LINEAR REGRESSION WITH MULTIPLE REGRESSORS

PERFORMING LINEAR REGRESSION IN PYTHON USING SCIKIT-LEARN

- Diptansu Poddar 210347
- Naman Mehrotra 210647
- Parthapratim Chatterjee 210705
- Soham Bharambe 210264

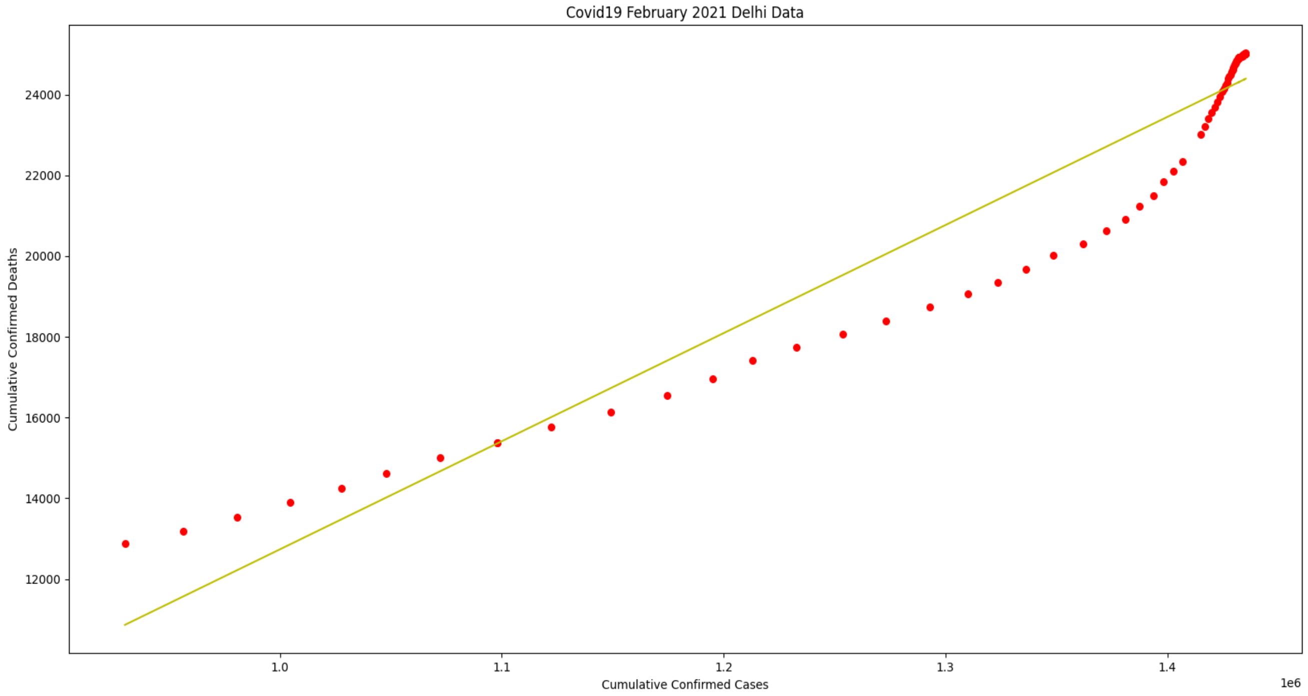
A BRIEF EXAMPLE OF SIMPLE REGRESSION

- We perform linear regression to express the linear relationship between variables.
- In this example we can foresee there exists a correlation between the number of confirmed cases and deaths due to covid19 in Delhi in February 2021.
- We used the popular python package skikit-learn to fit our data(Number of confirmed cases(independent variable) vs Number of deaths(dependent variable)

```
    Unnamed: 0                 Time  State  Confirmed  Recovered  Deaths  \
0            0  2021-04-22 22:46:02  Delhi     930179     831928   12887
1            1  2021-04-22 23:54:36  Delhi     956348     851537   13193
2            2  2021-04-23 23:24:43  Delhi     980679     875109   13541
3            3  2021-04-24 23:24:30  Delhi    1004782     897804   13898
4            4  2021-04-25 22:29:17  Delhi    1027715     918875   14248
..         ...                  ...    ...        ...        ...     ...
85          85  2021-07-11 19:02:20  Delhi    1435083    1409325   25015
86          86  2021-07-12 19:58:16  Delhi    1435128    1409417   25018
87          87  2021-07-14 13:56:28  Delhi    1435204    1409501   25020
88          88  2021-07-14 18:39:07  Delhi    1435281    1409572   25021
89          89  2021-07-15 19:34:29  Delhi    1435353    1409660   25022

    Active  New Cases
0    85364        NaN
1    91618    26169.0
2    92029    24331.0
3    93080    24103.0
4    94592    22933.0
..     ...        ...
85     743       53.0
86     693       45.0
87     683       76.0
88     688       77.0
89     671       72.0

[90 rows x 8 columns]
```



Intercept value = [-14037.2764932]
Slope value = [0.02677429]
r2 value = 0.9313659523193988
Standard Error of regression slope = [0.02892491]

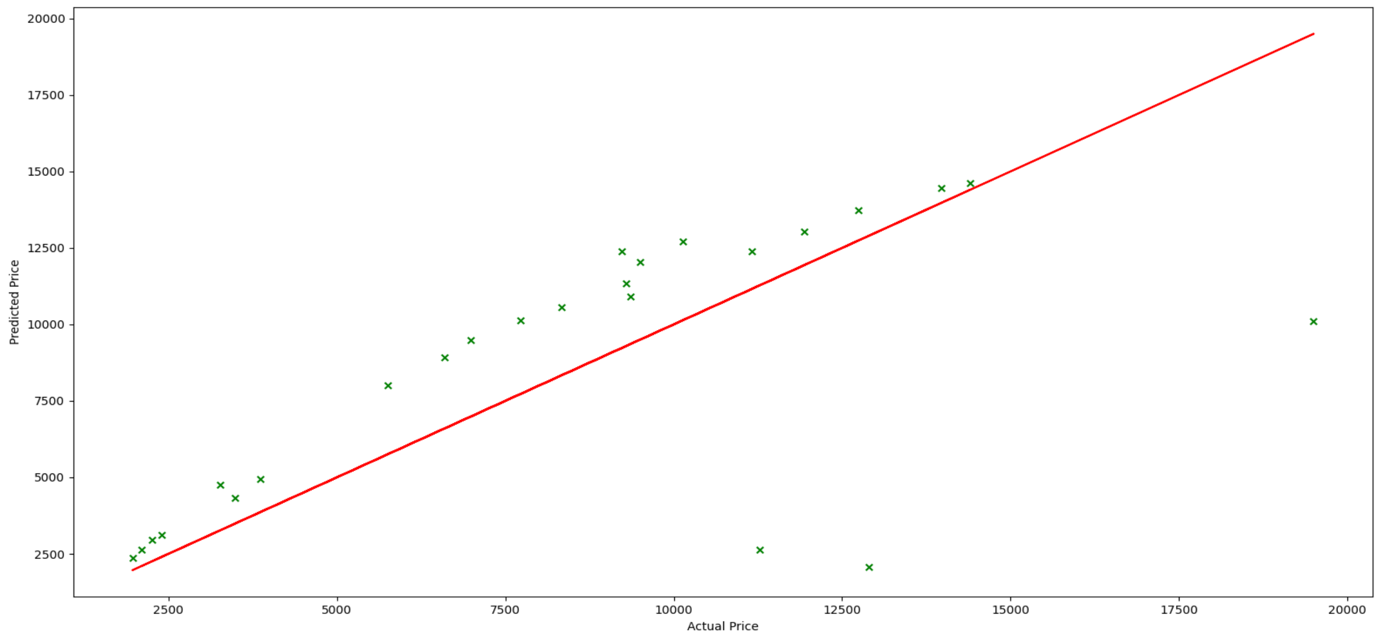
As we can see from the plotted data the relationship is mostly linear for a certain interval but when the total number of confirmed deaths increase the curve shows a steep rise in the end(At around 138000 confirmed cases) deviating from the linear behaviour. We can assume greater number of cases puts additional burden on the health care system causing death percentage to shoot up while lesser proportion of affected individuals gets access to quality medical treatment

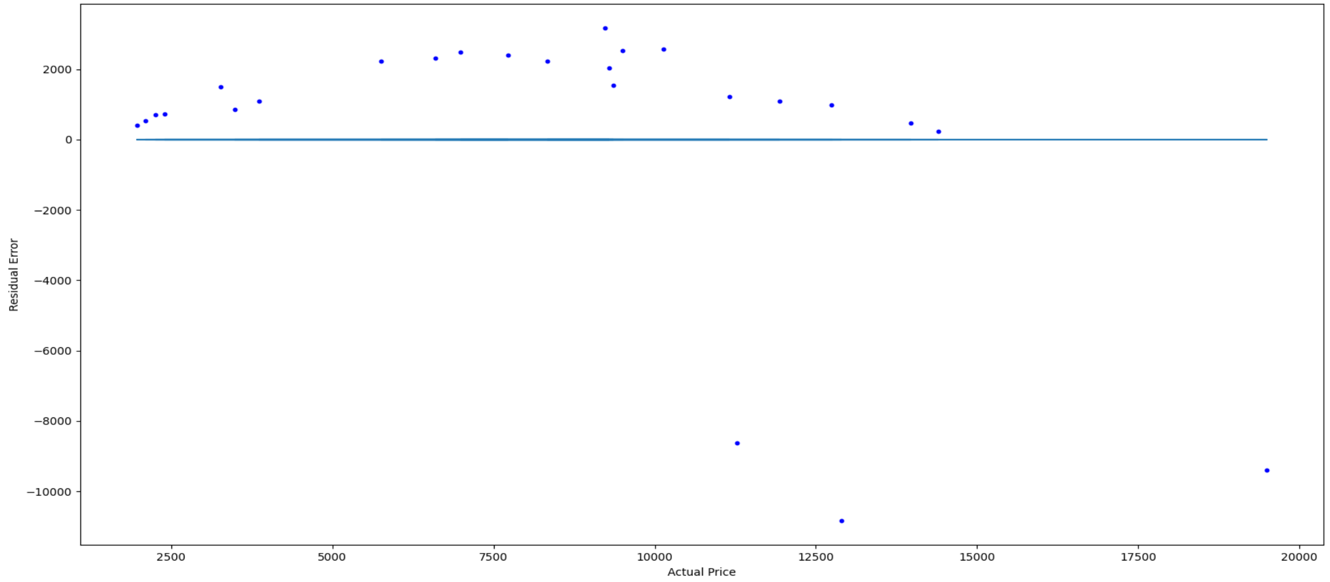
USING REGRESSION TO EVALUATE PRICE OF HEALTH INSURANCE
LINEAR REGRESSION WITH MULTIPLE REGRESSORS

- In the below example we have three independent variable age, bmi and number of children of the person which will be used to determine the dependent variable Insurance charge.
- In this we have split the data set into training and testing sets where 20 percent of the data is used for testing and 80 percent is used for training the model.
- We have plotted the suggested price line for the testing data set while showing the actual price
- This model then can be used to evaluate the health insurance price of a person with the given parameters of age, bmi and number of children.
- Categorical variables have been neglected/dropped and only male non-smokers residing in the north-east region has been considered.
- The algo even shows us the computed price when we enter requisite details.

```
Enter age: 45
Enter bmi: 23.55
Enter children: 1

Insurance Price as per users input:
[10081.98357339]
```






```
      age     bmi  children
8      37  29.830         2
10     25  26.220         0
17     23  23.845         0
44     38  37.050         1
60     43  27.360         3
...   ...     ...       ...
1294   58  25.175         0
1296   18  26.125         0
1315   18  28.310         1
1318   35  39.710         4
1325   61  33.535         0

[125 rows x 3 columns]


Intercept =
-3809.9095124767628

Coefficients =
[279.14545162  21.57634111 822.22492955]

r2 score for testing set =  0.3110662695971691
```

- We have plotted the graphs of predicted price vs actual price and residual error vs actual price.
- To generate the model, we have ignored the influence of categorical values such as sex, smoking habits and location
- As it is visible from the graph that it had very less data points hence our model is under trained and therefore gave an R2 value of 0.311.

DEALING WITH CATEGORICAL VARIABLES

- The r^2 value of the previous model is not satisfactory.
- To improve our model we can convert the categorical values using the get_dummies() method part of pandas package.
- The latest model is much better one which can determine the ideal price of insurance by crunching many more factors which take discreet value such as sex, smoking habits and region.
- As we can see this one has much better r^2 value.

```
      age     bmi  children      charges  female  no  northeast  northwest  \
0      19  27.900         0  16884.92400       1   0          0          0
1      18  33.770         1   1725.55230       0   1          0          0
2      28  33.000         3   4449.46200       0   1          0          0
3      33  22.705         0  21984.47061       0   1          0          1
4      32  28.880         0   3866.85520       0   1          0          1
...   ...     ...       ...          ...     ...  ..        ...        ...
1333   50  30.970         3  10600.54830       0   1          0          1
1334   18  31.920         0   2205.98080       1   1          1          0
1335   18  36.850         0   1629.83350       1   1          0          0
1336   21  25.800         0   2007.94500       1   1          0          0
1337   61  29.070         0  29141.36030       1   0          0          1

      southeast
0             0
1             1
2             1
3             0
4             0
...         ...
1333          0
1334          0
1335          1
1336          0
1337          0

[1338 rows x 9 columns]
```

Intercept = 11343.689963450903

Coefficients = [257.49024669 321.62189278 408.06102001 242.15306559 -23786.48604536 903.03300778 506.93644423 -135.34287187]

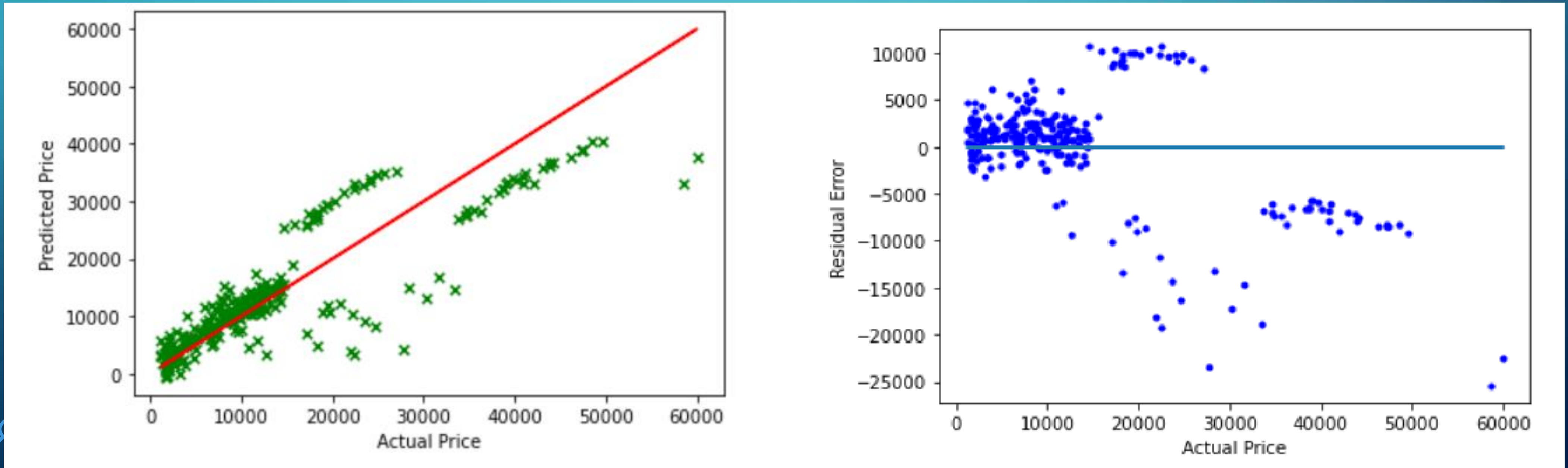
r2 score for testing set = 0.7623311844057112

SOME OBSERVATIONS

- Better r^2 value
- More useful as it can suggest prices for more categories of people based on categorical factors
- By looking at the coefficients we can see the nature of relationship. People who don't smoke and/or live in the southeast region will have lesser insurance prices.

Detecting Heteroscedasticity

To perform linear regression using the method of least squares we assume the data to be homoscedastic. But here when we plot the residual error vs actual price we see the points deviate from homoscedastic behaviour at around an actual price of 15000.



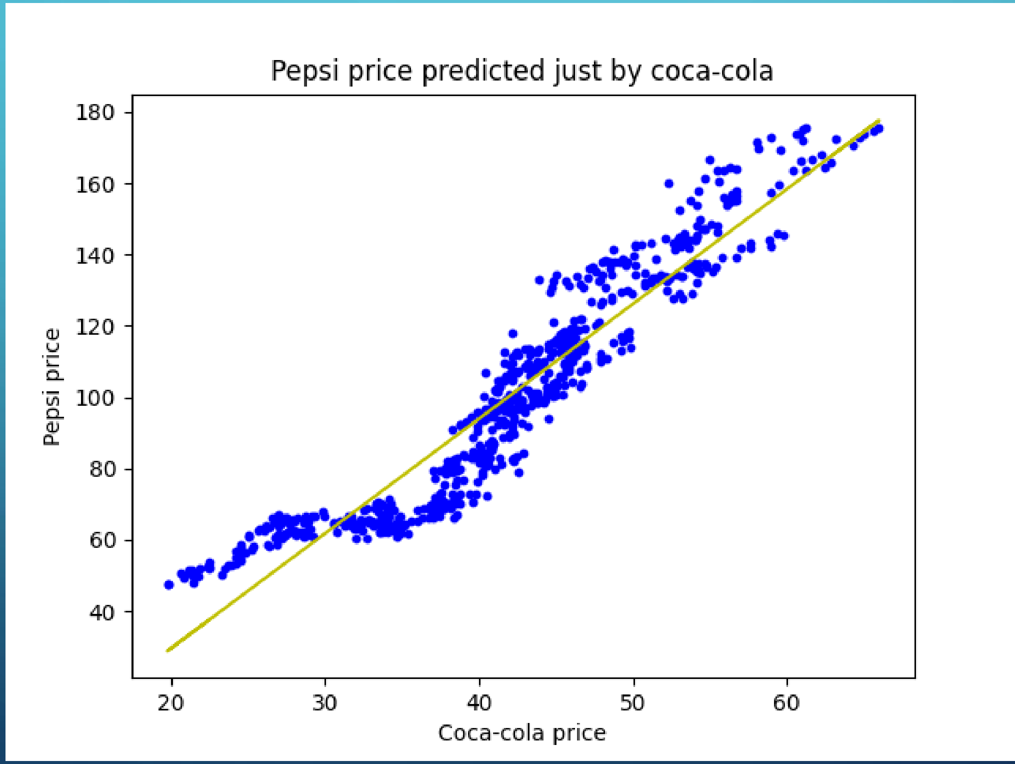
MEAN REVERSION

Explanation and Real-world applications

- Mean Reversion is an algorithm-based trading strategy often used by momentum-based trading firm and mid-frequency trading firms which involve comparing 2 similar equities of the same sector which have very similar price movement with respect to each other.
- This relies on the fact that the stocks with similar fundamental values will always converge to their regression mean .This will thus create a risk-free arbitrage strategy and will be a excellent source of diversification irrespective of the directions of the market.
- In the coming slides , We have demonstrated the use of regression analysis in comparing Coco-cola , Pepsi and SPY ETF.

SINGLE INDEPENDENT VARIABLE LINEAR REGRESSION

- The below code and chart shows the relations between the share price of Coca-Cola and Pepsi co . We can clearly notice that they follow the mean reversion and always converge to the mean (which is the regression line)
- This will create amazing short term opportunities which can produce better returns than the SPY even during the market sell off.
- We also notice that R2=0.89. We can improve it further. We shall include the variable of SPY as the second variable as it will help us include the market conditions even better as It represents the overall health of the market .



MULTIPLE VARIABLE LINEAR REGRESSION

- Here , The 2 independent variables are the price of Coca-Cola and SPY ETF, and the dependent variable is Pepsi price .
- We will obtain a 2-D graph/plane for this. We get a surprisingly very high value of R2 =0.96 ; This confirms that similar industry stocks move such that they will return to their mean .
- Addition of SPY , will help to take account of health of the market and prevent any extra error.
- The above result can be explained as both the companies are in their mature stage and do not have growth factor involved , So this is often used between commodity producing sectors

	p	c	s
0	53.889999	22.465000	90.669998
1	51.099998	21.309999	84.370003
2	50.439999	21.645000	84.050003
3	52.040001	21.930000	87.389999
4	51.490002	21.200001	83.330002
..	...	...	...
694	174.850006	65.559998	417.269989
695	173.860001	65.029999	429.059998
696	170.660004	64.309998	392.750000
697	163.649994	61.200001	391.859985
698	165.600006	62.860001	396.920013

[699 rows x 3 columns]

[1.09327951 0.21791734]

[3.22280723]

R2 value obtained by comparing pepsi with coca-cola and spy 0.9675867028976729

R2 value obtained by comparing pepsi with just coca-cola 0.8958902209787118

ACKNOWLEDGEMENTS

- We are profoundly grateful to our mentors Sandipan Mitra and Rohan Kumar for guiding us through each and every step and resolving our doubts.
- We are also grateful to our institute for giving us this opportunity to work on and learn about regression analysis at an early stage of our college journey.

THANK YOU!