

นวัตกรรมข้อมูลสำหรับงานธุรกิจดิจิทัล Data Innovation for Digital Business



การวิเคราะห์ความสัมพันธ์
(Correlation Analysis)



ประเภทของการวิเคราะห์ การถดถอยเชิงเส้น

- การวิเคราะห์การถดถอยเชิงเส้นเดียว
(Simple Linear Regression Analysis)
- การวิเคราะห์การถดถอยเชิงพหุคูณ
(Multiple Linear Regression Analysis)



การวิเคราะห์การถดถอยเชิงพหุคูณ (Multiple Linear Regression Analysis)

เป็นวิธีการวิเคราะห์การถดถอยที่ความสัมพันธ์ของตัวแปรเป็นเชิงเส้นตรง และมีตัวแปรประมาณการ (X) มากกว่า 1 ตัว และตัวแปรตอบสนอง (y) หนึ่งตัว ซึ่งมีความสัมพันธ์แทนด้วยสมการคณิตศาสตร์ดังนี้

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i + \varepsilon$$

โดยที่ y แทน ตัวแปรตอบสนอง

β_0 แทน ค่าคงที่ของสมการถดถอย ซึ่งเป็นค่าจุดตัด (Intercept) แกน y

$\beta_1, \beta_2, \beta_i$ แทน ค่าสัมประสิทธิ์การถดถอย (Regression Coefficient)

ที่ เกี่ยวข้องกับ x_1, x_2, x_i

x_1, x_2, x_i แทน ตัวแปรประมาณการ

ε แทน (epsilon)

Python กับ การวิเคราะห์

การถดถอยเชิงพหุคูณ

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
[ ] df = pd.read_csv('Advertising2.csv')
df
```

- คำสั่งเรียกใช้ไลบรารี
- อ่านข้อมูลจากไฟล์ Advertising2.csv
และแสดงผลลัพธ์จากการอ่านข้อมูล

Python กับการวิเคราะห์ การถดถอยเชิงเส้นเดียว

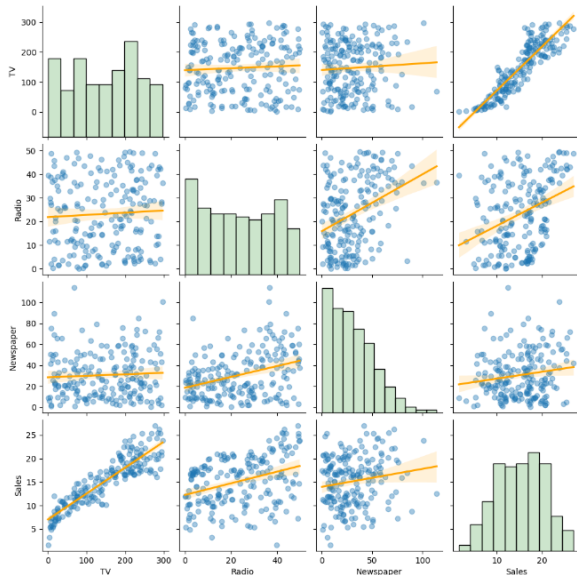
```
df.corr()
```

	TV	Radio	Newspaper	Sales
TV	1.000000	0.054809	0.056648	0.901208
Radio	0.054809	1.000000	0.354104	0.349631
Newspaper	0.056648	0.354104	1.000000	0.157960
Sales	0.901208	0.349631	0.157960	1.000000

- คำสั่งหาค่าสหสัมพันธ์
- พบว่าค่าสหสัมพันธ์ ระหว่าง Sales กับ TV มีค่ามากที่สุดคือ 0.901208 ซึ่งมีค่าความสัมพันธ์สูงในเชิงบวก รองลงมาคือ Radio และน้อยที่สุดคือ Newspaper

Python กับ การวิเคราะห์ การถดถอยเชิงเส้นเดียว

```
sns.pairplot(df,  
    kind='reg',  
    plot_kws={'scatter_kws': {'alpha': 0.4},  
              'line_kws': {'color': 'orange'}},  
    diag_kws={'color': 'green', 'alpha': 0.2});
```



- pairplot เป็นคำสั่งสำหรับพล็อตแผนภาพจากคอลัมน์ทั้งหมดที่อยู่ในตัวแปร df กำหนดให้เป็นแผนภาพประเภท reg เพื่อให้แสดงเส้นความสัมพันธ์
- scatter_kws กำหนดคุณสมบัติให้กับจุดที่พล็อตมีความทึบของสีระดับ 0.4
- line_kws กำหนดคุณสมบัติให้กับเส้นความสัมพันธ์ให้มีสีส้ม
- diag_kws กำหนดคุณสมบัติให้กับกราฟแท่งให้มีสีเขียว และความทึบของสีระดับ 0.2



Pythonกับการวิเคราะห์ การถดถอยเชิงเส้นเดียว

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

```
X = df[['TV', 'Radio', 'Newspaper']]  
y = df[['Sales']]
```

- เรียกใช้คลาส LinearRegression เพื่อนำมาใช้ในการหาความสัมพันธ์ระหว่างตัวแปร จากโมดูล Scikit-learn
- กำหนดคอลัมน์ TV , Radio , Newspaper เป็นตัวแปรประมาณการที่ X และคอลัมน์ Sale เป็นตัวแปรตอบสนองที่ y



Python กับ การวิเคราะห์ การถดถอยเชิงเส้นเดียว

```
lrm = LinearRegression()  
lrm.fit(X, y)
```

- ให้ตัวแปร lrm เก็บแบบจำลอง LinearRegression ที่จะสร้างขึ้น
- กำหนดให้แบบจำลองเรียนรู้ข้อมูลจากตัวแปร X และ y

Python กับ การวิเคราะห์ การถดถอยเชิงเส้นเดียว

```
lrm.intercept_
```

```
array([4.62512408])
```

```
lrm.coef_
```

```
array([[0.05444578, 0.10700123, 0.00033566]])
```

- ค่าคงที่ของสมการถดถอย ซึ่งเป็นค่าจุดตัด (Intercept) y

ค่าสัมประสิทธิ์การถดถอย (Regression Coefficient) ของตัวแปรประมาณการ

Python กับ การวิเคราะห์ การถดถอยเชิงเส้นเดียว

```
predictions = lrm.predict(X)
predictions
```

```
array([[21.22097209],
       [11.26824775],
       [10.49620897],
       [17.31244651],
       [15.64413685],
       [10.35633677],
       [11.2732847 ],
       [13.27062458],
       [ 5.31839603],
       [15.78871013],
       [ 8.85277202]])
```

- นำแบบจำลองที่ได้มาพยากรณ์จาก
คอลัมน์ TV , Radio , Newspaper ที่ตัว
แปร X

Python กับการวิเคราะห์ การถดถอยเชิงเส้นเดียว

```
from sklearn import metrics  
metrics.mean_absolute_error(y, predictions)
```

1.2363919943957848

- เรียกใช้ฟังก์ชัน metrics เพื่อนำมาใช้ในการประเมินประสิทธิภาพแบบจำลอง จากโมดูล Scikit-learn
- ใช้คำสั่ง mean_absolute_error เพื่อหาค่าความคลาดเคลื่อน (epsilon) ระหว่างผลลัพธ์จริงในตัวแปร y กับค่าที่ระบบพยากรณ์ออกมา

Python กับการวิเคราะห์ การถดถอยเชิงเส้นเดียว

```
lrm.predict([[200, 40, 70]])  
array([[19.81782533]])
```

```
lrm.predict([[200, 40, 70],  
             [100, 80, 50],  
             [ 40, 20, 10]])  
  
array([[19.81782533],  
       [18.64658327],  
       [ 8.94633644]])
```

- พยากรณ์ข้อมูลใหม่โดยกำหนดค่า X คือ ค่าโฆษณาช่องทาง TV , Radio , Newspaper เป็น 200 , 40 , 70 ยอดขายจะเป็นเท่าไร
- พยากรณ์ข้อมูลใหม่โดยกำหนดค่า X คือ โฆษณาช่องทาง TV , Radio , Newspaper ทั้งหมด 3 ชุดตามภาพ ยอดขายจะเป็นเท่าไร

Python กับ การวิเคราะห์ การถดถอยเชิงเส้นเดียว

```
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

```
model_a = smf.ols(formula='Sales ~ TV + Radio + Newspaper', data=df).fit()
```

- คำสั่งเรียกใช้ไลบรารี statsmodels
- กำหนดให้ตัวแปร model_a เก็บค่าสำหรับกำหนดความสัมพันธ์ ระหว่างคอลัมน์ Sales ที่ TV , Radio , Newspaper ที่เป็นตัวแปรประมาณการ แล้วนำข้อมูลจากตัวแปร df มาสร้างโมเดล

Pythonกับการวิเคราะห์ การถดถอยเชิงเส้นเดียว

```
print(model_a.summary())
```

OLS Regression Results

```
=====
Dep. Variable:      Sales    R-squared:                0.903
Model:              OLS      Adj. R-squared:            0.901
Method:             Least Squares    F-statistic:           605.4
Date:               Tue, 30 May 2023    Prob (F-statistic):    8.13e-99
Time:               10:59:24    Log-Likelihood:       -383.34
No. Observations:   200        AIC:                   774.7
Df Residuals:       196        BIC:                   787.9
Df Model:            3
Covariance Type:    nonrobust
=====
```

```
=====
              coef    std err          t      P>|t|      [0.025    0.975]
-----
Intercept    4.6251     0.308     15.041     0.000     4.019     5.232
TV            0.0544     0.001    39.592     0.000     0.052     0.057
Radio        0.1070     0.008    12.604     0.000     0.090     0.124
Newspaper    0.0003     0.006     0.058     0.954    -0.011     0.012
=====
```

```
=====
Omnibus:            16.081    Durbin-Watson:           2.251
Prob(Omnibus):      0.000    Jarque-Bera (JB):        27.655
Skew:               -0.431    Prob(JB):                9.88e-07
Kurtosis:           4.605    Cond. No.                 454.
=====
```

- แสดงผลจาก statsmodels ที่ได้



Reference

<http://pws.npru.ac.th/chalida/data/files/บทที่%208สหสัมพันธ์และการถดถอย.pdf>

https://www.nectec.or.th/schoolnet/library/snet2/knowledge_math/relation/relate2a.html

<http://oservice.skru.ac.th/ebookft/368/chapter10.pdf>

<https://bigdata.go.th/big-data-101/correlation-analysis-in-big-data/>

<https://www.coraline.co.th/single-post/basic-statistic-2>

<https://csit.nu.ac.th/kraisak/ds/ds/chapter04/Chapter04.pdf>

<https://www.youtube.com/watch?v=MNu32bm8dos&t=1120s>

https://www.youtube.com/watch?v=PC7Zk_u6g4w