X^T y$$

Car Price = -14145.8082 + 5.0921*Car Weight + 113.5414*Engine Size

Statsmodels Python 🐍

```python
import statsmodels.api as sm
X = sm.add_constant(data.iloc[:,:2])
model = sm.OLS(data.iloc[:,2],X).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.793
Method:                 Least Squares   F-statistic:                     390.7
Date:                Sat, 21 May 2022   Prob (F-statistic):           3.77e-70
Time:                        11:33:18   Log-Likelihood:                -1970.2
No. Observations:                 205   AIC:                             3946.
Df Residuals:                     202   BIC:                             3956.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.415e+04   1387.924    -10.192      0.000   -1.69e+04   -1.14e+04
curbweight     5.0921      0.930      5.472      0.000       3.257       6.927
enginesize   113.5415     11.635      9.759      0.000      90.601     136.482
==============================================================================
Omnibus:                       38.837   Durbin-Watson:                   0.689
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               86.341
Skew:                           0.871   Prob(JB):                     1.78e-19
Kurtosis:                       5.659   Cond. No.                     1.43e+04
==============================================================================
```

Take-a-ways 🥡

Multiple Regression: Linear regression with more variables

➕ & ➖ of Multiple Regression: Pros: great accuracy and as many independent variables as you want; Cons: problems with overfitting and multicollinearity

Scene behind of Multiple Regression: Least Square Method is used behind the multiple regression

Multiple Regression in Python 🐍: Easy to use with libraries such as sklearn and statsmodels.api

References

https://www.kaggle.com/datasets/hellbuoy/car-price-prediction

multiple-regression-ANOVA.pdf (open.ac.uk)

Multiple Linear Regression (umn.edu)

multiple-regression.pdf (statstutor.ac.uk)

Design Matrix & Normal Equations for Simple & Multiple Linear Regression.

Multiple Regression | Ch. 4, Linear Regression

Scan me!

Thank you for your attention