
The Curious
DataRockie Basecamp

Correlation and Regression

Slide Version 1.0 | 31 Dec 2023

Welcome

สวัสดีคร้าบทุกคน 😚

คอร์สนี้เรามาเรียนสถิติพื้นฐานที่ใช้กัน**เยอะสุด**ในงาน data analyst เวลาหาความสัมพันธ์ของตัวแปร

ลืมทุกอย่างที่เคยเรียนมา แล้วมา**เริ่ม**ใหม่ในคอร์สนี้ 555+

The Contents

- Welcome
- Correlation
- Linear Regression
- Error Calculation
- Model Interpretation
- Hands-On Tutorials

Your Instructor

แอดทอย คนดี คนเดิม

Bachelor of Economics, TH
Master of Economics, UK
1 Learning Hour Every Day

เรียน เขียน แชร์ 💯

Data

ข้อมูลในทางสถิติหลักๆจะแบ่งออกเป็น 2 แบบ

1. เชิงปริมาณ เช่น ยอดขาย กำไร
2. เชิงคุณภาพ เช่น จังหวัด ประเทศ ประเภทสินค้า

Ad Data

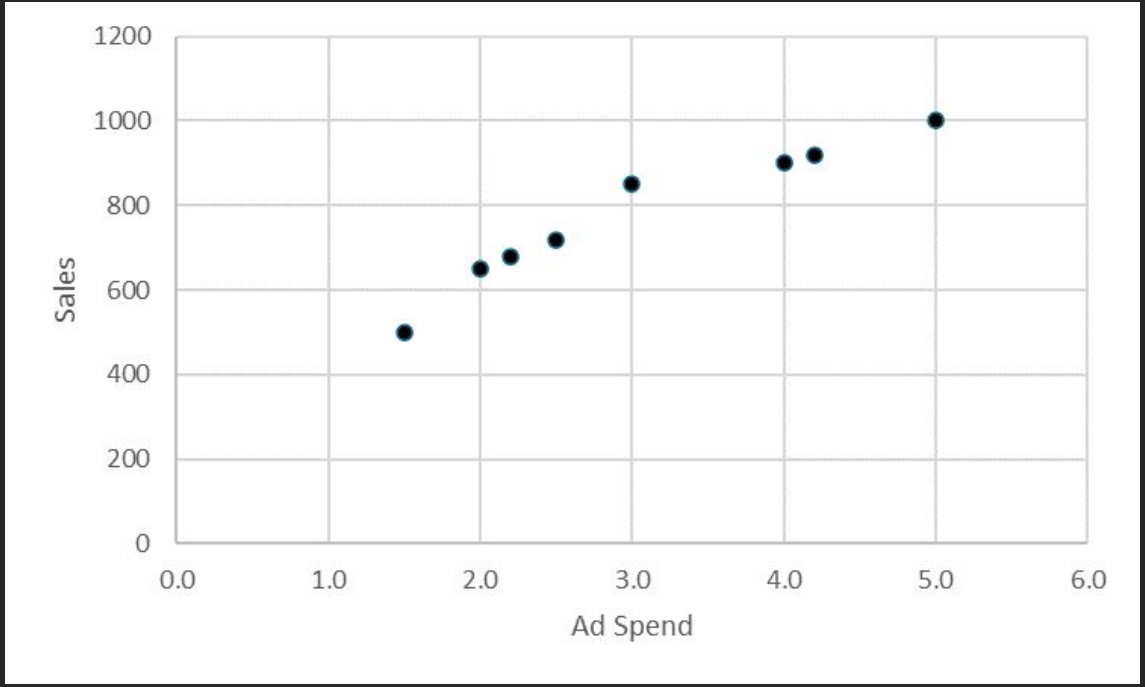
ข้อมูลตัวอย่างในคอร์สนี้เป็นแบบตัวเลขเชิงปริมาณ

ตัวแปรต้นคือ **$ Ad Spend**
ตัวแปรตามคือ **$ Sales**

Sales = f(Ad Spend)

Ad	Sales
1.5	500
2.0	650
2.2	680
2.5	720
3.0	850
4.0	900
4.2	920
5.0	1000

Basic Scatter Plot




Before Correlation

Covariance

ยุคแรก นักสถิติใช้ค่า COV เพื่อหาความสัมพันธ์**เชิงเส้นตรง**ของตัวแปรสองตัว

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$cov_{x,y}$ = covariance between variable x and y

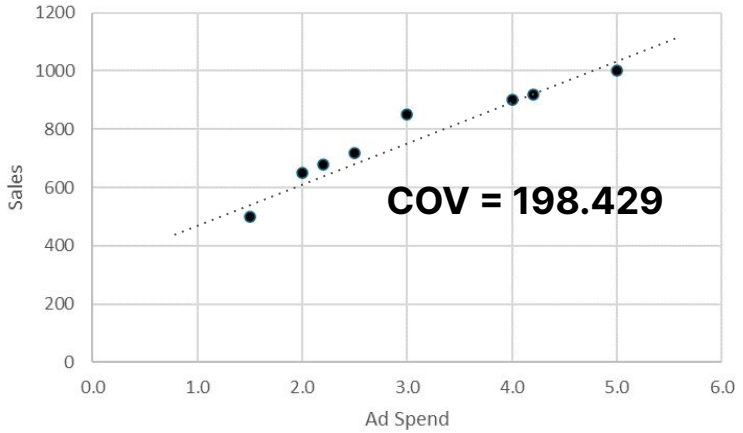
x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values



Limitation of Covariance

แต่ข้อจำกัดของ COV คืออธิบายผลยาก เพราะมีค่าได้ตั้งแต่ **-infinity** ถึง **+infinity**


Correlation

The Birth of Correlation

Karl Pearson จึงได้คิดวิธีการปรับสูตร Covariance ให้อ่านค่าได้ง่ายขึ้น

กำเนิดเป็น **Pearson Correlation** ที่มีค่าอยู่ระหว่าง **-1** ถึง **+1**

Karl Pearson (1857-1936)

Correlation Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

simple formula →

$$\frac{\text{COV(x,y)}}{\text{sd.x * sd.y}}$$

Simple Functions

ทุกวันนี้เราไม่ต้องคำนวณมือเองเหมือนสมัยก่อนแล้ว
แค่เรียกใช้ **function** ได้เลย

=COVARIANCE.S()

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$cov_{x,y}$ = covariance between variable x and y
x_i = data value of x
y_i = data value of y
$\bar{x}$ = mean of x
$\bar{y}$ = mean of y
N = number of data values

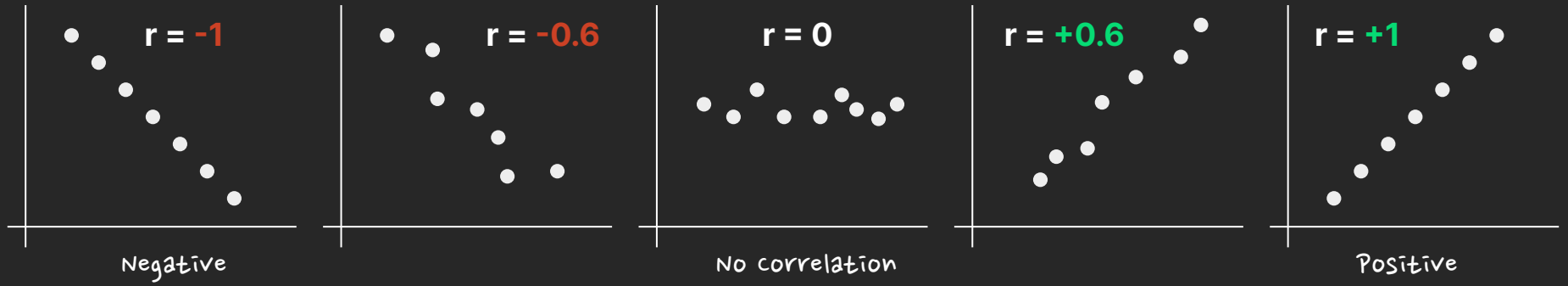
=CORREL()

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Interpretation

ค่า correlation หรือ r บอกความสัมพันธ์ของตัวแปรแบบตัวเลขสองตัว (เหมือน covariance)



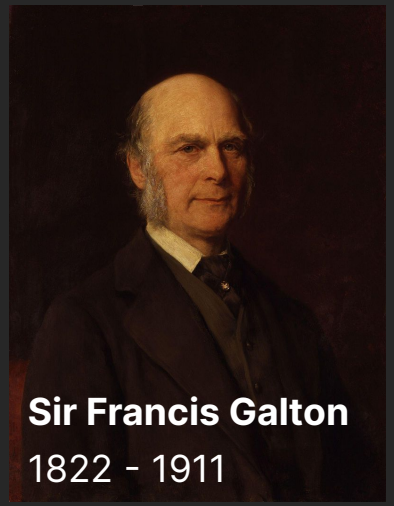
Use Case

Correlation ใช้ตอบคำถามต่อไปนี้ว่าตัวแปรสองตัวมี ...

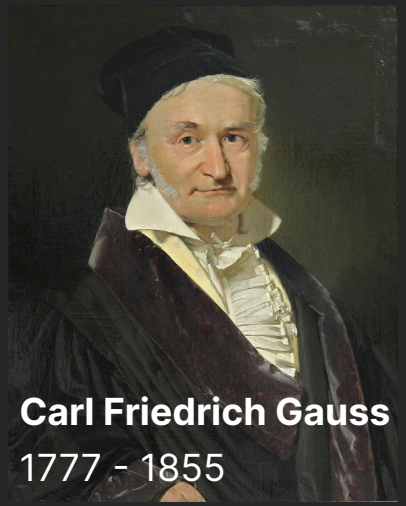
- ความสัมพันธ์เชิงเส้นตรงกันหรือเปล่า
- ความสัมพันธ์เป็นเชิง + หรือเชิง -
- ความสัมพันธ์นั้นเข้มแค่ไหน strong หรือ weak


Regression

The Inventors



Sir Francis Galton
1822 - 1911



Carl Friedrich Gauss
1777 - 1855

ปี **1809** Gauss พัฒนาเทคนิค **Least Squares Method**

ปี **1885** Galton ได้นำเสนอไอเดียเรื่อง **Regression** ให้กับโลกสถิติ

“Regression Towards The Mean”

Regression

นักสถิติคิดโมเดล Regression เพื่อใช้ตอบคำถามที่ Correlation ตอบไม่ได้

ตัวอย่างเช่น Sales = f(Ad Spend)

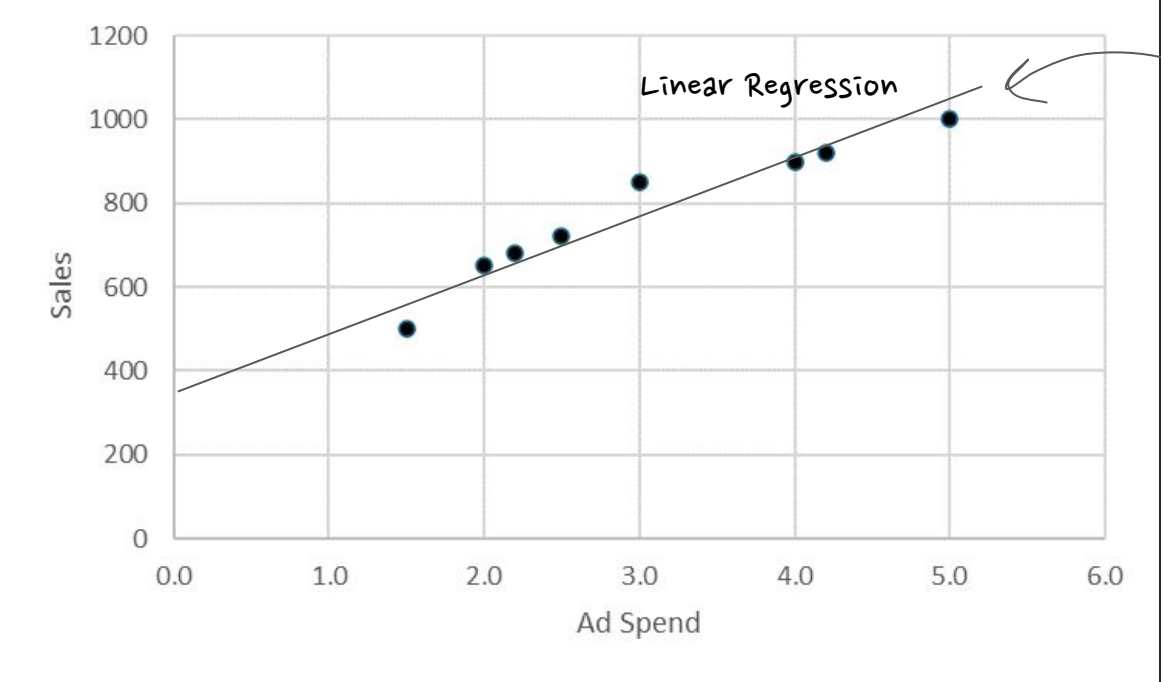
✅ ถ้า Ad Spend เปลี่ยน 1 หน่วย แล้ว Sales จะเปลี่ยนเท่าไหร่ ปัจจัยอื่นๆคงที่

A Primer to Linear Regression

และโมเดล Regression รูปแบบที่ง่ายที่สุด

ที่สอนในคลาสสถิติพื้นฐานทั่วโลกคือ **Linear Regression**

It’s Just A Straight Line

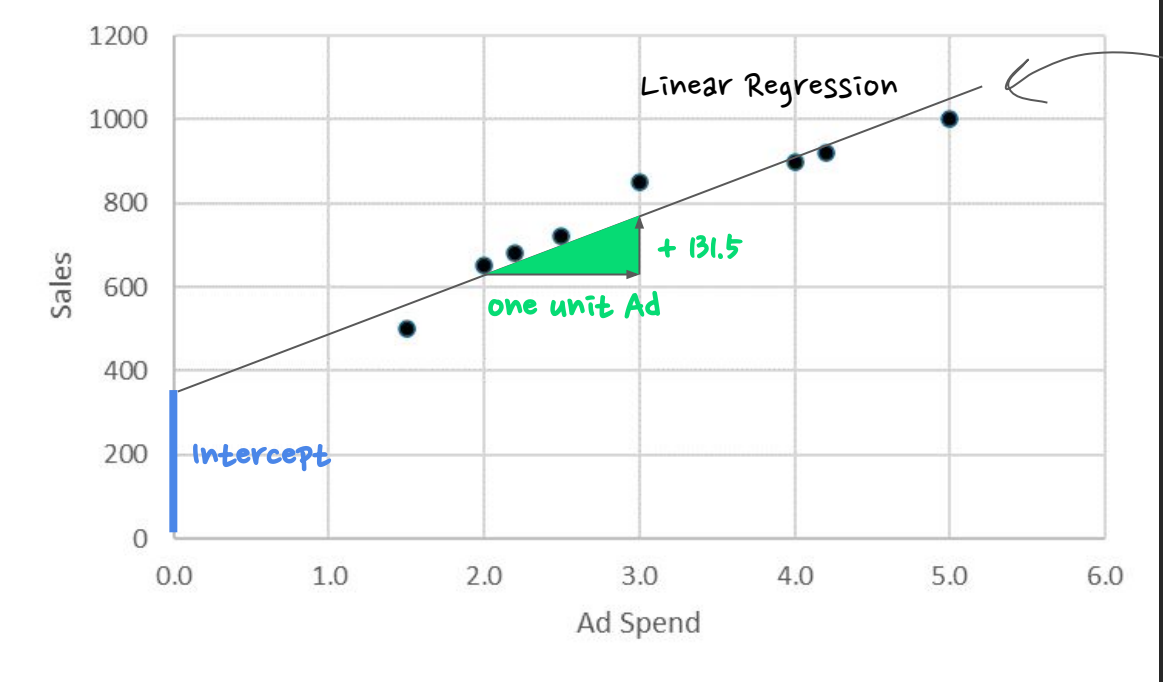


Sales = f(Ad)

Sales = **intercept** + **slope***Ad

Sales = **376** + **131.5***Ad

A Simple Idea

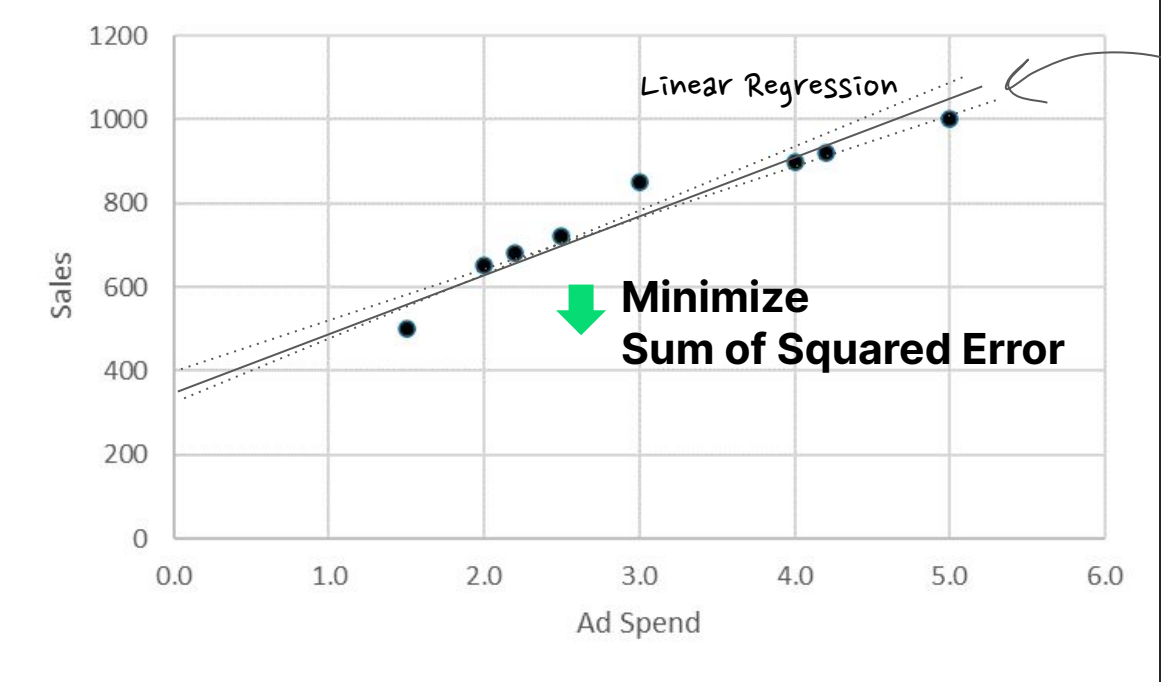


Sales = f(Ad)

Sales = **intercept** + **slope***Ad

Sales = **376** + **131.5***Ad

The Best Line



Sales = f(Ad)

Sales = **intercept** + **slope***Ad

Sales = **376** + **131.5***Ad

ทำให้ค่า **Total Error** ของโมเดลมีค่าต่ำที่สุด (minimize)

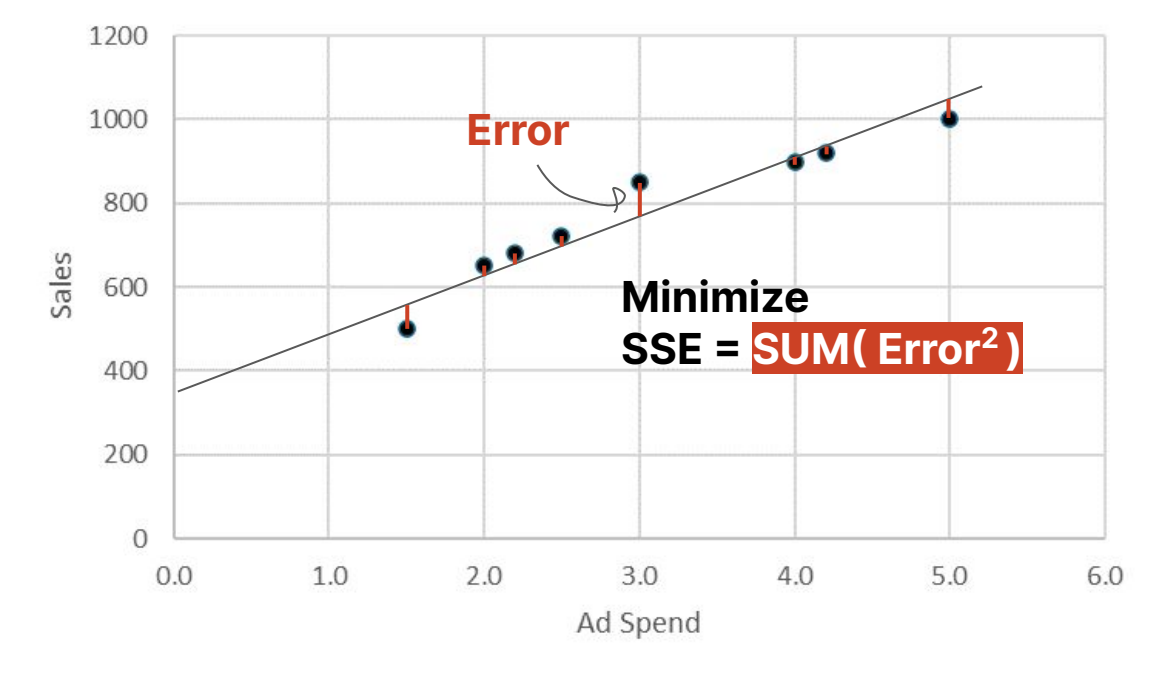
Error Calculation

Error คือ**ความแตกต่าง**ระหว่างค่าจริง กับสิ่งที่โมเดลทำนาย

Error = Actual - Prediction

✅ Error โดยรวมของโมเดล Linear Regression เรียกว่า **SSE** (sum of squared error)

Error Calculation



Sales = f(Ad)

Sales = **intercept** + **slope***Ad

Sales = **376** + **131.5***Ad

ทำให้ค่า **Total Error** ของโมเดลมีค่าต่ำที่สุด (minimize)

Our Model Tries To ..

หาค่า intercept และ slope ที่ทำให้ค่า sum of squared error ของโมเดลที่ค่าต่ำที่สุด

Find {**intercept**, **slope**}
That **Minimize SSE**

ง่ายจนงง 555+

R-Squared

คือค่าที่ใช้วัดสิ่งที่โมเดล linear regression **อธิบายได้** โดยมีค่าวิ่งอยู่ระหว่าง [0 - 1]

ค่ายิ่ง**เข้าใกล้ 1** แปลว่าโมเดลเราทำงานได้ดี -> ตัวแปรต้น X อธิบายตัวแปร Y ได้ดี

PS. R2 มีชื่อเต็มๆว่า Coefficient of Determination

R-Squared

เราสามารถคำนวณค่า R-Squared ได้ด้วยสูตร

R-Squared = SS_m / SS_t

หรือ **1** - SS_r / SS_t

ตัวย่อที่เราใช้ในสูตร
m = model
r = residual
t = total variance

และ $SS_m + SS_r = SS_t$

R-Squared

เราสามารถคำนวณค่า R-Squared ได้ด้วยสูตร

R-Squared = **4** / **10** = 40%

หรือ **1** - **6** / **10** = 40%

ตัวย่อที่เราใช้ในสูตร

m = model

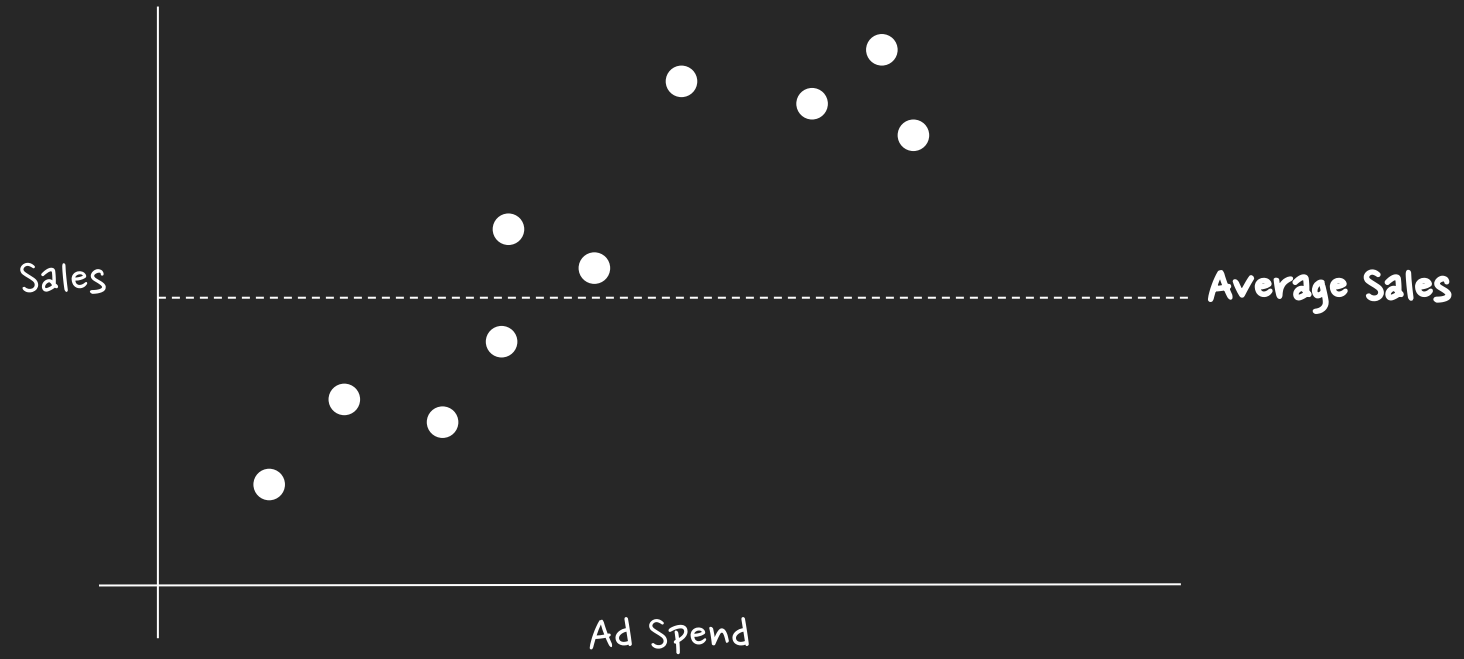
r = residual

t = total variance

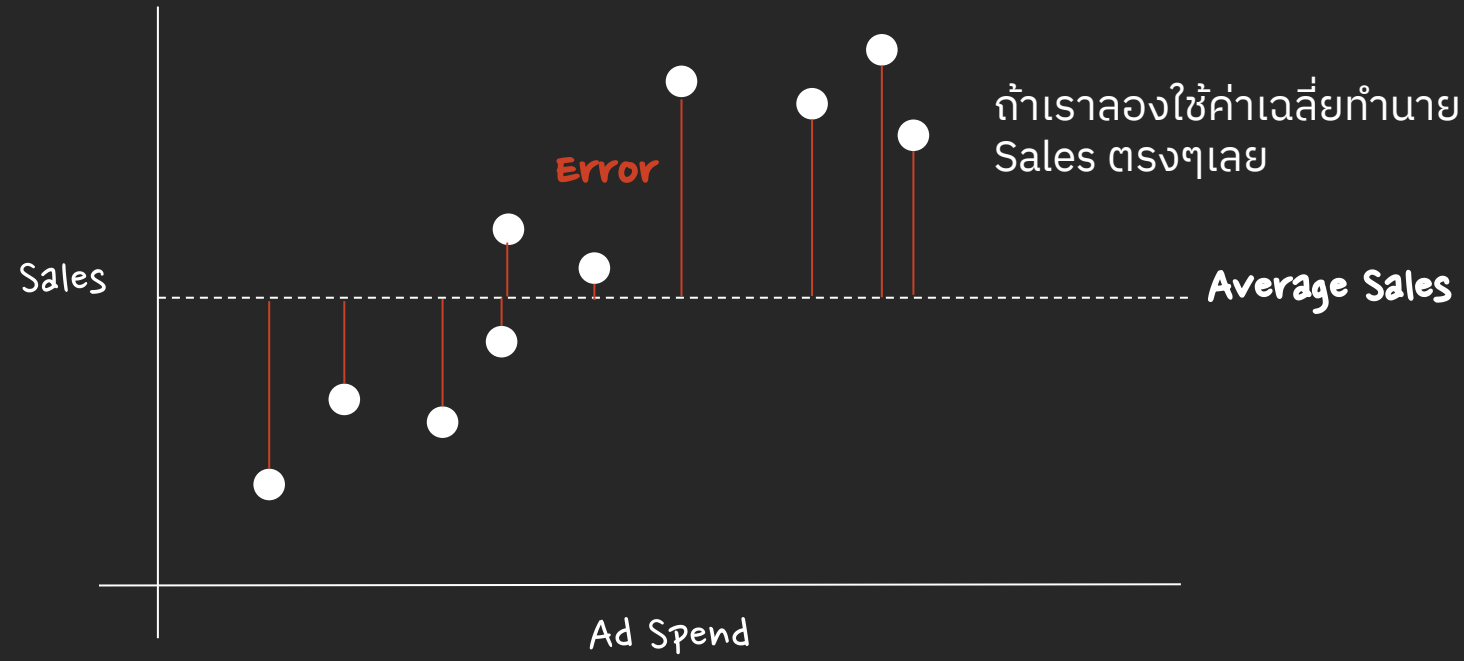
และ $SS_m + SS_r = SS_t$


Geek Mode

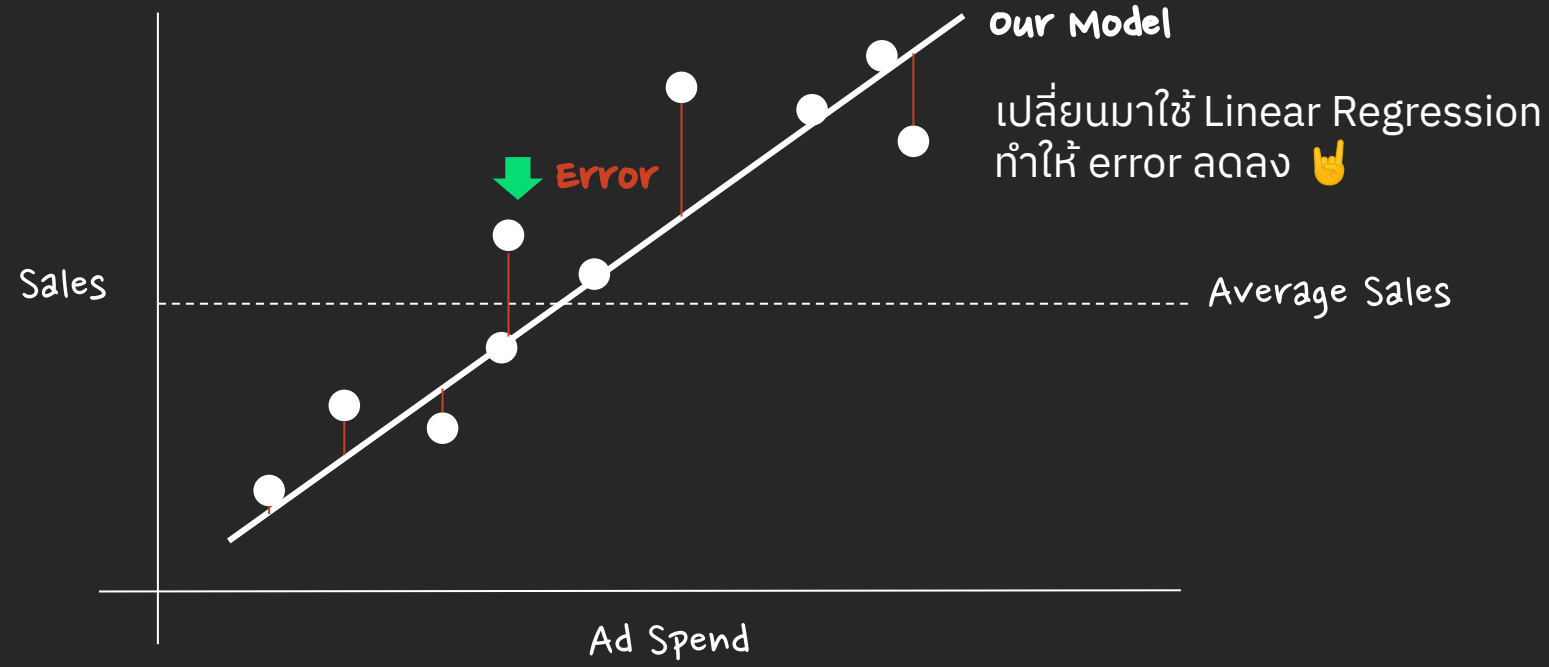
R-Squared Visualization



R-Squared Visualization

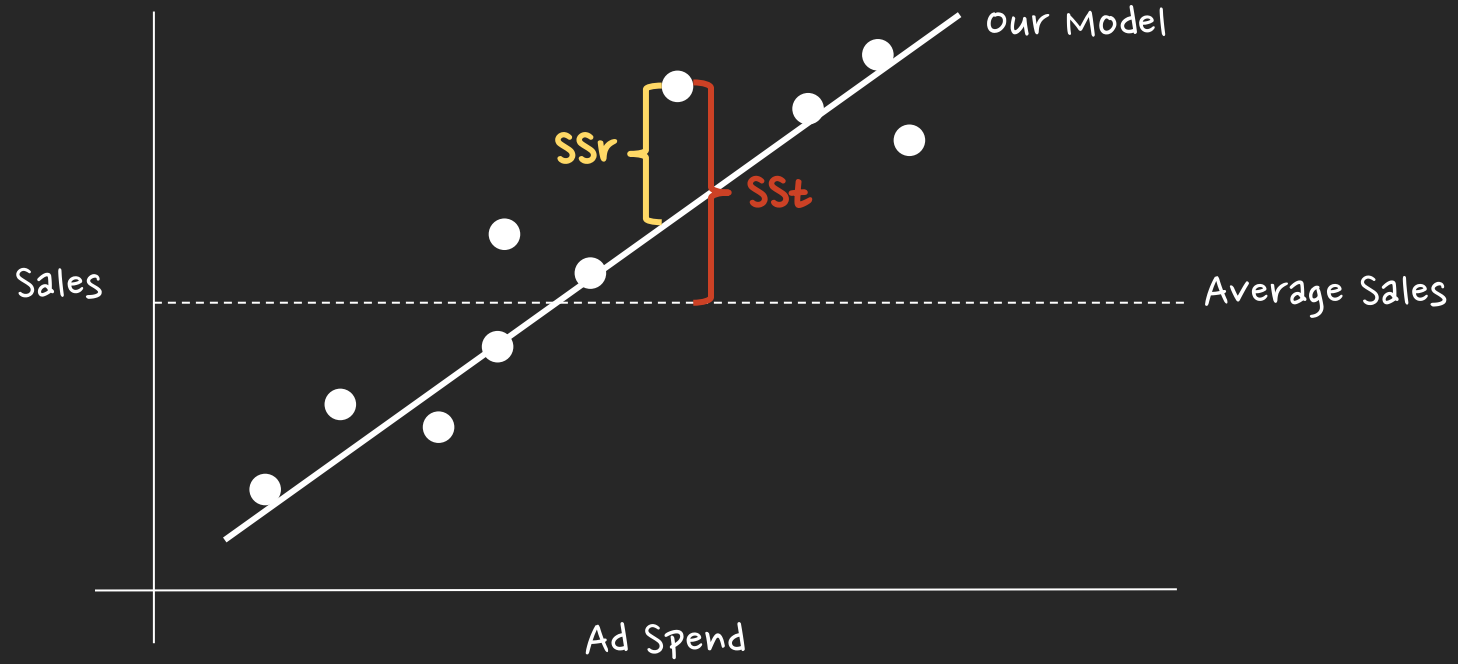


R-Squared Visualization

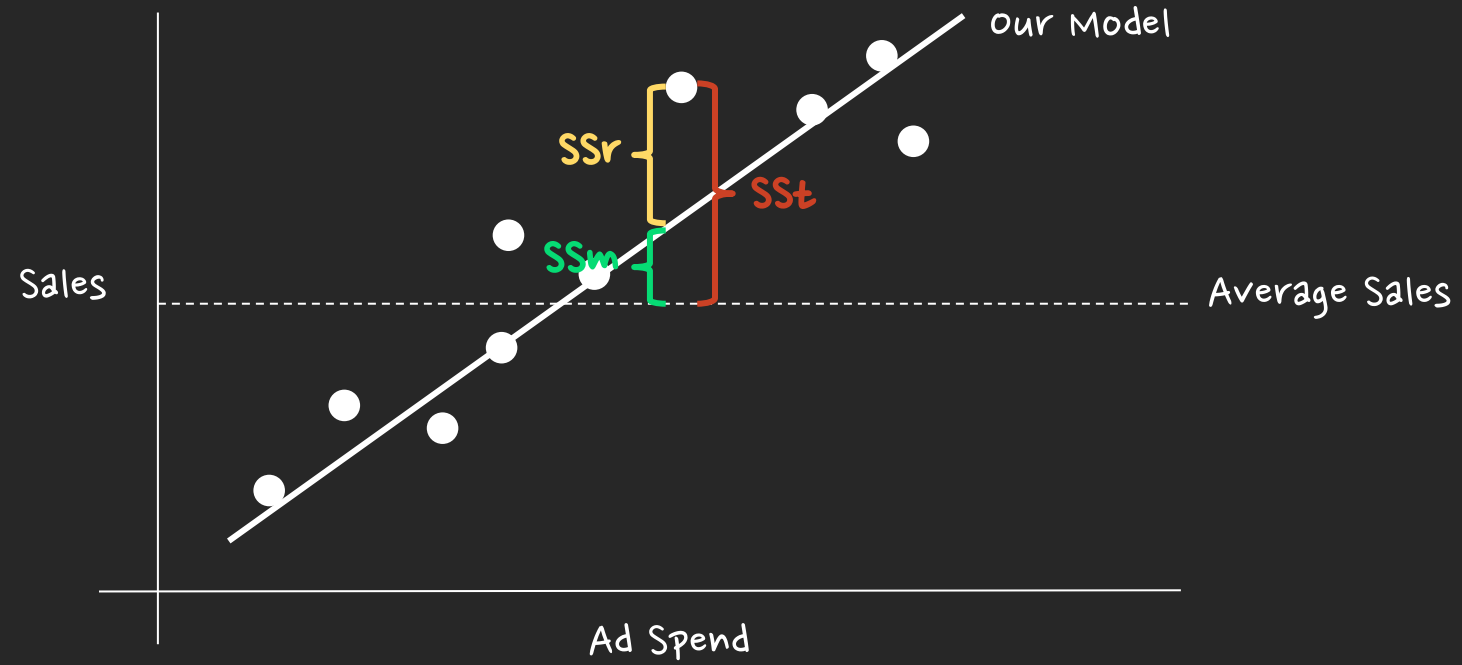


เปลี่ยนมาใช้ Linear Regression
ทำให้ error ลดลง 🤘

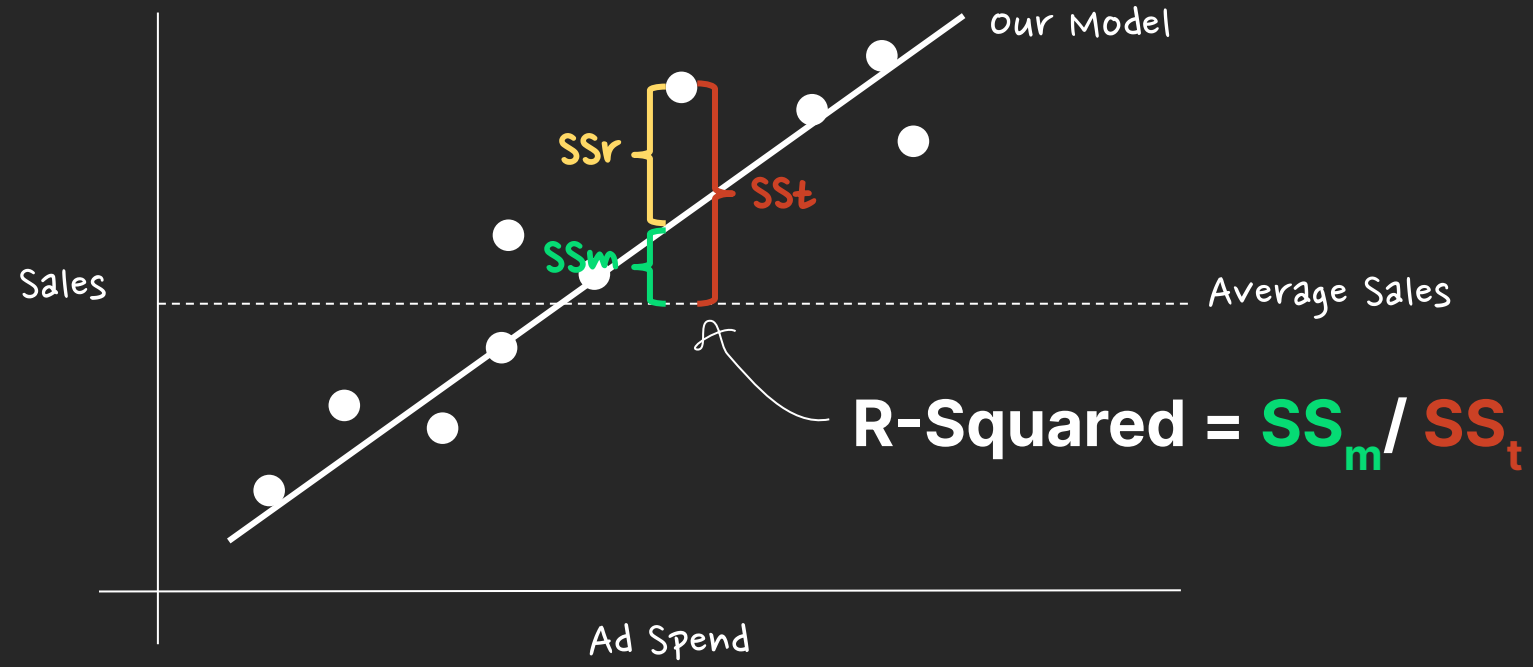
R-Squared Visualization



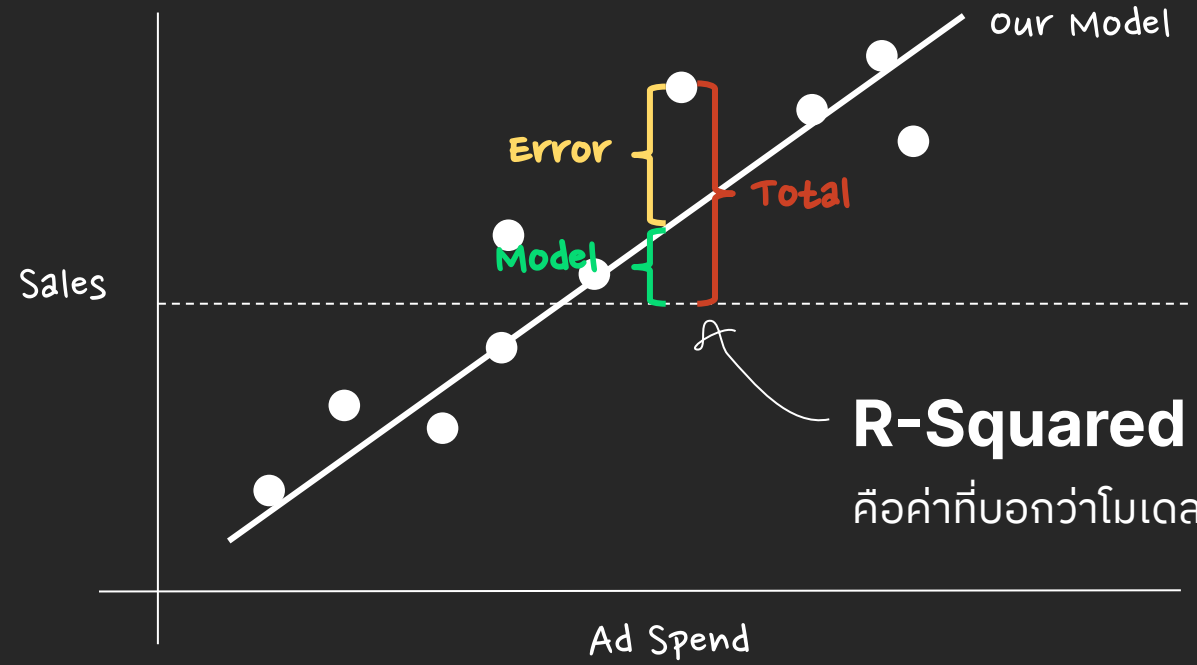
R-Squared Visualization



R-Squared Visualization



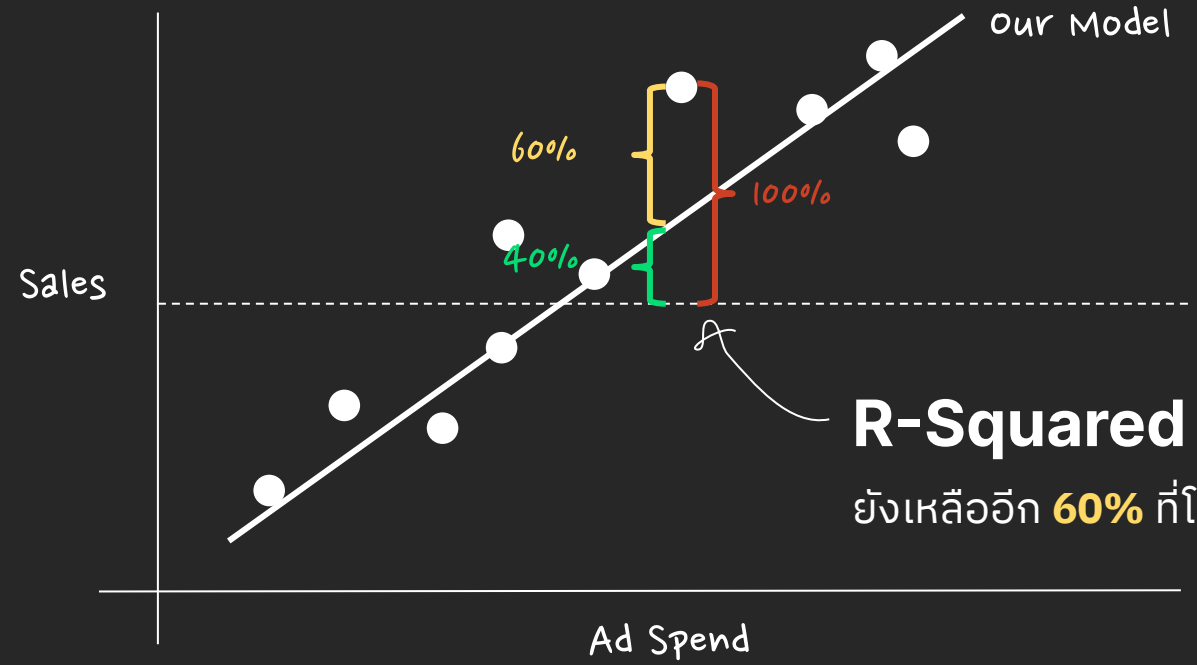
R-Squared Visualization



R-Squared = SS_m / SS_t

คือค่าที่บอกว่าโมเดลเราทำได้ดีกว่าการใช้ค่าเฉลี่ย

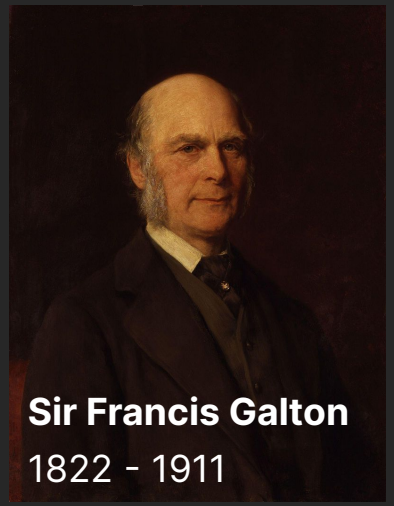
R-Squared Visualization



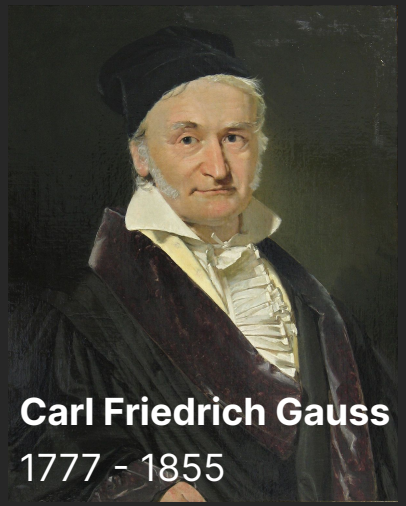
R-Squared = 40%

ยังเหลืออีก **60%** ที่โมเดลเรายังอธิบายไม่ได้

Recap: The Inventors



Sir Francis Galton
1822 - 1911



Carl Friedrich Gauss
1777 - 1855

ปี **1809** Gauss พัฒนาเทคนิค **Least Squares Method**

ปี **1885** Galton ได้นำเสนอไอเดียเรื่อง **Regression** ให้กับโลกสถิติ

“**Regression** Towards **The Mean**”


Mind Blown 🤯
ง่ายจนงง 555+


Simple vs. Multiple

Two Types of LR

[1] **Simple Linear Regression**
Sales = f(Ad)

[2] **Multiple Linear Regression**
Sales = f(Ad, TV, Radio)

The Model

[1] **Simple Linear Regression**
Sales = b0 + b1*Ad

[2] **Multiple Linear Regression**
Sales = b0 + b1*Ad + b2*TV + b3*Radio

โดยที่ b0 คือ **intercept** และ b1, b2, b3, .. คือ **slope**


Prediction

Predict New Data

Simple Linear Regression

Sales = f(Ad)

Sales = 100 + 50*Ad

✅ ถ้าเราใช้เงินโฆษณา **$5 Million USD** เราจะได้ยอดขายกลับมาเท่าไหร่?

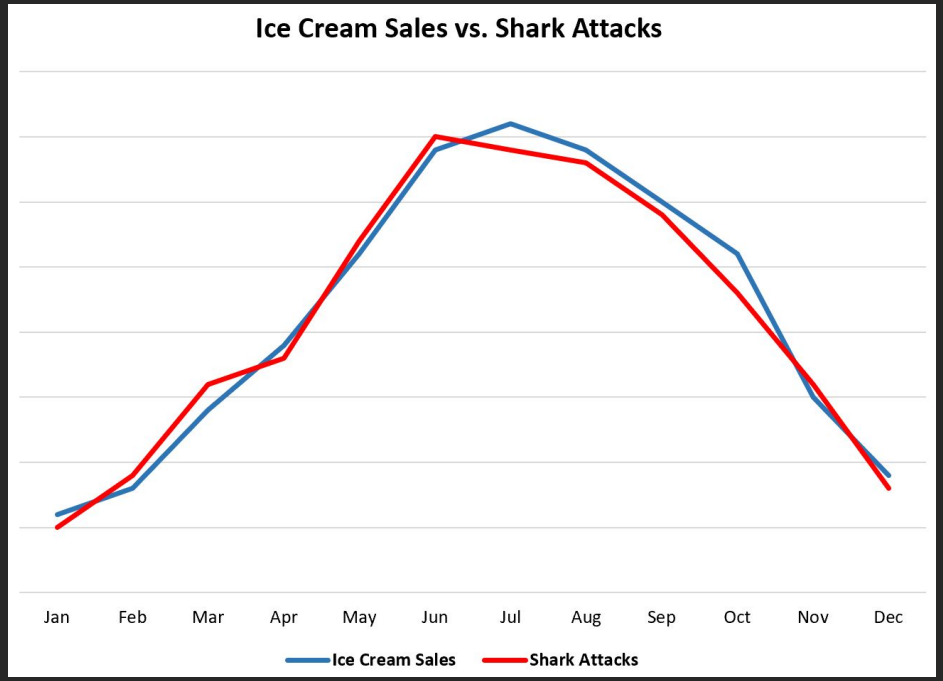
Sales = 100 + 50*5 = **350**


Caution

Correlation Does Not Imply Causation

ตัวแปรสองตัวมีความสัมพันธ์กัน (Correlation)
ไม่ได้แปลว่า X ทำให้เกิด Y (Causation)

Ice Cream vs. Shark



ชอบกินไอติม **ไม่ได้**แปล
ว่าจะถูกฉลาดกัด

Source: Correlation Does Not Imply Causation: 5 Real-World Examples - Statology

Ice Cream vs. Weight

กินไอติมเยอะ น้ำหนักขึ้น?

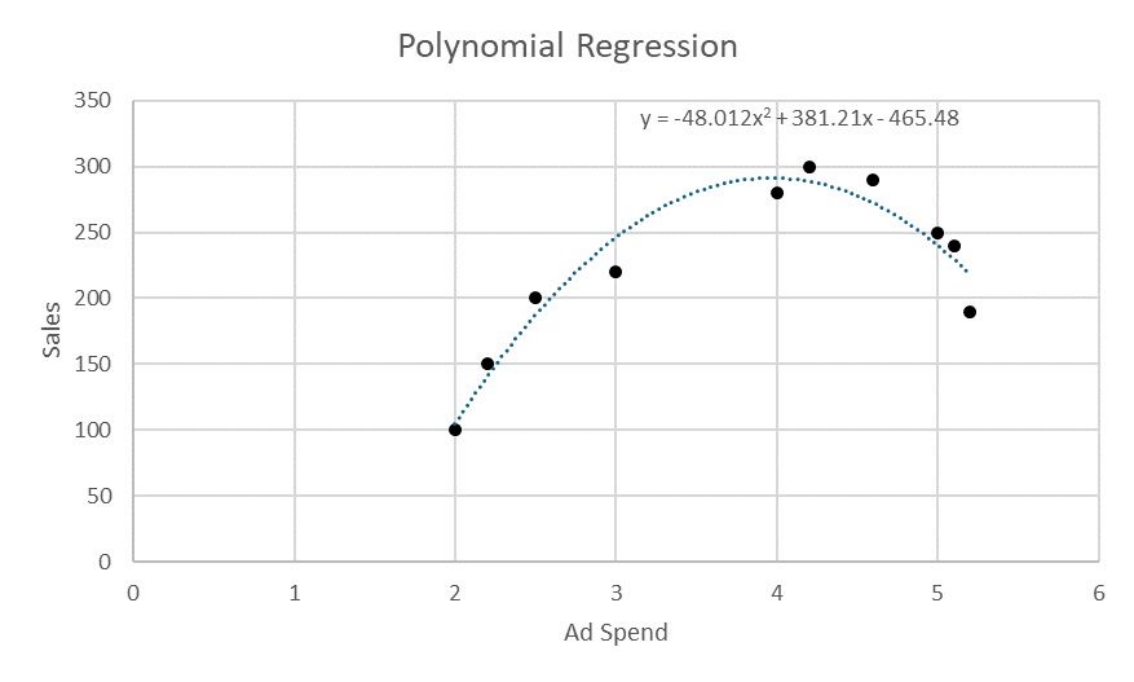
Remember This

ทุกความสัมพันธ์แบบ Causation **ต้องมี** Correlation เสมอ

แต่**ไม่ใช่**ทุกความสัมพันธ์แบบ Correlation จะเป็น Causation


Other Types
of Regression

Polynomial Regression



เวลาเจอกับ **Non-Linear** Relationship

Logistic Regression

เวลาเจอกับตัวแปรตาม y แบบ **binary** (0/1)

เช่น สอบผ่านหรือไม่ผ่าน

Take-Home Cheat Sheets

นักสถิติใช้ Correlation และ Regression ในการโมเดลความสัมพันธ์ตัวแปรเชิงปริมาณ

Note - ถ้าเรียนต่อไปจะรู้ว่า Regression สามารถรับตัวแปรประเภทอื่นได้ด้วย เช่น dummy 0,1

Correlation หรือ r มีค่าอยู่ระหว่าง [-1, +1]

Linear Regression ตัวพื้นฐานมี 2 แบบคือ

- Simple มีตัวแปรต้นหนึ่งตัว
- Multiple มีตัวแปรต้นมากกว่าหนึ่งตัว

นักสถิติใช้ Regression เพื่ออธิบายว่าถ้า x เปลี่ยน 1 หน่วย y จะเปลี่ยนเท่าไหร่ ปัจจัยอื่นคงที่

Functions ที่ต้องใช้ให้เป็นใน Excel/ Sheets

- `COV.S()`
- `CORREL()`
- `INTERCEPT()`
- `SLOPE()`
- `LINEST()`

วิธีวัดความแม่นยำของโมเดล Linear Regression ทำได้หลายวิธี

- R-Squared ยิ่งเข้าใกล้ 1 ยิ่งดี
- MAE ยิ่งเข้าใกล้ 0 ยิ่งดี

สูตรของ R-Squared คือ SS_{model} / SS_{total}
หรือ $1 - SS_{residual} / SS_{total}$


The Curious
DataRockie Basecamp