Multiple linear regression (MLR)

1.

Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Also, MLR can do the classification. For example, MLR can be used to predict the sale of products in the future based on past buying behavior. The difference between simple and multiple is that the multiple will analysis two or more variables and simple linear regression just use only one independent variable so the result of MLR will be more accurate.

The algorithms are divided into five parts.

(a) Describe and explore the dataset: In this part, we explore and get a common knowledge about the dataset itself.
(b) Data visualization: We use visualization to demonstrate the relationship between each variable and filter the invalid data.
(c) Calculate the MLR: We use the sklearn method to calculate the formula.
(d) Prediction: It is an extension about the part c. In this part, we use observed data and predict data to test the algorithm.
(e) Note: We use OLS method to calculate the formula. The result is different.

From our conclusion, we find that the numbers of convenience shop and the distance to MRT station will influence the house price. Multiple linear regression can predict the trend about the house price, but it will still be affected by other variables which this dataset not included.

2.

This formula of MLR is Y=A+B1X1+B2X2+…+BnXn

Y is a dependent variable and X is an independent variable. Bis a coefficient, A is a intercept.

In this dataset, Y is the price about house of unit area. We finally get two valid variables:

X1= distances to MRT station

X2= and the number of convenience shops

The formula we conclude that:

Y=39.1299-0.0056(distance to the nearest MRT station) +1.1976(numbers of convenient stores)

This formula demonstrates that a less distance to MRT station and more numbers of the convenient stores will increase the house price. The coefficient of distance to MRT approach to 0, so from the importance point, the effect of numbers of convenience stores will be more significant.

3.

As the formula above, we can apply it in two parts and two restrictions:

I: Apply:

(a) Prediction:
We can predict the house price in an area according whether this area will build more convenient stores or open the new MRT line.

(b) Estimation:
We can use this formula to do a preliminary judgement about the house price and make a prediction on price potential.

II: Restrictions:

(a) Sample size:

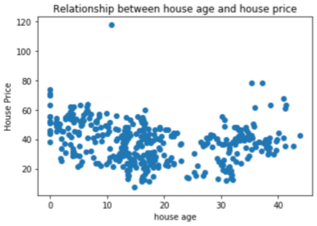
The sample size of the dataset is small so it may lead some error on the real situation.

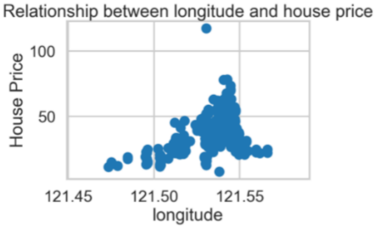
(c) Variable:
This formula has a little variable that we may ignore other factors which will affect the house price such as population density, crime rate, national conditions.

4.

Limitation:

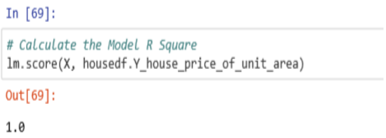
(a) Factor incomplete:
The dataset provides seven variables, just two variables are valid. We considered to put the house age and longitude into the formula but the relationship between those variables and house price is not a linear regression. They are curves.



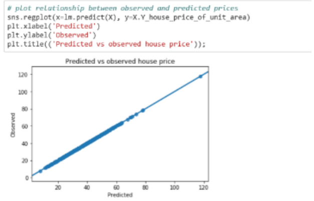


The effective linear relationship in this dataset is not so much that our formula may lead some limitation on prediction or estimation.

(b) Weak reference in real world



```
In [69]:
# Calculate the Model R Square
lm.score(X, housedf.Y_house_price_of_unit_area)
Out[69]:
1.0
```



```
# plot relationship between observed and predicted prices
sns.regplot(x=lm.predict(X), y=X.Y_house_price_of_unit_area)
plt.xlabel('Predicted')
plt.ylabel('Observed')
plt.title(('Predicted vs observed house price'));
```

This is the fitting model we use sklearn method to calculate. It is a perfect model so we think that this formula doesn't have a reference value in a real world.

We try to use OLS method the calculate the formula, but the R square is just 0.4, this is an index which means this formula has a weak reference.

5.

Libraries:

- Numpy

- Pandas
- Seaborn
- Matplotlib.pyplot
- Statsmodels.api
- Statsmodels.formula.api
- sklearn

Methods/functions:

- Print
- Plot
- Fit
- Prediction
- Visualization
- Im
- Observe

6.

a:

We use prediction method to do the evaluation.

Libraries:

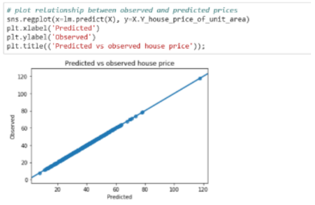
- Sklearn.linear_model

Method/Function:

- Plot
- Histogram

b:

We use plot histogram method to test the observe situation and prediction situation. The fit model is almost 100 percent fitting.



```
# plot relationship between observed and predicted prices
sns.regplot(x=lm.predict(X), y=X.Y_house_price_of_unit_area)
plt.xlabel('Predicted')
plt.ylabel('Observed')
plt.title(('Predicted vs observed house price'));
```

c:

It is a perfect dataset, so it has no value on reference.

7.

Section	Links	Notes

1	https://www.investopedia.com/terms/m/mlr.asp	Definition about MLR
2	https://en.wikipedia.org/wiki/Ordinary_least_squares	Definition about OLS
3	https://www.hindawi.com/journals/amete/2012/894714/	Application about MLR
4	https://pandas.pydata.org/pandas-docs/stable/getting_started/dsintro.html	Dataframe issue
5	https://www.kaggle.com/quantbruce/real-estate-price-prediction	Data source