

Machine learning theory vs classic approach in linear regression estimation

التقدير الخطي لنماذج الانحدار بين النظرية الكلاسيكية ونظرية التعلم الآلي

Medini Atmane

University of Souk Ahras, laboratory: Financial, accounting, collection and insurance -Souk Ahras (Algeria), a.medini@univ-jjel.dz

Received: 22/07/2023**Accepted:** 05/01/2024**Published:** 06/01/2024**Abstract:**

Regression theory in classical models assumes the unity of data in order to reach a standard relationship that explains the type and direction of the relationship between the variables. While machine learning models simulate human intelligence within a deeper style to give a stronger interpretation according to mathematical formulas based on artificial intelligence theory, especially when it comes to big data. The aim of this study is to compare the classical regression models using the sm and smf functions and machine learning models in Python by showing the relationship between the variables involved. Which found machine learning models that more accurately interpret the mathematical relationship between the data after dividing it into training and experimental data to reach that the final model (70%) of the experimental data was the best in estimating the standard relationship.

Keywords: Regression; Classic model; machine learning; intelligence artificial; Python.

JEL Classification Codes: C01, C19, C60, D85 ,D99

ملخص:

تفترض نظرية الانحدار في النماذج الكلاسيكية وحدة البيانات من أجل الوصول إلى علاقة معيارية تفسر نوع واتجاه العلاقة بين المتغيرات. بينما تحاكي نماذج التعلم الآلي تحاكي الذكاء البشري ضمن أسلوب أعمق لتعطي تفسيراً أقوى وفقاً للصيغ الرياضية القائمة على نظرية الذكاء الاصطناعي خاصة إذا تعلق الأمر بالبيانات الضخمة. حيث تهدف هذه الدراسة إلى المقارنة بين نماذج الانحدار الكلاسيكية الاعتيادية باستخدام دالة sm و smf ونماذج التعلم الآلي في برمجية بايثون بعرض التعرف على العلاقة بين المتغيرات المتضمنة. والتي توصلت إلى نماذج التعلم الآلي تفسر بصورة أدق العلاقة الرياضية بين البيانات بعد تقسيمها إلى بيانات تدريبية وأخرى تجريبية لتصل إلى أن النموذج النهائي (70%) من البيانات التجريبية كان الأفضل في تقدير العلاقة القياسية.

كلمات مفتاحية: الانحدار ، النموذج الكلاسيكي، التعلم الآلي ، الذكاء الاصطناعي ، بايثون

تصنيفات JEL : C01 ، C19 ، C60 ، D85 ، D99

Corresponding author: Medini atmane, e-mail: a.medini@univ-jjel.dz

INTRODUCTION:

Regression analysis models vary according to the data and the objective of the study. We can also distinguish between more than one model. The classical method deals with the data as a single block and then estimates the standard relationship according to the visualization of simple or multiple regression using the method of least squares. As for the regression models that rely on machine learning, which is part of the mechanisms of artificial intelligence, they adopt a completely different method that divides the data. And then make a preliminary assessment between the experimental evidence to reach a more powerful relationship between the study variables and between the experimental called tested data. So artificial intelligence build models that can mimic human intelligence in specific tasks, such as predicting housing prices, stock market trends, or customer behavior. Regression models provide a mathematical framework for learning and making predictions based on input data, contributing to the development of intelligent systems. machine learning, including regression models, plays a crucial role in the development of artificial intelligence by enabling systems to learn from data and make predictions or decisions based on patterns and relationships identified in the data.

1. Basic knowledge Linear Regression

Linear Regression is a statistical/machine learning technique that attempts to model the linear relationship between the independent predictor variables X and a dependent quantitative response variable Y . (Vishwa, 2020). In a simple regression model, the goal is to understand and model the relationship between a dependent variable and a single independent variable. It's worth noting that this explanation covers the basic steps involved in a simple regression model. There are additional considerations and techniques that can be applied to enhance the model's accuracy, such as checking assumptions, handling outliers, evaluating model diagnostics, and exploring more complex regression models. The basic steps involved in creating a simple regression model (Jason, 2016):

- ✓ **Gather Data:** Collect data that includes observations of both the dependent variable (often denoted as "Y") and the independent variable (often denoted as "X"). Each observation should have a pair of values for Y and X.
- ✓ **Visualize the Data:** Plot the data on a scatter plot to visually examine the relationship between the dependent variable and the independent variable. This helps determine if there is a potential linear relationship between the variables.
- ✓ **Choose the Regression Equation:** Based on the scatter plot, select an appropriate equation that represents a linear relationship between Y and X. In its simplest form, this equation is represented as $(Y = \beta_0 + \beta_1 X)$, where β_0 is the intercept and β_1 is the slope coefficient.
- ✓ **Estimate Model Parameters:** Use a method such as the method of least squares to estimate the model parameters (β_0 and β_1). The method of least squares aims to minimize the sum of squared residuals between the observed Y values and the predicted Y values based on the regression equation.
- ✓ **Assess Model Fit: Evaluate** the goodness of fit of the model by examining statistical measures such as the coefficient of determination (R-squared), which indicates the proportion of the variance in the dependent variable explained by the independent variable.
- ✓ **Interpret the Model:** Analyze the estimated coefficients (β_0 and β_1) to understand the relationship between the dependent variable and the independent variable. The intercept (β_0)

Machine learning theory vs classic approach in linear regression estimation

represents the expected value of Y when X is zero, and the slope (β_1) represents the change in Y associated with a one-unit change in X.

- ✓ **Make Predictions:** Once the model is built and validated, you can use it to make predictions by plugging in new values of X into the regression equation to estimate the corresponding Y values.

1.1. Simple regression models vs machine learning regression approach

Simple Regression Model is a regression model typically refers to a basic statistical model that assumes a linear relationship between the dependent variable and one independent variable. It aims to find the best-fit line that minimizes the sum of squared residuals, well Machine learning regression models, encompass a wide range of algorithms that can capture complex relationships between variables. These models are capable of handling multiple independent variables and can capture nonlinear patterns in the data. Examples of machine learning regression models include linear regression, decision trees, random forests, support vector regression, and neural networks. Also we can say that Machine Learning is the solution when data is large, and relation becomes difficult to quantify manually (Devashree , 2023).

1.2. Requirement assumptions in regression approach

The relationship between the dependent variable and independent variable(s) is defined by a predetermined equation, typically a straight-line equation

($y = b \cdot x + c$). The model estimates the values of the slope (b) and intercept (c) to fit the data.

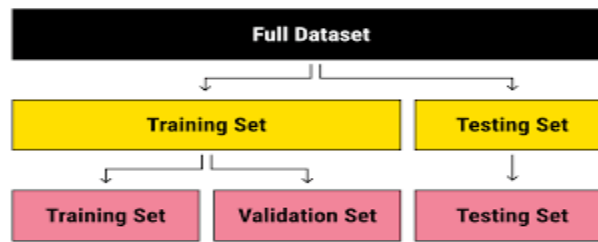
There are many Assumptions of Linear Regression (Devashree , Guide, 2023).

- ✓ **Linearity:** The models of Linear Regression models must be linear
- ✓ **Homoscedasticity:** Homoscedasticity means the standard deviation and the variance of the result or let's say model fitted $(y - \hat{y})^2$ must be the same for any values of independent variable(x).
- ✓ **Non-multicollinearity:** The data should not have multicollinearity
- ✓ **No Autocorrelation:** When data are obtained across time
- ✓ **Not applicable to Outliers:** The value of the dependent variable cannot be estimated for a value of an independent variable which lies outside the range of values in the sample data.
- ✓

1.3. Machine Learning Regression Model

Regression analysis is a predictive technique and modelling analyzes that study the relation between the target or dependent variable and independent variable in a dataset (Pavan Vadapalli, 2022). Simple regression model assumes a linear relationship between variables and follows a predetermined equation. However, machine learning regression are more flexible, capable of handling complex relationships, and learn patterns from the data itself. Machine learning regression models take a data-driven approach and learn the relationship between the variables from the provided data. These models automatically adjust their internal parameters during the training process by minimizing a chosen loss function. The model iteratively updates its parameters to optimize its predictions based on the available training data, machine learning approach often require more computational resources and larger datasets but can provide more accurate predictions in various scenarios.

Fig (02): Training vs testing data process in in ML approach



Source: <https://labelyourdata.com>

3. Algorithms of machine learning in python

There is several software can be used to explain the relation sheep between variables. But its Store the principal method to run the process python, R_studio or any other software require libraries and frameworks available that provide implementations of various algorithms for modeling. In objective to improve the model's performance (nikhilag, No date):. We shod also note that the libraries must be required and installed in the environment of work. Some of these commonly used machines learning libraries algorithm are (EMMANUEL , 2020) :

- ✓ **Tensor Flow:** Neural Networks: `import tensor flow as tf`
- ✓ **Keras:** Neural Networks: `from Keras. models import Sequential`
- ✓ **XGBoost:** Gradient Boosting: `import XGBoost as xgb`
- ✓ **LightGBM:** Gradient Boosting: `import LightGBM as lgb`
- ✓ **From sklearn.** model selection `import train_split_test`

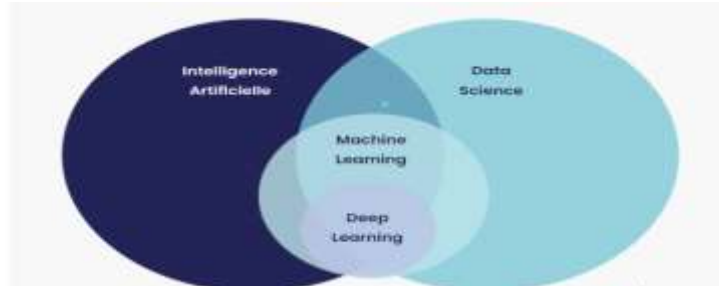
Machine learning & intelligence artificially(AI) in regression approach

Machine learning and artificial intelligence (AI) are closely related, and regression models are one type of machine learning algorithm used in AI. Regression is a type of machine learning algorithm used for predicting continuous numerical values based on input data. It aims to find the relationship between a dependent variable (the value to be predicted) and independent variables (features or attributes) which reduced the time consumption and cost of the process (Rohan Gupta, 2021). Well AI (Artificial Intelligence) and ML (machine learning) technology are used to improve the accuracy of prediction for screening both of many other sectors and science as infectious and non-infectious diseases (Samuel & others, 2020). Regression models learn from historical data and use it to make predictions or estimate values for new, unseen data points. Machine learning is a subfield of AI that focuses on creating algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed, it involves developing mathematical and statistical models that can identify patterns and relationships in data. Machine learning is a branch of Artificial Intelligence (AI) focused on building applications that learn from data (Partha Pratim Mandal & Reza Rezaee, 2019) .Artificial intelligence refers to the broader field of creating intelligent systems that can simulate human intelligence and perform tasks that typically require human intelligence, such as reasoning, problem-solving, and learning. Machine learning is a key component of AI, enabling the development of intelligent systems by leveraging data and algorithms. All that can explain the Relationship. Also regression approach can be used as part of AI systems to learn patterns and make predictions or decisions. In the context of artificial intelligence, the goal is to build models that can mimic human intelligence in specific tasks, such as predicting housing prices, stock market trends, or customer behavior. Regression models provide a mathematical

Machine learning theory vs classic approach in linear regression estimation

framework for learning and making predictions based on input data, contributing to the development of intelligent systems. In our case we take the next algorithm in order to run the machine learning (Ptreek , no date).

Fig(02):IA & Ma& Deep learning concepts



Source: <https://datavalue-consulting.com>

2.Applied Study

In the applied part of the study, continuous quantitative data representing two types of data will be relied upon, and loyer which represents the dependent renter and the area of the property that represents the independent variable. The model will be estimated through Python software. The approved algorithms can be indicated as the next work:

fig (03): Algorithm used in classic and machine learning regression model

```
1  'Generation of machine learning model in python'
2  "Created on Sun Jul  9 09:29:37 2023"
3  "@author: Dr.Medini atmane"
4  import pandas as pd
5  import numpy as np
6  import matplotlib.pyplot as plt
7  ##package of machine learning regression modeling
8  from sklearn.linear_model import LinearRegression
9  from sklearn.model_selection import train_test_split
10 from sklearn.metrics import mean_squared_error
11
12
13 ##importation of data
14 data=pd.read_csv('house.csv')
15 y=data['loyer']
16 x=data['surface']
17 #####Classic regression modeling using sm function
18 import statsmodels.api as sm
19 classicmodel=sm.OLS(y,x).fit()
20 classicmodel.summary()
21
22
23
24 # #####Split the data into training and testing sets machineLearning model
25 x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
26 ## Create a Linear Regression model
27 MachineLearningmodel = LinearRegression()
28 MachineLearningmodel.fit(x_train, y_train)
29 ## Train the model on the training data
30 MachineLearningmodel.fit(x_train, y_train)
31 ##Make predictions on the testing data
32 y_pred=MachineLearningmodell.fit(x, y)
33
34
```

Source: prepared by the research using python 3.9

The main objective in our work consists to detect the difference between classic regression modeling and process algorithm using python software in objective to analysis the relation sheep between the variables. That need to prove that from the statistic's criteria of the models.

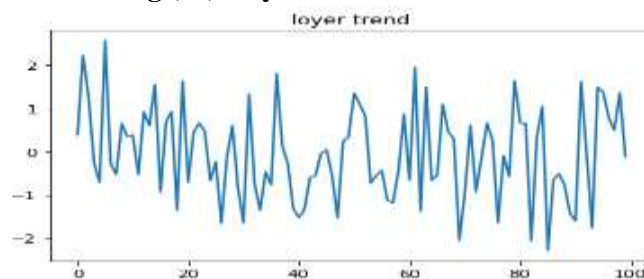
1. Classic approach Regression

The basic steps involved in creating a simple regression model, begin from the visualization of data in order to classify and located the depended and the in depended variables. The second steep should be interested to choose the equation model based on the number of the variable's in the data. after the estimation of the parameters of the model we shod assets and fitting it in order to interpreted it and make decision about the nature of the relation can be between the variable's.

1.1. Sm classic function regression

Sm function fitting the model without intercept. All the result can be taken generate the relation between the y as depended variable and x as in depended variable.

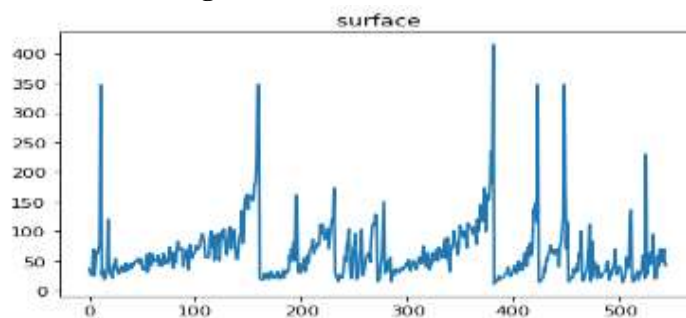
fig.(04):Loyer trend séries



Source: prepared by the research using python 3.9

```
plt. Plot(x)
plt.title(" surface ")
plt.show()
```

fig(05):Surface trend series



Source: prepared by the research using python3.9

▪ génération of the distribution

```
# Générâtes a Random dataset
y = np.random.normal(loc=0, scale=1, size=100)
# Perform Shapiro-Wilk test
statistic, p_value = stats.shapiro(y)
p_value
# Interpret the results
alpha = 0.05 # significance level
if p_value > alpha:
    print("The data follows a normal distribution (fail to reject H0)")
```

Machine learning theory vs classic approach in linear regression estimation

```
else:
    print("The data does not follow a normal distribution (reject H0)")
```

X out_put

Fig(06):static out put criteria in classic model regression

```
In [19]: statistic, p_value = stats.shapiro(x)

In [20]: p_value
Out[20]: 0.0663667842745781

In [21]: alpha = 0.05 # significance level

In [22]: print("The data follows a normal distribution (fail to reject H0)")
The data follows a normal distribution (fail to reject H0)
```

Source: prepared by the research using python 3.9

Y_ outputs:

Fig (07): normality output test of depended variable

```
In [24]:
...: y = np.random.normal(loc=0, scale=1, size=100)
...: # Perform Shapiro-Wilk test
...: statistic, p_value = stats.shapiro(y)
...: p_value
Out[24]: 0.32253333926200867
```

Source: prepared by the research using python 3.9

1.2. Classic regression model using sm and smf function

Classic regression approach can be predicting from Sm_ or smf_ fuction in python software. the main difference between the two models can be summarized in the existence or not of the intercept. In our case we are going to make the two model regression (sm and smf) in objective to detect the difference between the two fitted models.

- **Smf _function in classic regression approach**

```
import statsmodels.formula.api as smf
formula=('y~x')
modell=smf.ols(formula=formula, data=data).fit()
modell.params
modell.summary()
import statsmodels.api as sm
# in sm model we shoold not use formula, we just have to wright y,x
model0=sm.OLS (y, x,data=data).fit()
model.Params
```

table (01): smf_ fuction regression model between (Loyer and surface)

OLS Regression model between (Loyer and surface using smf_ fuction			
Dep. Variable:	y (Loyer)	R-squared:	0.822
Model:	OLS	Adj. R-squared:	0.822

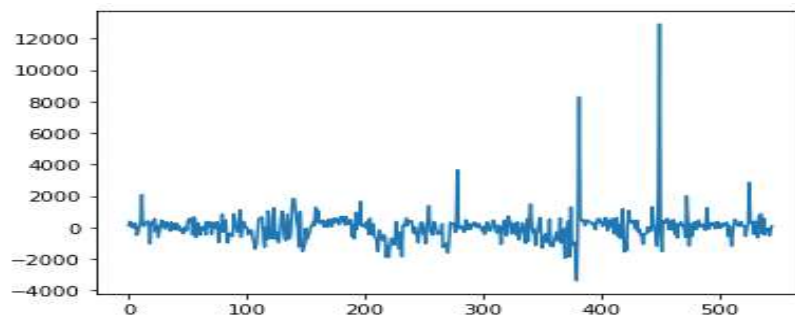
Medini Atmane

Method:	Least Squares			F-statistic:	2509.	
Date:	Sat, 08 Jul 2023			Prob (F-statistic):	9.99e-206	
Time:	08:55:35			Log-Likelihood:	-4497.1	
No. Observations:	545			AIC:	8998.	
Df Residuals:	543			BIC:	9007.	
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	-283.3784	65.268	-4.342	0.000	-411.587	-155.170
x	40.9712	0.818	50.093	0.000	39.365	42.578
Omnibus:	704.425			Durbin-Watson:	1.597	
Prob(Omnibus):	0.000			Jarque-Bera (JB):	145756.842	
Skew:	6.206	Prob(JB):	0.00			
Kurtosis:	82.149	Cond. No.	131.			
OLS Regression model between (Loyer and surface using Smf_fuction						
Dep. Variable:	loyer			R-squared (uncentered):	0.912	
Model:	OLS Adj.			R-squared (uncentered):	0.912	
Method:	Least Squares			F-statistic:	5664.	
Date:	Sat, 08 Jul 2023			Prob (F-statistic):	8.83e-290	
	coef	std err	t	P>		

Source: prepared by the research using python 3.9

plt. Plot (modelsmf. resid)

fig (08): resides values in classic regression model



Source: prepared by the research using python 3.9

- **Generate a random reside_ values**

```
# Perform Shapiro-Wilk test
statistic, p_value = stats.shapiro(y)
p_value
# Interpret the results
alpha = 0.05 # significance level
if p_value > alpha:
    print ("The data follows a normal distribution (fail to reject H0)")
else:
    print ("The data does not follow a normal distribution (reject H0)")
```

fig (09): Normality test of the depended variable(Loyer)

Machine learning theory vs classic approach in linear regression estimation

```
In [76]:
...: statistic, p_value = stats.shapiro(y)
...: p_value
...: # Interpret the results
...: alpha = 0.05 # significance level
...: if p_value > alpha:
...:     print("The data follows a normal distribution (fail to reject H0)")
...: else:
...:     print("The data does not follow a normal distribution (reject H0)")
The data does not follow a normal distribution (reject H0)
```

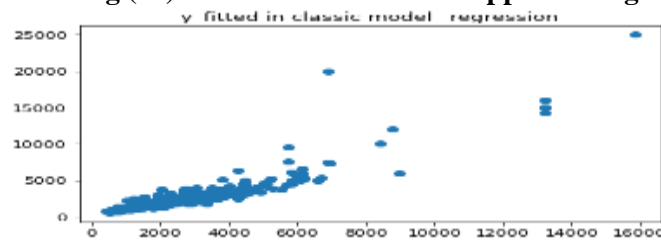
Source: prepared by the research using python 3.9

3.Diagnostic of fitted_ values

The goodness of fit test explains the quality of the model estimated, however we should include some statistic criteria in order to diagnostic the results, as homogeneity of the variance, normality distribution of the residue. multicollinearity of the independent variable's.

```
print(Fitted_values)
plt. scatter(Fitted_values)
plt. title ("fitted values distribution")
plt. show ()
```

fig (10): fitted values in classic approach regression



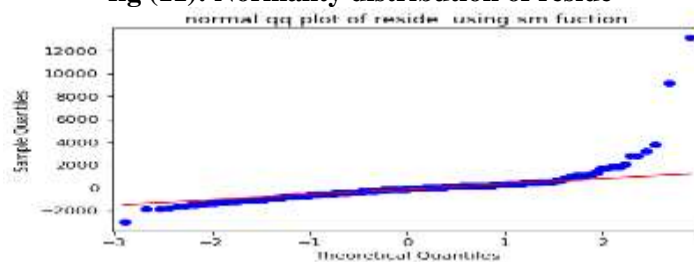
Source: prepared by the research using python 3.9

Diagnostic of residue

Normality distribution of residue

```
Import statsmodels.api as sm
smmodel=sm.OLS(y,x).fit()
reside=smmodel. resid
fig1=sm.qqplot(data=reside, line="q")
plt. title ("normal qq plot of reside using sm function")
plt. Show ()
```

fig (11): Normality distribution of residue

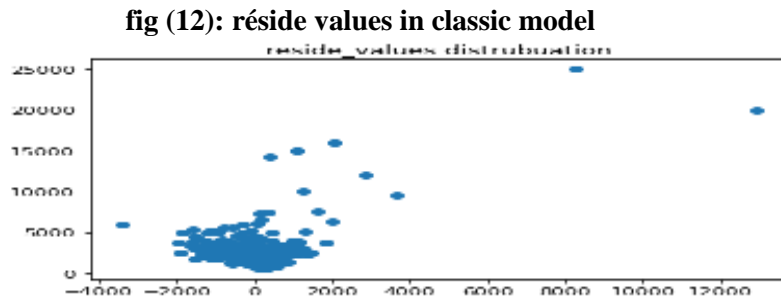


Source: prepared by the research using python 3.9

The equality of the residue distribution refer to diagnostic the nature of Rside_values in the model created. or when they cluster together. Our decision of homoscedasticity case will be taken when the residuals are equally distributed. but when are not we are in heteroscedasticity case.

2.reside_values distribution

```
print(Fitte_values)
plt. scatter(Fitte_values)
plt. title ("Fitte_values distribution")
plt. show ()
```



Source: prepared by the research using python 3.9

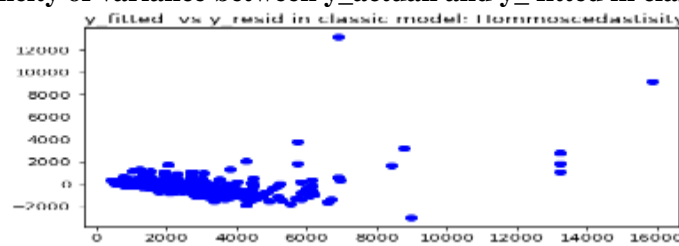
Homogeneity of variance

▪ Fitte_values vs resides values

Homogeneity of Variance is a statistic criterion that will be taken in order to detect the nature of variation between fitted values (predictive _values) and reside values. well, it does not mean that must be equal in all the time, but we interested at the degree of the variance over difference data points which should be the same. If we get null sum between the Fitte_values and the predict values, that mean we have zero reside _values, and oud model is perfectly build.

```
input
plt. Scatter (y_fittedclassic, y_residclassic)
plt. scatter (y_fittedclassic, y_residclassic, color='blue')
plt. title (" y_fitted vs y_resid in classic model: Homoscedasticity")
plt. show ()
```

Fig(13): homogeneity of variance between y_ actual and y_ fitted in classic regression model



Source: prepared by the research using python 3.9

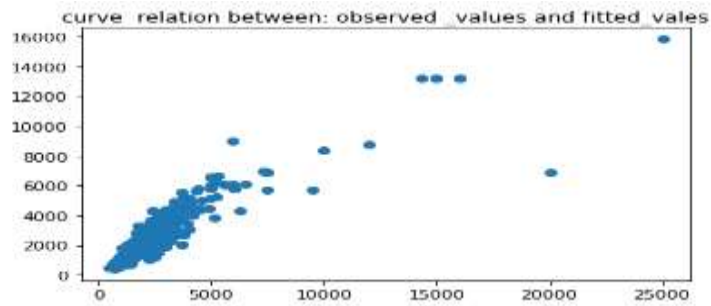
The points are localized and not strongly uniformly distributed. So, we can't confirm the case of homoscedasticity ho in this model (sm_model).

2.observed_values vs reside_values

Under the same theory of homogeneity of variance, and in order to refer the homoscedasticity or the heteroscedasticity Case we present the next scatter plot.

Fig(14): Homogeneity of variance in classic regression model

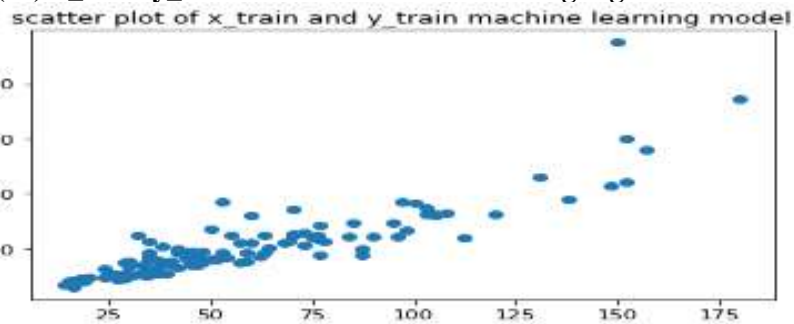
Machine learning theory vs classic approach in linear regression estimation



Source: prepared by the research using python 3.9

The points are localized and not strongly uniformly distributed. So, we can't confirmed the case of homoscedasticity h_0 in this model (sm_model).

Fig(15):x_train y_train trend in machine learning regression model



Source: prepared by the research using python 3.9

The main objective in this kind of estimation is to develop a model regression using the small repartition of data called x_{train} and y_{train} before make the final model estimated model using the teste data x_{test} , y_{test} .

4.Machine Learning approach regression

4.1Sm and smf functions using 20%of training data in Machine Learning regression

```

import statsmodels.api as sm
from sklearn. model selection import train_test_split
x_train, y_test, y_train, y_test=train_test_split(y, x,train_size=0.2,random_state=50)
MachineLearning_model=sm.OLS (x_train, y_train). fit()
MachineLearning_model. params
MachineLearning_model. summary ()
    
```

Table (02): OLS Regression Results using training data (x_train, y_train)

OLS Regression Results using training data at 20% of level			
Dep. Variable:	loyer	R-squared (uncentered):	0.932
Model:	OLS	Adj. R-squared (uncentered):	0.931
Method:	Least Squares	F-statistic:	1475.

Medini Atmane

Date:	Sun, 09 Jul 2023	Prob (F-statistic):	8.65e-65			
Time:	19:24:40	Log-Likelihood:	-862.98			
No. Observations:	109	AIC:	1728.			
Df Residuals:	108	BIC:	1731.			
Df Model:	1					
Covariance Type :	nonrobust					
	Coef	std err	t	P> t	[0.025 0.975]	
Surface(x)	34.8194	0.907	38.401	0.000	33.022 36.617	
Omnibus :	85.611	Durbin-Watson :	2.208			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	992.985			
Skew:	2.383	Prob(JB):	2.38e-216			
Kurtosis:	16.997	Cond. No.	1.00			
OLS Regression Model Result using test data at 70% of level						
Dep. Variable:	loyer	R-squared (uncentered):	0.911			
Model:	OLS	Adj. R-squared (uncentered):	0.911			
Method:	Least Squares	F-statistic:	4454.			
		Prob (F-statistic):	1.15e-230			
	Coef	std err	t	P> t	[0.025 0.975]	
Surface	38.7738	0.581	66.741	0.000	37.632 39.916	
Omnibus :	626.219	Durbin-Watson :	1.959			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	143587.042			

Source: prepared by the research using python 3.9

Table (03): Statistic Criteria in Machine Learning Regression process (y_test, x_test)

X_train values	20%	30%	40%	50%
Coefficient	38.1570	37.4673	38.2505	38.3068
Confidence_Interval	37.161-39.153	36.614-38.921	36.944-39.524	36.838-39.776
Statistical	75.262	64.378	59.082	51.345
P values	0.000***	0.000***	0.000***	0.000***
R_squared	0.912	0.916	0.915	0.906

Source: prepared by the research using python 3.9

The most R_square fitted is in 30% of training data, witch main the existence of perfect model between all the other models tested.

4.2. Diagnostic of the model estimated

In order to reach the optimal trade-off between the estimated experimental models, we have to make some relevant statistical measurements. In order to detect the model estimated at 30% of teste data in both data (x and y), we should know about the fitted and the reside values.

- **Fitte_values using tested data**

Fitte_values explain the new values of y_had from the equation estimated

Machine learning theory vs classic approach in linear regression estimation

$$(y_had=b0+38.7738*X_test)$$

- **Rside_values using tested data**

Rside_values explain the difference between actual and y_fitted values from the model. If the sum of residue equal to zero, that main attempt the perfect model between all the trained models.

4.3. Sm Testing Machine Learning Model based on 30% level of repartition

In this steop of our study and after accepting 30% level of data repartition, we are going to predict a new model regression based on machine learning philosophy, and on 70% of tested values of the data.

4.4. Machine Learning Model accepted

```
modell=sm.OLS (x_test, y_test). fit ()
modell.params
modell.summary()
```

table (04): OLS regression model accepted as 70% of Y_test and y_test

OLS Regression Results						
Dep. Variable:	loyer			R-squared (uncentered):	0.905	
Model:	OLS Adj.			R-squared (uncentered):	0.905	
Method:	Least Squares			F-statistic:	3625.	
Date:	Sat, 08 Jul 2023			Prob (F-statistic):	9.70e-197	
Time:	19:18:00			Log-Likelihood:	-	
3199.1						
No. Observations:	382			AIC:		
6400.						
Df Residuals:	381			BIC:		
6404.						
	coef	std err	t	P> t	[0.025	0.975]
surface(x)	39.1901	0.651	60.206	0.000	37.910	40.470
Omnibus:	537.266			Durbin-Watson:	1.921	
Prob(Omnibus):	0.000			Jarque-Bera (JB):	100109.748	
Skew:	6.934			Prob(JB):	0.00	
Kurtosis:	81.085			Cond. No.	1.00	

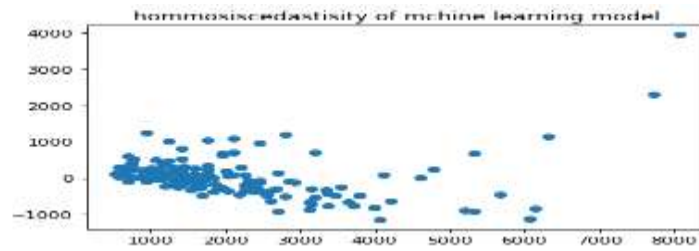
Source: prepared by the research using python 3.9

- **Homoscedasticity test**

In order to test the homogeneity of the variance in the model accepted, we build the next plot using the next algorithm.

```
MachineLearning_model.fittedvalues
MachineLearning_model.resid
plt.scatter(MachineLearning_model.fittedvalues,MachineLearning_model.resid)
plt.title("hommosiscedastisity of mchine learning model")
plt.show()
```

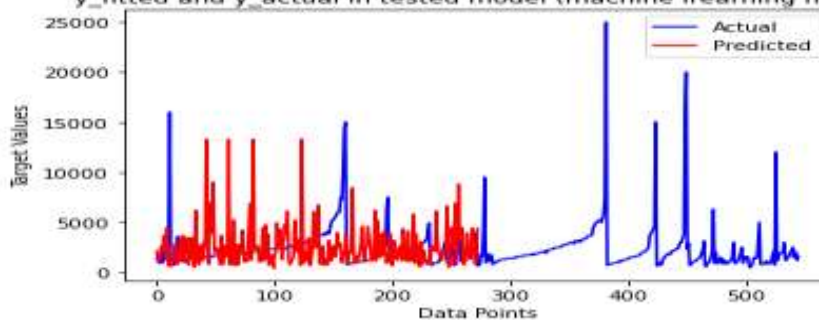
Fig(16):Homogeneity of variance in machine learning model



Source: prepared by the research using python 3.9

The point are not located in the same place, they are uniformly distributed, witch explain homoscedasticity case in the model estimated.

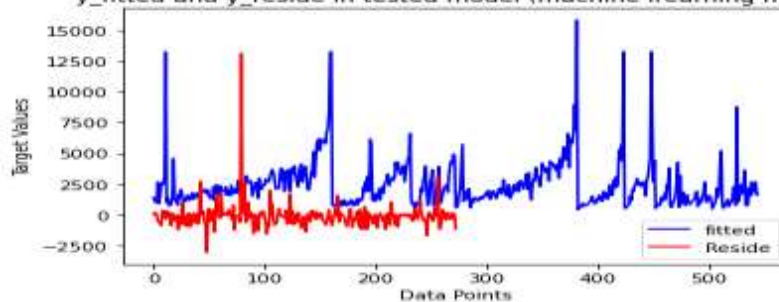
Fig (17): fitted values vs actual values in machine Learning model
y_fitted and y_actual in tested model (machine lrearning model)



Source: prepared by the research using python 3.9

We record through the following chart the nature of the distribution of the original data to the data that was estimated through the best model.

Fig(18): fitted values vs reside values in Machin learning model
y_fitted and y_reside in tested model (machine lrearning model)



Source: prepared by the research using python 3.9

In the figure, the nature of the residual distribution can be seen with respect to the original data.

5. Results of the comparative models

The comparative analysis of the regression models between the classical model and machine learning models revealed, that the results of the regression are linked to one level of analysis. Where the data was treated as an integrated unit. To reveal one equation that determines the kind and nature of relationship .While we were able, after adopting machine learning models,

Machine learning theory vs classic approach in linear regression estimation

to reveal a broader possibility to identify the assumed relationship between the two types of data, after dividing it into training and experimental data, which is the method that is appropriate for big data, as we reached a stronger representation - through the large coefficient of determination - For training data at 30%, which means a stronger explanation for the rest of the data, which constitutes 70%, and the following table explains it.

Table (): result of comparative estimation (classic & machine learning models)

	Classic model		Machine learning model
	Smf_fuction	smf_fuction	
Sloop	40.9712	40.9712	Sloop= 39.1901
Constant	-283.3784	/	/
R_square	82.2%	91.2%	90.5%
Adi_R_square	82.2%	91.2%	90.5%

Source: prepared by the research using python 3.9

6. Conclusion

The main objective of this research is the comparative analysis of the results of the regression analysis model according to two analysis theories, in classic analysis, the data is considered as one unit, that is divided into explanatory and dependent variables. However, this method may seem unrealistic, especially when it comes to Big data, where the number of variables is very large. Which necessitated addressing a different method aimed at adopting an experimental model through a training model, which is the technique known as the machine learning technique that somewhat simulates human capacity under the idea of artificial intelligence. also in classic method the estimation integrates all the data in one using the all the data selected (depended and in depended variable). But in machine learning model method we develop our model from two levels. The algorithm can have divided the selected data on two principal categories, train and test. So the results taken in classic model shoood of positive relation sheep between Loyer and Surface with positive sloop (surface=40.9712 and intercept= -2380378) using python software. With explanatory capacity of the determinant coefficient as 82.2%. Machine learning is the process which include and developed artificial intelligence by enabling systems to learn from data and make predictions or decisions based on patterns and relationships identified in the data. Well process showed more than model can be estimated as levels of depended variable's train. The most perfect regression model is in 70% of x_test and 30% of x_train (surface= **39.1901**) with **90.5** % of explanatory capacity of the determinant coefficient and p_value=.000***.

6. Bibliography List:

1. Devashree , M. (2023, 06 15). Retrieved from Linear Regression in Machine Learning: www.knowledgehut.com

2. Devashree , M. (2023, 06 15). *Linear Regression in Machine Learning: A Comprehensive Guide*. Retrieved from www.knowledgehut.com
3. EMMANUEL , J. (2020, 10 06). *Faire une régression linéaire avec R et avec python*. Retrieved from Stat Decision.
4. Jason, B. (2016, 03 15). *Machine Learning Mastery*. Retrieved from Linear Regression for Machine Learning: <https://machinelearningmastery.com/>
5. nikhilag, g. (No date). *Statistics with Python*. Retrieved from Statistics with Python
6. Partha Pratim Mandal, P., & Reza Rezaee. (2019). AEGC 2019: From Data to Discovery–Perth, Australia1Facies classification with different machine learning algorithm Anefficient artificial intelligence technique for improved classification. *AEGC Data to Discovery*, p. 02.
7. Pavan Vadapalli. (2022, 11 14). *6 Types of Regression Models in Machine Learning You Should Know About*. Retrieved from <https://www.upgrad.com/>
8. Ptreek , N. (no date). *Regression Analysis Using Artificial Neural Networks*. Retrieved from <https://www.scaler.com/>
9. Rohan Gupta, a. (2021, 01 29). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity* , p. 1317.
10. Samuel, L., & others, a. (2020, 11). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *journal homepage*.
11. Vishwa, P. (2020, 06). *Linear Regression Model for Machine Learning*. Retrieved from <https://towardsdatascience.com/>