

Linear Regression

ເນັດງາງ \downarrow ເນັດງານ 1 ເນັດ 0: ນັກ ລົມ, ອົກສົກ ໄກນ y

\downarrow slope, coeff \downarrow intercept

ສົນກາຣ໌ວິນດາຣາງ: $y = mx + b$

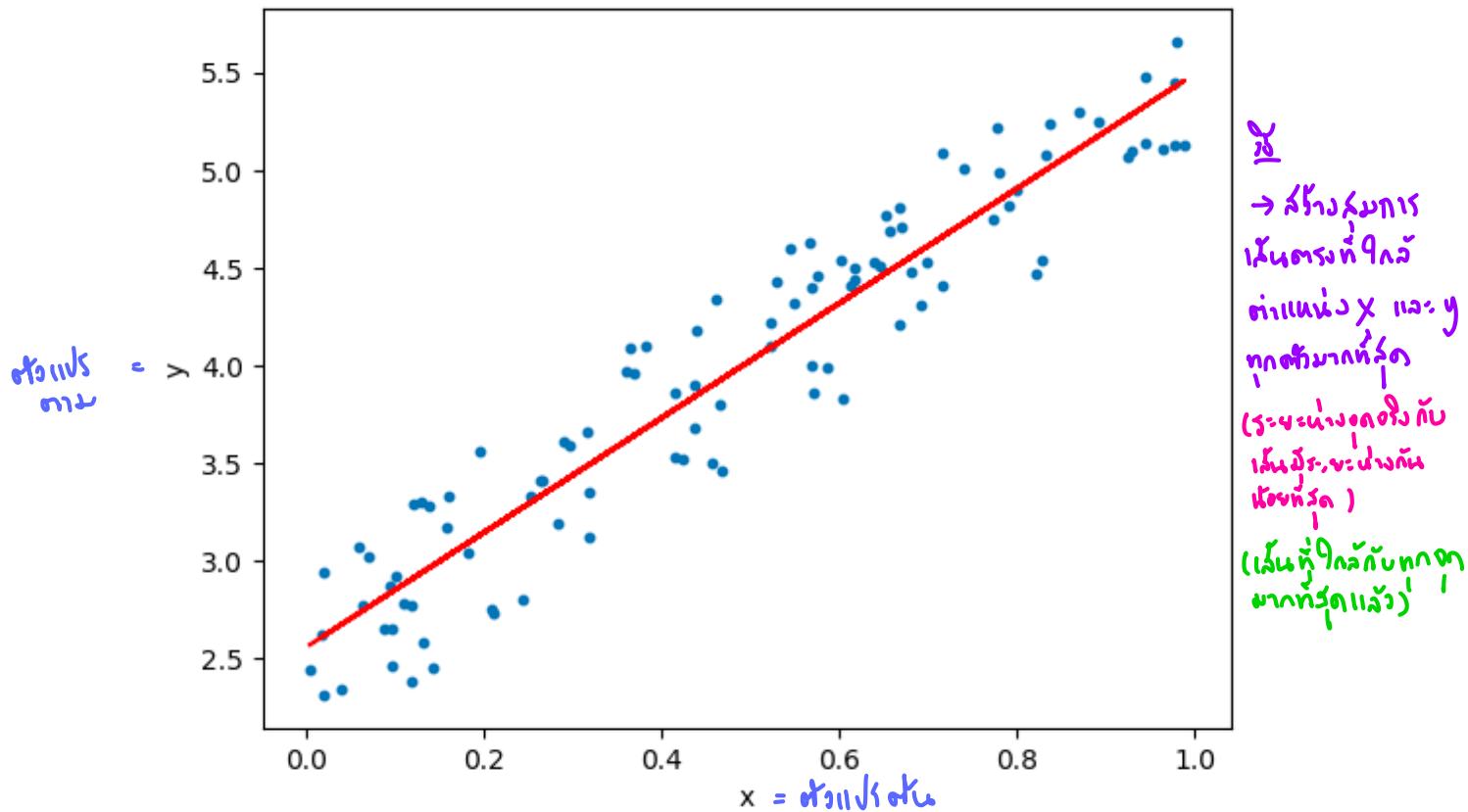


image from: [Linear regression hypothesis testing: Concepts, Examples](#)

- by  **scikit-learn** [scikit-learn \(LinearRegression\)](#) \rightarrow ບໍ່ຈະ SK-learn ໃຈນ multiLR ແກ້ໄຂLR ເຊັ່ນ ດີກັບ
ແກ່ຈຸດກັບທີ່ 1
- Linear Regression ເປັນເທົນີກພື້ນຖານສໍາຫັບການພຍາກຮົນຄ່າຕົວເລີຂໂດຍໃຊ້ຄວາມສັນພັນຮ່ວ່າງ
ຕົວແປຣ ໂດຍການຫາເສັນຕຽນທີ່ ແນະສນໃນການແສດງແນວໂນັມຂອງຂ້ອມລຸ, **ກົດກັບ ຂົບນູ້ທີ່ ເປັນຕົວ ລົກທີ່ 1**.
- $Y = \text{coefficients} * X + \text{intercept}$ \therefore ຜົກສົກ coeff, intercept
ມີດີນິຕິ ເນັດງານທີ່ 1 (ຮະບະດີເກີນ
ກົດກັບທີ່ 1)

Understanding Linear Regression

ເປັນເທົນີກພື້ນຖານສໍາຫັບການພຍາກຮົນຄ່າຕົວເລີຂໂດຍໃຊ້ຄວາມສັນພັນຮ່ວ່າງຕົວແປຣ ໂດຍການຫາເສັນຕຽນທີ່
ແນະສນໃນການແສດງແນວໂນັມຂອງຂ້ອມລຸ

Simple Linear Regression X ຜົກສົກ

Simple linear regression ມັງເນັນໄປທີ່ ຕົວແປຣອີສະແບນເດືອນແລະ ຕົວແປຣຕາມແບນເດືອນ ສູຕຣ໌ສໍາຫັບເສັນ
ຄດຄອຍຄືວ່າ:

$$y = mx + b$$

ໂດຍທີ່ m ຄືວ່າຄວາມລາດຊັ້ນແລະ b ຄືວ່າຈຸດຕັດຂອງເສັນ

Multiple Linear Regression

× มากกว่า 1 ตัว (叫做 feature) e.g. จำนวน หน. ร่างกาย..

Multiple linear regression ขยายความความคิดไปยังตัวแปรอิสระมากกว่าหนึ่งตัว สูตรกล้ายเป็น:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

โดยที่ b_0 คือจุดตัดและ b_i คือค่าสัมประสิทธิ์ที่ i

- ตัวที่ y ซึ่งมีจุดเด่น
- ตัว coeff คือค่าสัมประสิทธิ์ที่ x มีผล

สร้าง DataFrame

- เรียกใช้ pandas

```
import pandas as pd
```

- อ่านไฟล์ `california_housing_train.csv` จาก `sample_data` โดยคลิกขวาที่ชื่อไฟล์ และเลือก `Copy path`

```
data_train = pd.read_csv('/content/sample_data/california_housing_train.csv')
```

```
data_train
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	median_income
0	-114.31	34.19	15.0	5612.0	1283.0	1015	5.0
1	-114.47	34.40	19.0	7650.0	1901.0	1129	6.0
2	-114.56	33.69	17.0	720.0	174.0	333	7.0
3	-114.57	33.64	14.0	1501.0	337.0	515	8.0
4	-114.57	33.57	20.0	1454.0	326.0	624	9.0
...
16995	-124.26	40.58	52.0	2217.0	394.0	907	12.0
16996	-124.27	40.69	36.0	2349.0	528.0	1194	13.0
16997	-124.30	41.84	17.0	2677.0	531.0	1244	14.0
16998	-124.30	41.80	19.0	2672.0	552.0	1298	15.0
16999	-124.35	40.54	52.0	1820.0	300.0	806	16.0

17000 rows × 9 columns

- สำรวจ `data_train`

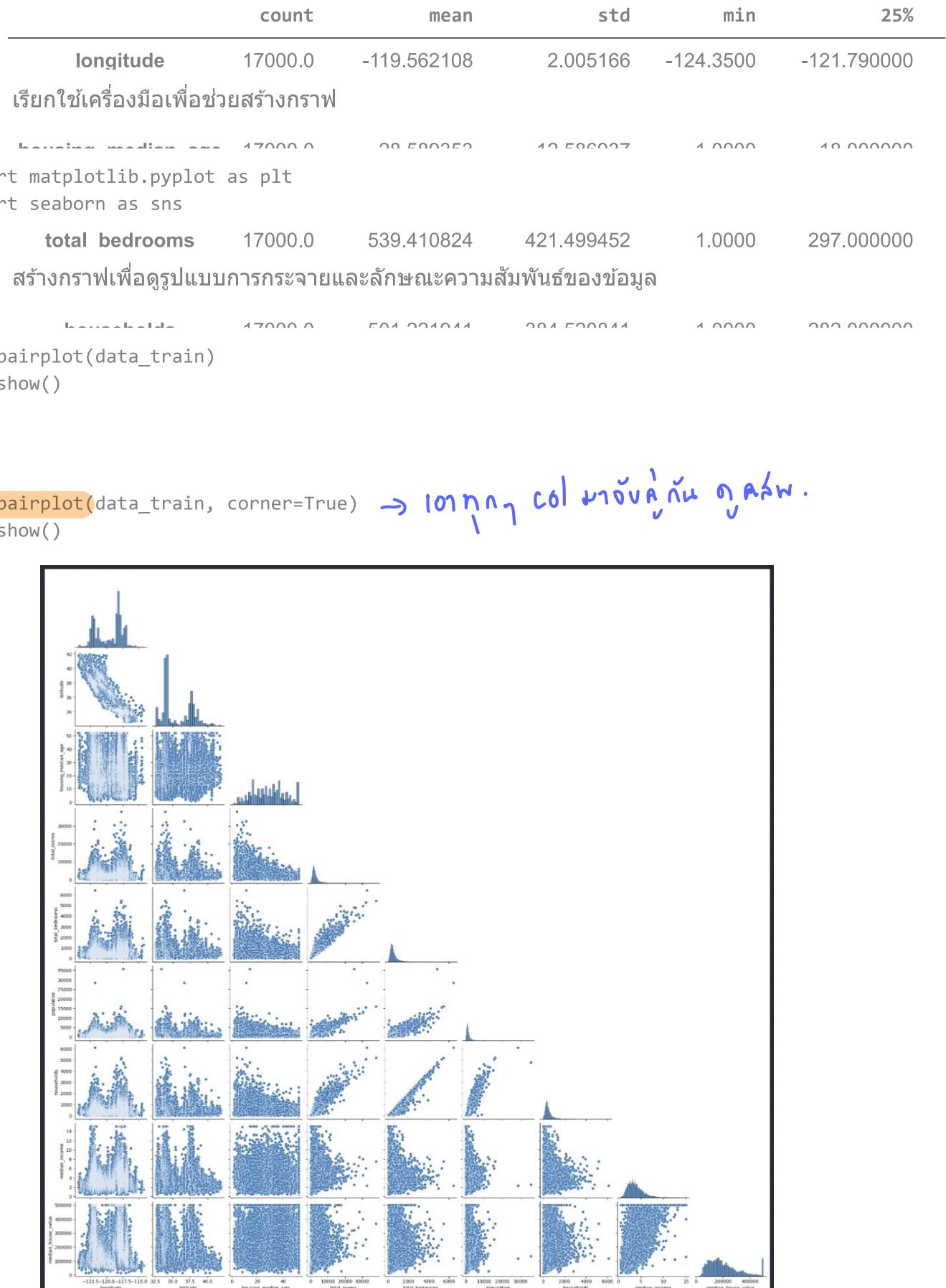
```
data_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17000 entries, 0 to 16999
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   longitude        17000 non-null   float64
 1   latitude         17000 non-null   float64
 2   housing_median_age 17000 non-null   float64
 3   total_rooms      17000 non-null   float64
 4   total_bedrooms   17000 non-null   float64
 5   population       17000 non-null   float64
 6   households       17000 non-null   float64
 7   median_income    17000 non-null   float64
 8   median_house_value 17000 non-null   float64
dtypes: float64(9)
memory usage: 1.2 MB
```

```
data_train.describe()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	
count	17000.000000	17000.000000	17000.000000	17000.000000	17000.000000	17
mean	-119.562108	35.625225	28.589353	2643.664412	539.410824	1
std	2.005166	2.137340	12.586937	2179.947071	421.499452	1
min	-124.350000	32.540000	1.000000	2.000000	1.000000	
25%	-121.790000	33.930000	18.000000	1462.000000	297.000000	
50%	-118.490000	34.250000	29.000000	2127.000000	434.000000	1
75%	-118.000000	37.720000	37.000000	3151.250000	648.250000	1
max	-114.310000	41.950000	52.000000	37937.000000	6445.000000	35

```
data_train.describe().T
```



▼ สร้าง Linear Regression

- เรียกใช้ LinearRegression library

```
from sklearn.linear_model import LinearRegression
```

- ในที่นี่ต้องการทำนาย median_house_value → ห้อง. คาดการณ์ราคาบ้าน

1. (เตรียม)
2. ตั้งค่า: y_train = data_train['median_house_value']
3. ตั้งค่า: x_train = data_train.drop('median_house_value', axis=1) → เป็น multi LR แนว X หลายตัว (9 ตัว)
x_train = x หลายตัว
y_train

0 66900.0
1 80100.0
2 85700.0
3 73400.0
4 65500.0
...
16995 111400.0
16996 79000.0
16997 103600.0
16998 85800.0
16999 94600.0
Name: median_house_value, Length: 17000, dtype: float64

x_train

inplace = True

ก็คือ: ห้ามแก้ไขในตัวเดิม (ให้เป็น False)

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population
0	-114.31	34.19	15.0	5612.0	1283.0	1015
1	-114.47	34.40	19.0	7650.0	1901.0	1129
2	-114.56	33.69	17.0	720.0	174.0	333
3	-114.57	33.64	14.0	1501.0	337.0	515
4	-114.57	33.57	20.0	1454.0	326.0	624

- สร้างโมเดล Linear Regression แบบฝึกหัด ชั้นปี 9 ชั้นปี 9

```
model_lr = LinearRegression()
```

- เริ่มการเรียนรู้

model_lr.fit(X_train, y_train) จัดข้อมูล

LinearRegression
LinearRegression() - ต้องรู้ว่ามีโมเดล หรือไม่ (มีตัวบ่งบอกแล้ว)

- ดูผลลัพธ์ ค่า coefficients (ค่าเส้น)

จัดส่ง X (ค่าตัวแปร)

```
model_lr.coef_
```

array([-4.31396373e+04, -4.29256731e+04, 1.15069493e+03, -8.37825121e+00, 1.17648543e+02, -3.84887721e+01, 4.54360026e+01, 4.05070684e+04])

X_train.columns ชื่อตัวแปร

Index(['longitude', 'latitude', 'housing_median_age', 'total_rooms', 'total_bedrooms', 'population', 'households', 'median_income'],
dtype='object')

- ดูค่า intercept (ค่าตัดส่วน y) คือค่าตัวคงที่

```
model_lr.intercept_
```

-3620600.8929739078 → รากฐานของผลลัพธ์ เมื่อใส่ค่า X ค่า 0 (X เป็น 0)

Predictions

ใช้ model ในการทำนายผล

ເລື່ອງທ່ານລັບໜັງ train

- ອ່ານໄຟລ໌ `california_housing_test.csv` ຈາກ `sample_data` ເພື່ອໃຊ້ໃນການທົດສອບ ໂດຍຄລືກຂວາທີ່
ຊື່ໄຟລ໌ ແລ້ວເລືອກ `Copy path`

```
data_test = pd.read_csv('/content/sample_data/california_housing_test.csv')
data_test.head()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	h
0	-122.05	37.37	27.0	3885.0	661.0	1537.0	
1	-118.30	34.26	43.0	1510.0	310.0	809.0	
2	-117.81	33.78	27.0	3589.0	507.0	1484.0	
3	-118.36	33.82	28.0	67.0	15.0	49.0	
4	-119.67	36.33		19.0	1241.0	244.0	850.0

- ໃນທີ່ນີ້ຕ້ອງການທຳນາຍ `median_house_value`

```
y_test = data_test['median_house_value']
X_test = data_test.drop('median_house_value', axis=1)
```

- ທຳນາຍຜລຈາກຊຸດຂອ້ມລ

```
predictions = model_lr.predict(X_test)
predictions
```

ກົດໄໝ test ທີ່ໄດ້ຮັບ X ທີ່ໄດ້ຮັບ
ຈົ່ງເປັນຕົວຂອງ col ແລ້ວ ພັນຍາວົງກົນ train
(ກົດໄໝຮັບ col ລຳຕັບຕົວທົງກົນ train ດັ່ງ)

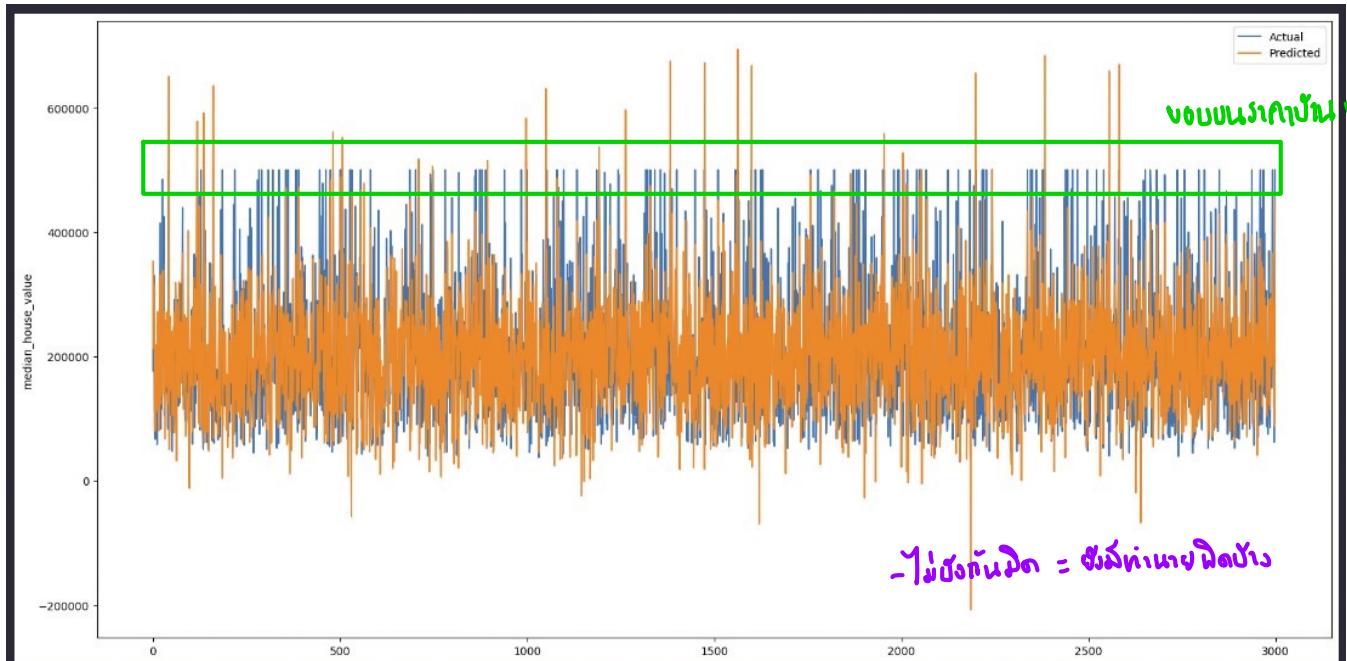
ຈຳລວງທີ່
ກ່າວປ້ານທີ່: array([352812.31112454, 212717.70074518, 272344.69951888, ...,
88220.58643733, 146374.67481457, 456779.61107787]) → 3000 ລາຍການ

data_test.shape **ກົດໄໝ test**
(3000, 9) **3000 ກ່າວ** **9 col.**

- ລອງດູເປັນກາຟເປົ້າຍນເທີຍນ

```
plt.figure(figsize=(20,10))
sns.lineplot(data=y_test, label='Actual')
sns.lineplot(data=predictions, label='Predicted')
plt.show()
```

ກົດໄໝເກີນໄພເປົ້າຍນນີ້ເພື່ອ:



‐ Evaluating Model Performance

వార్షికమైని.

mean absolute error
 $MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$

$$2. \text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

↓
square
(square root)

↓
mean of error

3. Root MSE = $\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$

4. R-squared (Coefficient of Determination)

R-squared หรือค่าหมายความสัมพันธ์ เป็นตัวชี้วัดทางสถิติที่ให้ข้อมูลเกี่ยวกับความเหมาะสมของโมเดล คดถอย มันแสดงถึงสัดส่วนของความแปรปรวนในตัวแปรตามที่สามารถทำนายได้จากตัวแปรอิสระ กล่าวอีกนัยหนึ่ง R-squared แสดงให้เห็นความสามารถของตัวแปรอิสระในการอธิบายการแปรปรวนในตัวแปรตาม

Formula

The formula for calculating R-squared is:

ເປັນພາກທ່າງງົດໃນອົງກົດ
ແລ້ວໄປໄຫຼື່ອງ
ອະລັດຕູ້ປິດຫົວໜ້າ 2 ມອດ
ຈະມອດທີ່ມີກ່າວ

మంగः

R-squared = 1 - (SSR / SST) ก็จะอยู่ bt. 0-1

Where:

- SSR (Sum of Squared Residuals) = the error សរុបកន្លែង រក្សាសម្រាប់ 2
(តើម.នូវ bt. ការណែនាំ នឹង ការរួមចាប់)

$$SSR = \sum (y_i - \hat{y}_i)^2$$

- y_i คือค่าตัวแปรตามที่สังเกตได้สำหรับจุดข้อมูลที่ i ค่าจริง
 - \hat{y}_i คือค่าที่ทำนายได้สำหรับตัวแปรตามสำหรับจุดข้อมูลที่ i

SSR แสดงถึงความแปรปรวนที่ไม่สามารถอธิบายได้ด้วยเส้นก่อกรอย่างเดียว

- SST (Total Sum of Squares)

$$SST = \sum (y_i - \bar{y})^2$$

- ² និង 1 នៅក្នុង:
- □² នៅក្នុង សង្គមភីរោង
- e.g. $0.5^2 \rightarrow 0.25$
- នៅក្នុងការ > 1 នៃ □² នៅក្នុង
- e.g. 0.98^2 និង 1.01^2 , 0.99^2
- និងការកូរ error ដែល
 បានក្នុង

- y_i คือค่าตัวแปรตามที่สังเกตได้สำหรับจุดข้อมูลที่ i ค่าจริง
- \hat{y}_i คือค่าเฉลี่ยของค่าตัวแปรตามที่สังเกตได้ ค่าเฉลี่ย

SST แสดงถึงการกระจายทั้งหมดของจุดข้อมูลรอบค่าเฉลี่ย โดยไม่สนใจความสัมพันธ์ใดๆ กับตัวแปรอิสระ

Interpretation

$$Q_{\text{total}} = 100\% \quad 0 = 100\% (0\%)$$

ค่า R-squared อยู่ในช่วงระหว่าง 0 ถึง 1 ค่าที่มี R-squared ใกล้เคียง 1 จะเป็นตัวแบบที่เหมาะสมกับข้อมูลมากที่สุด นี่คือวิธีการอธิบายค่า R-squared:

- R-squared = 1: ไม่เดลอธิบายการแปรปรวนในตัวแปรตามอย่างสมบูรณ์
- R-squared > 0.5: ไม่เดลถือเป็นตัวแบบที่เหมาะสม
- R-squared < 0.5: ไม่เดลอาจไม่เหมาะสมและอาจไม่สามารถอธิบายความแปรปรวนได้มากนัก

วัดค่า R^2

เขียนภาษา Python
 $\text{model_lr.score(X_test, y_test)}$ หาค่า R^2 ที่ได้รับ

0.6195057678312047 \rightarrow $R^2 = 61.95\%$

```
from sklearn.metrics import mean_absolute_error
```

\rightarrow เสือภาก MAE

```
mean_absolute_error(y_test, predictions)
```

50352.22825794297 \rightarrow ค่า MAE คือค่าที่บ่งบอกถึงค่าความไม่ถูกต้องของค่าที่ได้รับจากโมเดล

```
import pickle
```

\rightarrow เก็บ model ไปใน binary file

```
pickle.dump(model_lr, open('/content/sample_data/lr.pkl', 'wb'))
```

\rightarrow ดูไฟล์

```
model_lr1 = pickle.load(open('/content/sample_data/lr.pkl', 'rb'))
```

\rightarrow ดูค่า coefficient

```
model_lr1.coef_
```

```
array([-4.31396373e+04, -4.29256731e+04, 1.15069493e+03, -8.37825121e+00,
       1.17648543e+02, -3.84887721e+01, 4.54360026e+01, 4.05070684e+04])
```