The Curious : Essential Statistics

Theory

ข้อมูลทางสถิติหลักๆจะแบ่งเป็น 2 แบบ

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

→ ที่เราเรียนจะเน้นเป็น Numeric number data (ad spending/sales..) เนื่องจากจะเหมาะกับการรัน correlation และ linear regression มากกว่าพวกตัวแปร category

ยกตัวอย่าง Sales =f(Ad Spend)

- ตัวแปรต้น(x)คือ $ Ad Spend จะอยู่ที่แกนนอน
- ตัวแปรตาม(y)คือ $ Sales จะอยู่ที่แกนตั้ง

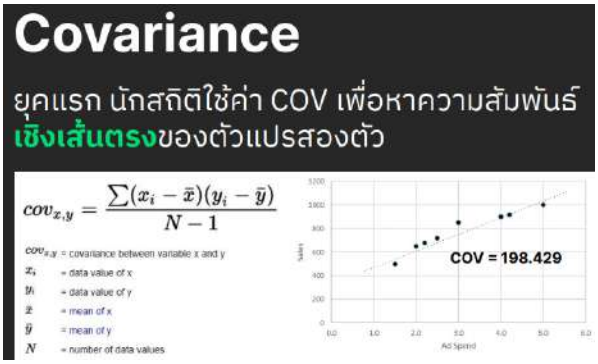
: ใช้เงินโฆษณา a ล้านบาท ได้ยอดขาย b ล้านบาท

:นิยมใช้ Scatter plot ในการทำกราฟ หาความสัมพันธ์(เชิงเส้นตรง)

-

ยุคเริ่มต้นใช้ค่า Covariance

สมัยแรกสุดจะใช้ Covariance หา**ความสัมพันธ์เชิงเส้นตรง**ของตัวแปร numeric สองตัว



มีข้อดี คือ

: สามารถดูทิศทางความสัมพันธ์ของตัวแปร 2 ตัว

มีข้อจำกัด คือ

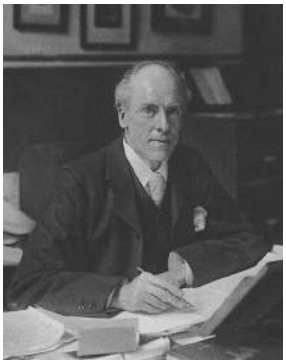
: มีช่วงกว้างมากๆ คือจาก -infinity ถึง +infinity (ไม่มี Unit ที่แท้จริง)

→ เปรียบเทียบค่ายาก นำไปใช้ต่อได้ยาก

**Noted ปัจจุบันยังคงใช้อยู่ ในการคำนวณสถิติและโมเดลขั้นสูงหลายๆตัว ⇒ Covariance Matrix, PCA หรือ Portfolio Theory ที่นักลงทุนใช้ดูว่าหุ้นหรือ asset ตัวไหนที่มีผลตอบแทนเคลื่อนที่ในทิศทางเดียวกันบ้าง etc.

การใช้งาน จะใช้ฟังก์ชัน `=COVARIANCE.S()`

เกิดค่า Correlation ขึ้นมา เพื่อการใช้งานที่ดีขึ้น



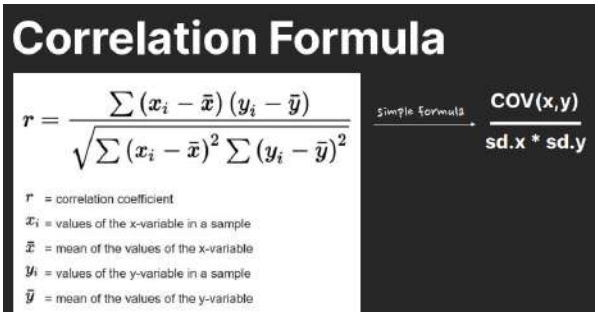
→ ด้วยค่า Covariance ใช้งานยาก นักสถิติ Karl Pearson จึงคิดค้นการ Normalize Covariance ให้อยู่ในช่วง [-1, +1] จึงเป็นที่มาของค่า **"Pearson Correlation"**

Note นอกจากนี้ Karl Pearson ยังมีส่วนร่วมในการพัฒนาเทคนิคสถิติอีกมากมาย เช่น PCA และ Chi-Squared และเป็นบุคคลสำคัญในการก่อตั้ง School of Biometrics อีกด้วย

Correlation หรือก็คือ Standardized Covariance (หรือ Normalized) เรียกอีกอย่างว่า ค่า r

ค่าจะมีช่วงอยู่ระหว่าง [-1, +1]

เกิดมาจาก นำค่า Covariance หารด้วย (ส่วนเบี่ยงเบนมาตรฐานของ x คูณกับส่วนเบี่ยงเบนมารฐาน y)



พอช่วงมีความแคบลง จึงเป็นที่นิยมในการใช้งาน สามารถบอกได้ทั้ง direction และ strength ของความสัมพันธ์นั้นๆ เช่น r=0.85 แปลว่า positive + strong relationship

การใช้งาน จะใช้ฟังก์ชัน `=CORREL()`

Interpretation

ค่า Correlation หรือ ค่า r บอกความสัมพันธ์ของตัวแปรแบบตัวเลข 2 ตัว (เหมือนกับ Covariance)

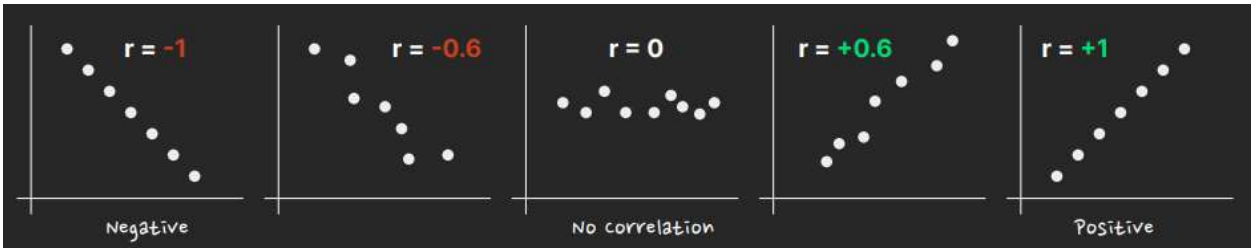
เครื่องหมาย (-) r มีค่าเป็นลบ : มีความสัมพันธ์กัน โดยที่ตัวแปรหนึ่งเพิ่ม ตัวแปรหนึ่งจะตก

→ -1 เป็น Perfectly negative correlation

เครื่องหมาย (+) r มีค่าเป็นบวก : มีความสัมพันธ์กัน โดยที่ตัวแปรหนึ่งเพิ่ม ตัวแปรหนึ่งจะเพิ่มตาม

→ +1 เป็น Perfectly positive correlation

r = 0 ก็คือ r มีค่าเป็น 0 : ตัวแปรสองตัว ไม่มีความสัมพันธ์กัน → No correlation



สรุปการใช้งาน (Use Case)

Correlation ใช้ตอบคำถามว่าตัวแปรสองตัวมี...

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

The Inventors of Regression

นักสถิติสองคนที่วางรากฐานของ Linear Regression ให้เราใช้งานทุกวันนี้คือ Carl Friedrich Gauss และ Sir. Francis Galton

ปี 1809 Gauss เป็นผู้พัฒนาเทคนิค Least Squares Method ก็คือ Linear Regression ที่เราใช้ปัจจุบัน (Ordinary Least Square)

ปี 1885 Galton เป็นคนเริ่มใช้คำว่า "Regression" ตอนค้นพบปรากฏการณ์ Regression Towards The Mean จากการศึกษาส่วนสูงของพ่อกับลูก

→ นักสถิติคิดค้นโมเดล Regression เพื่อใช้ตอบคำถามที่ Correlation ตอบไม่ได้ เช่น

จาก Sales = f(Ad Spend)

`?` ถ้า Ad Spend เปลี่ยนหนึ่งหน่วย แล้ว Sales จะเปลี่ยนเท่าไหร่ เมื่อปัจจัยอื่นๆคงที่

**Correlation จะไม่สามารถบอกเลขของตัวแปรตามที่เปลี่ยนไปได้

Regression Crash Course

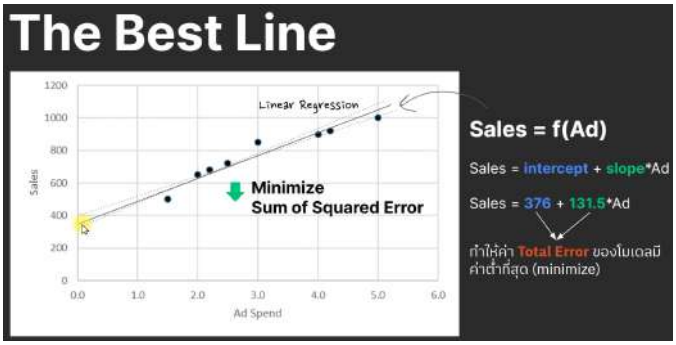
โมเดล Regression รูปแบบที่ง่ายที่สุด ในสถิติพื้นฐานคือ Linear Regression

Linear Regression ก็คือสมการเส้นตรงเหมือนกับ y = mx + c แต่ในโมเดลสถิติจะเขียน y = b0 + b1*x

สมการเส้นตรงสามารถเขียนขึ้นมาด้วย parameters สองตัว
คือ `intercept` และ `slope`

`?` รู้ได้อย่างไรว่าต้องเป็นเส้นตรงเส้นนี้

→ เป็นเหตุผลให้ Gauss คิดค้น Least Square Method ก็คือเส้นตรงที่ลากตัดผ่านข้อมูลได้ดีที่สุด เกิด Square Error น้อยที่สุด



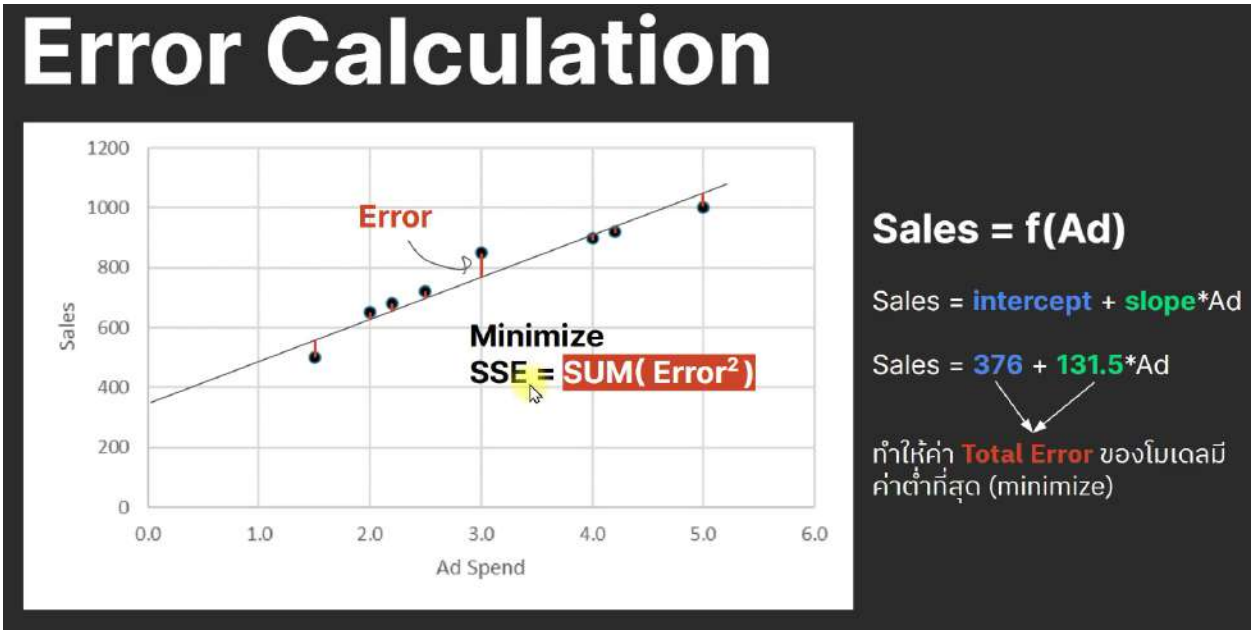
The Core Idea Behind Regression

Error คืออะไร?

คือความแตกต่างระหว่างค่าจริง กับสิ่งที่โมเดลทำนาย Error = Actual - Prediction

→ Error โดยรวมของโมเดล Linear Regression เรียกว่า SSE(Sum of squared error) หรือ Residual

Least Square Method ก็คือการ Minimize SSE



หน้าที่ของ Linear Regression จึงเป็น

หาค่า intercept และ Slope ที่ทำให้ค่า sum of squared error ของโมเดลที่มีค่าต่ำที่สุด

Find {`intercept` , `slope`} that minimize `SSE` (sum of squared errors)

R-Squared

ชื่อเต็มคือ **Coefficient of Determination** เป็นค่าสถิติที่เรานิยมใช้วัด Overall Performance ของโมเดล Linear Regression

- ใช้วัดสิ่งที่โมเดล Linear regression อธิบายได้

- จะมีค่าวิ่งอยู่ระหว่าง [0-1] ยิ่งเข้าใกล้ 1 แปลว่าโมเดลเราทำงานได้ดี → หมายความว่า ตัวแปรต้น x อธิบายตัวแปรตาม y ได้ดี

- สามารถหาได้จาก correlation^2 (** ค่า correlation คือค่า r)

**R-Squared (R2) มีอีกชื่อเต็มๆว่า Coefficient of Determination

สูตร R-Squared

SSE : Sum of Squared errors

R-Squared = sum of squared จาก Model (SSm)/ sum of squared of total (SSt)

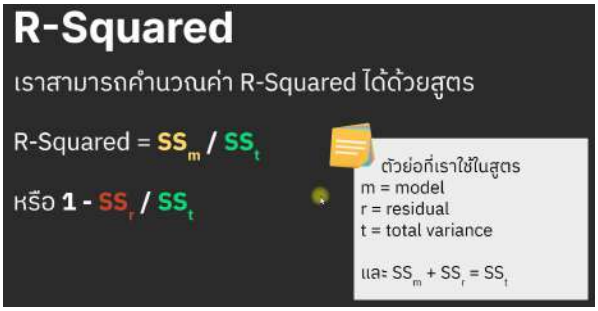
หรือ

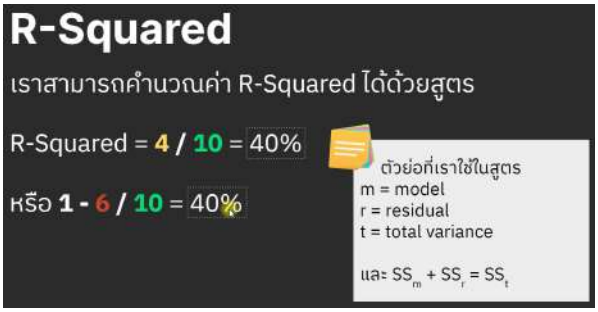
R-Squared = 1 - sum of squared residual (SSr)/ sum of squared of total (SSt)

โดยที่ SSm (โมเดลอธิบายได้) + SSr (โมเดลอธิบายไม่ได้) = 1

สองสูตรด้านบนจึงได้ผลลัพท์เหมือนกัน

ใน Excel และ Google Sheets เราไม่ต้องคำนวณมือเอง เพราะฟังก์ชัน `=LINEST()`





Note

เมื่อนักสถิติอยากวัดผลโมเดล Linear Regression สามารถใช้ Metrics ในการวัดผลได้หลากหลาย เช่น

- R-Squared หรือ SSE(Sum of squared errors)
- MAE หรือ mean absolute error
- MSE หรือ mean squared error
- RMSE หรือ root mean squared error

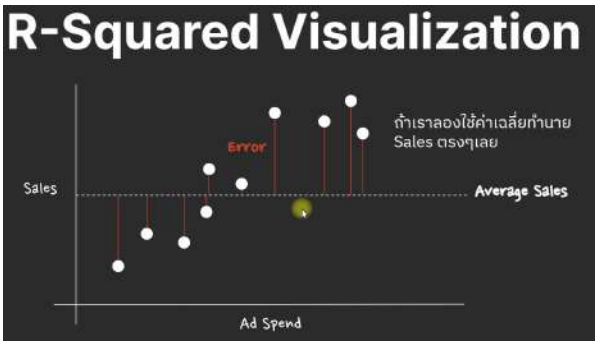
ฟังก์ชันที่ใช้ในการเขียนสูตร Excel/Google sheets : `ABS()` `AVERAGE()` `SUM()` `SQRT()` หรือยกกำลังสอง `error^2`

Tip - เวลาเห็นคำว่า "Error" ในชื่อ metrics ค่าพวกนี้ ยิ่งต่ำ ยิ่งเข้าใกล้ศูนย์ยิ่งดี แปลว่าโมเดลทำนายได้แม่นยำขึ้น actual values ใกล้กับ predicted values

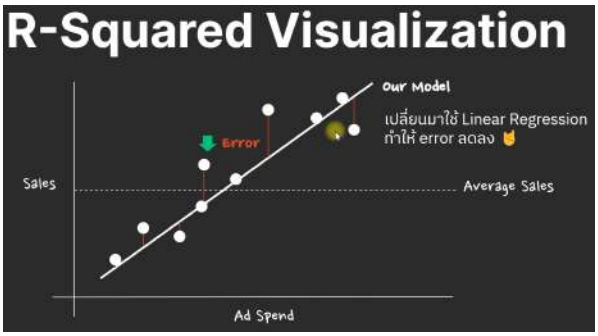
Geek Mode - R Squared Calculation

สมมติเรามี data พล็อตได้ดังภาพ

ในโลกที่ไม่มี Linear Regression เราอาจใช้ค่าเฉลี่ยในการหาค่าตรงกลาง → ผิดเยอะมาก

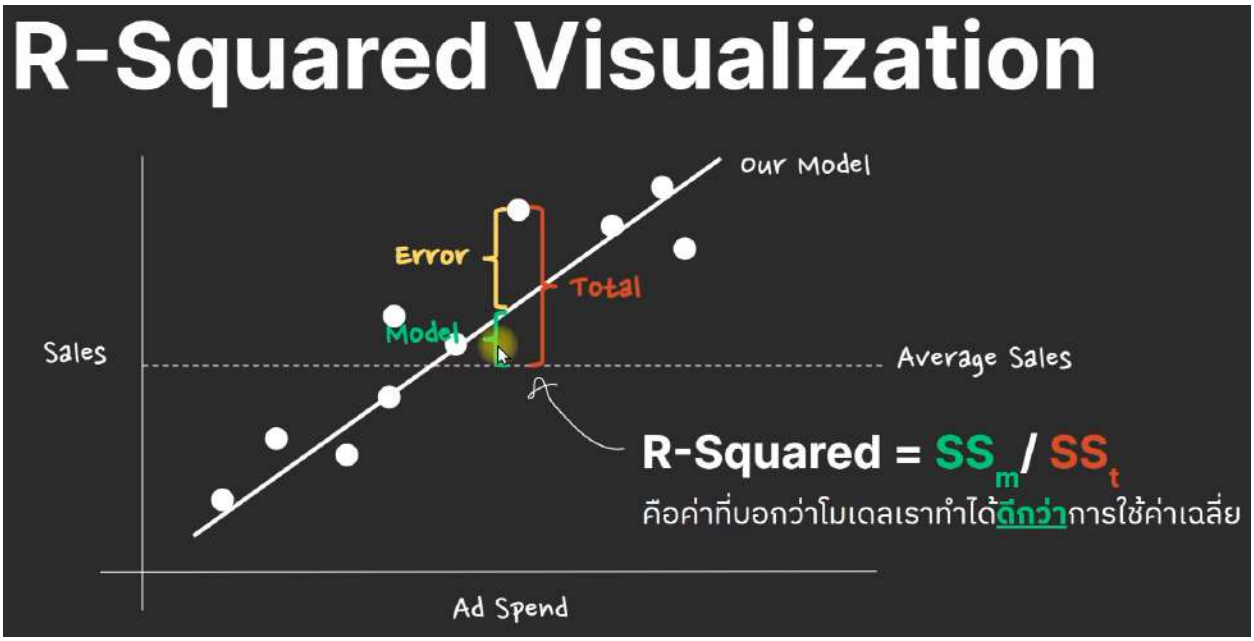


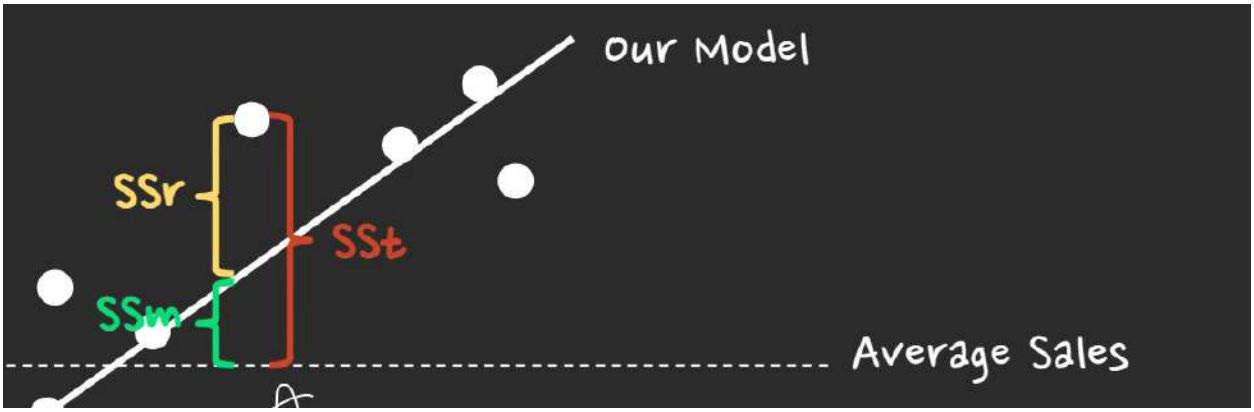
Galton จึงหาวิธี จึงได้ Linear Regression ออกมา พบว่าค่า Error ลดลงเป็นอย่างมาก



→ สามารถอธิบาย Error และ Variance ใน data ของเราได้ดียิ่งขึ้น

total error ที่เกิดจากการวัดด้วยค่าเฉลี่ย(SSt) ลดลง และ Error ที่เกิดจากการใช้ Model ของเรา คือ SSr ซึ่งโมเดลอธิบายไม่ได้ และจะมี SSm เป็นสิ่งที่โมเดลของเราสามารถอธิบายได้





นิยามของ R-Squared จึงเป็นการที่ SSm/SSt หรือ 1 - SSr/SSt

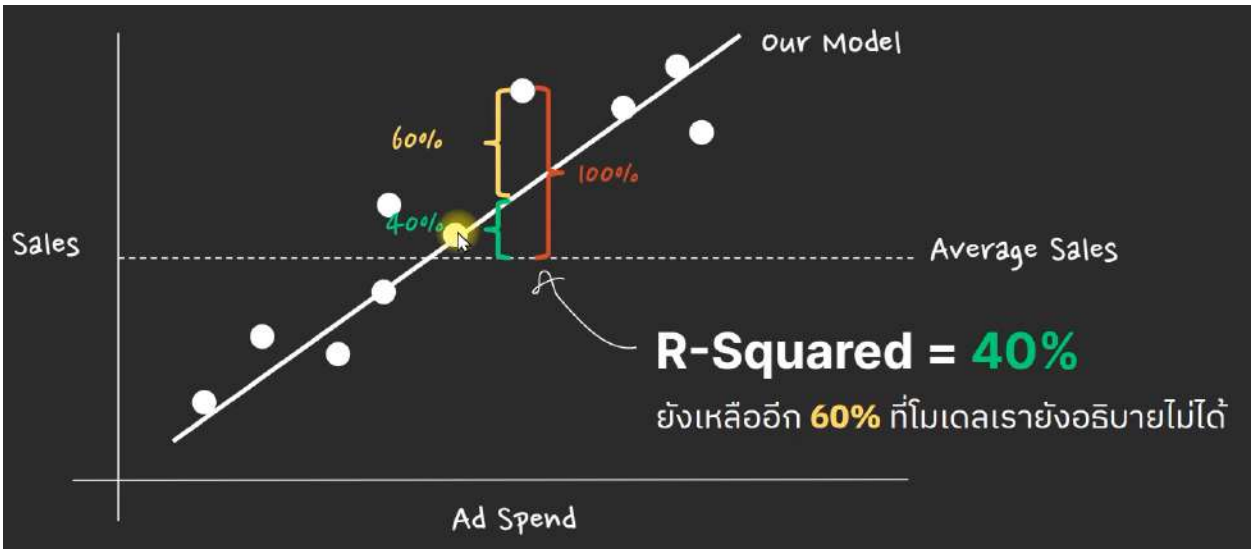
R-Squared จึงเป็นค่าที่บอกว่า Model ของเราอธิบายค่า data ดังภาพได้กี่ % เมื่อเทียบกับการใช้ค่าเฉลี่ยของ data ดังภาพ(ดีดว่ากี่ %)

ยกตัวอย่าง

SSr = 60% SSm =40% SSt = 100%

R-Squared = 40% เหลืออีก 60% ที่โมเดลอธิบายไม่ได้

“Regression Towards The Mean” by Galton and Guass



Let's Do The Work

Program : Excel

รายชื่อ Functions ที่แอดทอยสอนใช้ในคอร์สนี้

- `COVARIANCE.S()` - S ย่อมาจาก Sample สำหรับหาค่า Covariance
- `CORREL()` คำนวณค่า Pearson Correlation
- `INTERCEPT()` หาค่า Intercept ของโมเดล SLR
- `SLOPE()` หาค่า Slope ของโมเดล SLR
- `LINEST()` หาค่า Beta ทั้งหมดของโมเดล Linear Regression ได้ทั้ง SLR, MLR

นอกนั้นจะเป็น Functions ด้าน Math ทั่วไป เช่น `ABS()` `SUM()` หรือการยกกำลังสอง `x^2`

Correlation

	A	B
1	Ad	Sales
2	1.5	500
3	2	650
4	2.2	680
5	2.5	720
6	3	850
7	4	900
8	4.2	920
9	5	1000

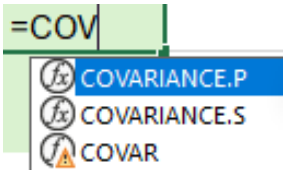
ตัวแปรต้น(x) คือ Ad

ตัวแปรตาม(y) คือ Sales

หาความสัมพันธ์ของตัวแปร 2 ตัวที่เป็นตัวเลข และเป็นความสัมพันธ์เชิงเส้นตรง

ฟังก์ชัน `=COVARIENCE.P` ใช้กับกลุ่ม data ที่เป็น Population (ประชากรทั้งหมดทั้งหมด) **ใช้น้อยมาก**

ฟังก์ชัน `=COVARIENCE.S` ใช้กับกลุ่ม data ที่เป็น Sample (กลุ่มตัวอย่างที่สุ่มมา)



ค่า Covariance ใช้สูตร `=COVARIANCE.S(array1,array2)`

เขียน สูตร `=COVARIANCE.S(A2:A9,B2:B9)` จะได้ 198.429

ค่า Correlation(ตัวย่อ r) คำนวณแบบ ใช้สูตรสำเร็จรูป(Build-in function)

-ค่า Correlation ใช้สูตร `=CORREL(array1,array2)`

เขียน สูตร `=CORREL(A2:A9,B2:B9)` จะได้ 0.96511

ค่า Correlation(ตัวย่อ r) คำนวณแบบ Manual

step 1 คำนวณ standard deviation(sd) ของ Ad `=STDEV.S(A2:A9)` = 1.2284

step 2 คำนวณ standard deviation(sd) ของ Sales `=STDEV.S(B2:B9)` = 167.396

step 3 Correlation = `Covariance/(SD_Ad*SD_Sales)`

`198.429/(1.2284 * 167.396)` จะได้ 0.96511

ค่า R-Squared ในกรณีที่เป็น Simple Linear Regression ใช้สูตร `=(ค่า correlation)^2`

เขียน สูตร `=E3^2` จะได้ 0.93143

Correlation Matrix

เป็นการหาค่า Correlation แบบเป็นคู่ๆ มีความสมมาตรกัน เมื่อพับเป็นสามเหลี่ยม จึงสามารถคำนวณครั้งเดียวได้ ใช้สูตร `=CORREL(array1,array2)`

	A	B	C	D
1	Ad	TV	Radio	Sales
2	1.5	0.2	1	500
3	2	0.5	1	650
4	2.2	0.6	2	680
5	2.5	1	2.2	720
6	3	1.2	3	850
7	4	1.5	3	900
8	4.2	2	3.5	920
9	5	2.1	4	1000

step 1 List ตัวแปรทั้งหมดที่ต้องการหาค่า Correlation ไว้ทั้งแกนนอน - แกนตั้ง หาเป็นคู่ๆ

	Ad	TV	Radio	Sales
Ad	1	0.978986		
TV	0.978986	1		
Radio	0.947157	0.961392	1	
Sales	0.965108	0.964794	0.961496	1

Correlation Test

เราสามารถทดสอบนัยสำคัญของค่า Correlation ด้วย T-Test โดย Hypothesis ของ Correlation เขียนได้ดังนี้

Ho: Correlation = 0
Ha: Correlation ≠ 0

ถ้าค่า `p-value` ของ T-Test มีค่าน้อยหรือเท่ากับ 0.05 (p-value <= 0.05) เราจะ `Reject Ho` และสรุปผลตาม Ha ค่า Correlation ของตัวแปรสองตัวมีนัยสำคัญทางสถิติ

หรือแปลภาษาคนง่ายๆว่า ถ้า p-value <= 0.05 ความสัมพันธ์ของตัวแปรสองตัวที่เราวิเคราะห์น่าจะไม่ใช่เรื่องบังเอิญ เช่น cor(ad, sales) is significance การเพิ่ม/ลดเงินโฆษณามีความสัมพันธ์กับยอดขายของบริษัท

Correlations

[DataSet1] C:\Users\eskim\Desktop\correlation_data.sav

Correlations

		ad	sales
ad	Pearson Correlation	1	.965**
	Sig. (2-tailed)		<.001
	N	8	8
sales	Pearson Correlation	.965**	1
	Sig. (2-tailed)	<.001	
	N	8	8

**. Correlation is significant at the 0.01 level (2-tailed).

ตัวอย่างนี้ `p-value` < 0.001 (reject Ho) ค่า correlation = 0.965 ระหว่าง ad, sales มีนัยสำคัญทางสถิติ

- Positive correlation - ad และ sales มีความสัมพันธ์เชิงบวก เปลี่ยนแปลงในทิศทางเดียวกัน ad เพิ่ม sales มีแนวโน้มจะเพิ่มขึ้นเช่นกัน
- Strong correlation - 0.965 เข้าใกล้ 1 แปลว่าความสัมพันธ์แข็งแกร่งมากๆ เข้ม!

Note - โปรแกรมสถิติส่วนใหญ่จะชอบใช้ `*` เพื่อบอกว่าค่าสถิติตัวใดที่มีนัยสำคัญบ้าง

Two Types of Linear Regression (LR)

What is Y-Intercept

ยกตัวอย่าง โมเดล sales = f(ad spend) เขียนเป็นสมการได้ sales = b0 + b1*ad_spend

b0 หรือ y-intercept ในสมการใช้ดูว่าค่าเฉลี่ยของยอดขาย `average sales` เวลาที่**ไม่ได้**ใช้เงินโฆษณาเลยจะอยู่ที่เท่าไหร่ (ad_spend = 0) (จุดตัดแกน y)

**Note - เวลาทำงานจริง เราไม่ค่อยดูค่า y-intercept เท่าไหร่ ส่วนใหญ่เราจะโฟกัสที่ค่า slope มากกว่า เพราะ slope อธิบายการเปลี่ยนแปลง (ความสัมพันธ์) ของ x และ y ได้

What is Slope

slope ใช้บอกอัตราการเปลี่ยนแปลงของ y ต่อ x เขียนเป็นสมการได้ change in y / change in x one unit

แปลภาษาไทยอีกทีว่า "ถ้า x เปลี่ยนหนึ่งหน่วย y จะเปลี่ยนเท่าไหร่" หรือ "ถ้าเราปรับเงินโฆษณา (ad_spend) 1 ล้านบาท ยอดขาย (sales) จะเปลี่ยนแปลงกี่บาท" อันนี้เรา assume ว่า 1 unit = 1 ล้านบาท

→ ถ้าเราปรับ ad_spend 1 ล้าน แล้วยอดขาย +3.5 ล้าน `slope` จะมีค่าเท่ากับ 3.5/1 = 3.5

Linear Regression

สูตร run Linear Regression เบื้องต้น(ได้ intercept และ slope ในทีเดียว) : `=LINEST(ค่า y, ค่า x)` จะได้ผลลัพธ์เป็น Slope และ Intercept ตามลำดับ

parameter เพิ่มเติม

`=LINEST(ค่า y, ค่า x,TRUE หรือ FALSE)`

- TRUE คือ ให้แสดงค่า Intercept → ใช้บ่อย สามารถ keep เป็น Default ได้เลย
- FALSE คือ ไม่ต้องแสดงผลค่า Intercept

`=LINEST(ค่า y, ค่า x,TRUE หรือ FALSE, TRUE หรือ FALSE)`

- TRUE คือ ให้แสดงค่า Additional Regression statistic ซึ่งรวมถึง R-Squared → ให้ใช้ TRUE
- FALSE คือ ไม่ต้องแสดงผลค่า

9	5	1000		LINEST	131.5341	376.321023	=LINEST(B2:B9,A2:A9,TRUE,TRUE)
10					14.56934	47.4846419	
11				R-Squared	0.931434	47.3447423	
12					81.5074	6	
13					182700.9	13449.1477	196150
14					SSm	SSr	SSt
15				R-Squared =	SSm/SSt		
16					0.931434		
17				R-Squared =	1-SSr/SSt		
18					0.931434		

1. Simple Linear Regression(SLR)

คือโมเดล Linear Regression แบบที่เรียบง่ายที่สุด แต่เป็นพื้นฐานสำคัญในการต่อยอดสถิติขั้นสูงตัวอื่นๆอีกมากมาย

สามารถเขียนด้วยสมการดังนี้

สมการเส้นตรงที่เราเคยเรียน y = mx + c

โดยที่ c คือค่า constant หรือ `y-intercept`

ส่วน m คือ `slope`

สมการแบบนักสถิติคือ y = b0 + b1*x

โดยที่ b0 คือ `y-intercept`

b1 คือ `slope` หรือความชันของเส้นตรง

ตัวอย่างการเขียนโมเดล :

Sales = f(Ad)

→ Sales = b0 + b1*Ad

**มี 1 ตัวแปรต้น

การใช้สูตร formular

หาค่า intercept ด้วยสูตร `=INGTERCEPT(ค่า y, ค่า x)`

หาค่า slope ด้วยสูตร `=SLOPE(ค่า y, ค่า x)`

สูตร run Linear Regression เบื้องต้น(ได้ intercept และ slope ในทีเดียว) :

`=LINEST(ค่า y, ค่า x)` จะได้ผลลัพธ์เป็น Slope และ Intercept ตามลำดับ

หรือใช้สูตร `=LINEST(ค่า y, ค่า x,TRUE หรือ FALSE, TRUE หรือ FALSE)` เพื่อแสดงค่าทางสถิติทั้งหมด

	A	B	C	D	E	F	G	H	I
1	Ad	Sales							
2	1.5	500		**COV**	198.4286		=COVARIANCE.S(A2:A9,B2:B9)		
3	2	650		**COR**	0.965108		=CORREL(A2:A9,B2:B9)		
4	2.2	680		**R2**	0.931434		=E3^2		
5	2.5	720							
6	3	850		**INTERCEPT**	376.321		=INTERCEPT(B2:B9,A2:A9)		
7	4	900		**SLOPE**	131.5341		=SLOPE(B2:B9,A2:A9)		
8	4.2	920							
9	5	1000		**LINEST**	131.5341	376.321023	=LINEST(B2:B9,A2:A9,TRUE,TRUE)		
10					14.56934	47.4846419			
11				**R-Squared**	0.931434	47.3447423			
12					81.5074	6			
13					182700.9	13449.1477	196150		
14					**SSm**	**SSr**	**SSt**		
15				**R-Squared =**	**SSm/SSt**				
16					0.931434				
17				**R-Squared =**	**1-SSr/SSt**				
18					0.931434				

2. Multiple Linear Regression

ตัวอย่างการเขียนโมเดล :

Sales = f(Ad, TV, Radio)

**มีตัวแปรต้นมากกว่า 1 ตัว

→ Sales = b0 + b1*Ad + b2*TV + b3*Radio

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

**ปัจจุบันยังไม่มีวิธีการ plot chart ด้วยค่าดิบ จะใช้เป็นการ plot ค่า error แทน (เรายังไม่เรียนลึกถึงขั้นนั้น)

สูตร run Linear Regression

```
=LINEST(ตัวแปร y, ตัวแปร x ทั้งหมด ทุกคอลัมน์,TRUE หรือ FALSE, TRUE หรือ FALSE)
```

**มัก set ให้เป็น TRUE ทั้งหมด

**เวลาแสดงผล ตัวแปรต้นจะเรียงจากขวามาซ้าย ดังภาพ

	A	B	C	D	E	F	G	H	I
1	Ad	TV	Radio	Sales		**Radio**	**TV**	**Ad**	**Intercept(b0)**
2	1.5	0.2	1	500		62.60151	33.69569	59.3771025	403.9147799
3	2	0.5	1	650		59.35669	148.3343	72.1931809	91.85579992
4	2.2	0.6	2	680	**R-Squared**	0.953853	47.5702	#N/A	#N/A
5	2.5	1	2.2	720		27.55996	4	#N/A	#N/A
6	3	1.2	3	850		187098.3	9051.696	#N/A	#N/A
7	4	1.5	3	900		**SSm**	**SSr**	**SSt**	
8	4.2	2	3.5	920	**R-Squared**	0.953853		196150	
9	5	2.1	4	1000		0.953853			

อ่านผล regression coefficients

Radio	TV	Ad	Intercept
62.6015	33.6957	59.3771	403.915

จะเห็นว่า ค่า slope ทั้ง 3 ตัว มีค่าเป็นบวกทั้งหมด จึงมีความสัมพันธ์ไปในทางเดียวกัน

วิธีการอ่าน

ถ้าสมมติเราเพิ่ม Ad 1 unit(หรือ 1 บาท) เราจะเพิ่มยอดขายได้ 59 บาท ปัจจัยอื่นคงที่ (all other variables held constant) ก็คือ Radio และ TV คงที่

ถ้าเราเพิ่มเงินโฆษณาในวิทยุ 1 บาท เราจะเพิ่มยอดขายได้ 62.6 บาท ปัจจัยอื่นๆคงที่ (TV Ad)

**ในการทำโมเดล ไม่สามารถทำให้ทำให้ทุกตัวเปลี่ยนพร้อมกันได้เหมือนในชีวิตจริง จึงได้เป็น Assumption ให้เรา fix ว่าตัวแปรอื่นๆไม่เปลี่ยนแปลง (Isolate เจาะจงตัวแปรออกมา)

Model Prediction

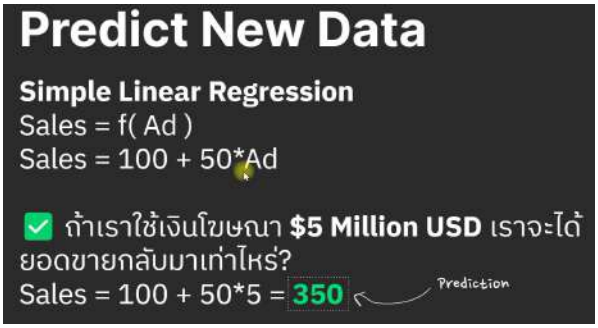
หลังจากได้สมการ Regression เราสามารถนำสมการ (หรือโมเดล) นี้ไปใช้ทำนาย Input ใหม่ได้เลย นี่คือเหตุผลที่ Regression Models ถูกนำไปใช้ในงาน Machine Learning เยอะมาก

ความแตกต่างของ Statistics vs. ML สำหรับแอดทอยคือ

- สถิติเน้นความเข้าใจ ตัวแปรในสมการส่งผลยังไงกับตัวแปรตาม
- ML เน้นการทำนายผล เอาแม่นไว้ก่อน ไส้ในค่อยว่ากัน (black box)

เวลาเราเพิ่ม หรือ ลด Ad จะมีแนวโน้มที่ Sales จะเปลี่ยนแปลง เท่าไหร่ (จากภาพคือ 50)

→ การ predict คือ การที่เราเอาเลขไปแทนใน Ad เพื่อทำนายออกมา

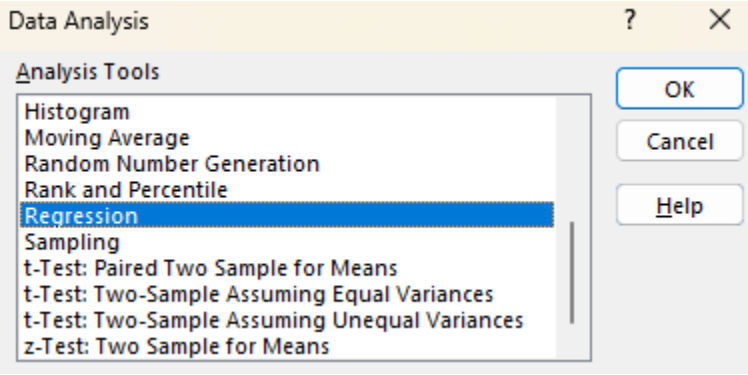


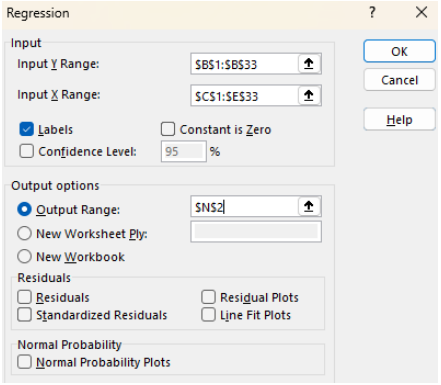
Analysis Toolpak

ใช้เพื่อ run Full Regression Result โดยที่ไม่ต้องเขียนโค้ด

เราจะใช้ hp wt am ในการทำนาย mpg → `mpg ~ f(hp + wt + am)`

Ribbon : Data → Data Analysis(ขวาสุด) → Regression (default จะเป็น Linear Regression)





	A	B	C	D	E	M	N	O	P	Q	R	S	T	U	V
1	model	mpg	hp	wt	am										
2	Mazda RX4	21	110	2.62	1		SUMMARY OUTPUT								
3	Mazda RX4	21	110	2.875	1										
4	Datsun 710	22.8	93	2.32	1		*Regression Statistics*								
5	Hornet 4 D	21.4	110	3.215	0		Multiple R	0.916455294							
6	Hornet Sp	18.7	175	3.44	0		R Square	0.839890305							
7	Valiant	18.1	105	3.46	0		Adjusted R Square	0.822735695							
8	Duster 360	14.3	245	3.57	0		Standard Error	2.53751194							
9	Merc 240D	24.4	62	3.19	0		Observations	32							
10	Merc 230	22.8	95	3.15	0										
11	Merc 280	19.2	123	3.44	0		ANOVA								
12	Merc 280C	17.8	123	3.44	0			*df*	*SS*	*MS*	*F*	*Significance F*			
13	Merc 450S	16.4	180	4.07	0		Regression	3	945.7561158	315.252	48.96003	2.90787E-11			
14	Merc 450S	17.3	180	3.73	0		Residual	28	180.2910717	6.438967					
15	Merc 450S	15.2	180	3.78	0		Total	31	1126.047188						
16	Cadillac F	10.4	205	5.25	0										
17	Lincoln Co	10.4	215	5.424	0			*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
18	Chrysler Ir	14.7	230	5.345	0		Intercept	34.00287512	2.642659337	12.86692	2.82E-13	28.58963286	39.41611738	28.58963286	39.41612
19	Fiat 128	32.4	66	2.2	1		hp	-0.037478726	0.009605422	-3.90183	0.000546	-0.057154541	-0.01780291	-0.05715454	-0.017803
20	Honda Civ	30.4	52	1.615	1		wt	-2.878575414	0.904970538	-3.18085	0.003574	-4.732323527	-1.0248273	-4.73232353	-1.024827
21	Toyota Co	33.9	65	1.835	1		am	2.08371013	1.376420152	1.513862	0.141268	-0.73575874	4.903179	-0.73575874	4.903179

Reading Full Regression Results

อันดับแรก ในตารางที่สอง ANOVA

- ถ้าค่า `F-Test` ไม่ซิก/Significant (p-value > 0.05) : ไม่ต้องดูตารางสุดท้าย เพราะไม่มีตัวแปรต้นตัวใดมีนัยสำคัญเลย (*"ตัวแปรต้น x1, x2, x3, ... ทั้งหมดที่ใส่ไปในโมเดล ไม่สามารถใช้ทำนาย y ได้เลย"*)

- ถ้า F-Test มีนัยสำคัญ (p-value <= 0.05) : ให้ไปดู coefficients ในตารางสุดท้าย

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.91645529								
R Square	0.83989031								
Adjusted R Square	0.82273569								
Standard Error	2.53751194								
Observations	32	จำนวนแถวที่ใช้สร้างโมเดล							
ANOVA คือการทดสอบ Overall significant of the model									
	df	*SS*	*MS*	*F*	*Significance F*				
Regression	3	945.7561158	315.252	48.96	0.00000	<0.05	แปลว่า มีตัวแปรต้น ตัวแปรใดตัวแปรหนึ่งที่มีนัยสำคัญทางสถิติ สามารถใช้ในการทำนายตัวแปรตามได้จริง		
Residual	28	180.2910717	6.43897						
Total	31	1126.047188							
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	
Intercept	34.00	2.642659337	12.8669	0.00000	28.5896329	39.4161174	28.5896329	39.416117	
hp	-0.04	0.009605422	-3.90183	0.00055	-0.0571545	-0.0178029	-0.0571545	-0.017803	
wt	-2.88	0.904970538	-3.18085	0.00357	-4.7323235	-1.0248273	-4.7323235	-1.024827	
am	2.08	1.376420152	1.51386	0.14127	-0.7357587	4.903179	-0.7357587	4.903179	
				ตัวแปรที่มี p-value < 0.05 จะมีนัยสำคัญทางสถิติ					
				ในตารางนี้ ได้แก่ hp, wt					

สรุป

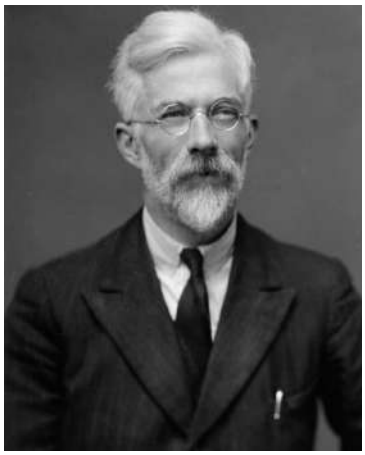
โมเดลตัวนี้ที่เราสร้างขึ้นมาจากตัวแปร 3 ตัว จะมีตัวแปร 2 ตัวที่มีนัยสำคัญทางสถิติ R-Squared ของ Overall model เท่ากับ 83.9%

ตัวแปร 2 ตัวที่มีนัยสำคัญ มีความสัมพันธ์เชิงลบกับ mpg

**ปกติแอดจะดู `Standard Error` ของโมเดลด้วย แต่ยังไม่กล่าวถึงในบทนี้เพราะจะเกี่ยวข้องกับ Central Limit Theorem (ยังไม่สอนใน The curious)

p-value 5% มาจากไหน ทำไมต้อง 5%?

ช่วงปี 1920s เลย ท่าน Sir. Ronald Fisher นักสถิติในตำนานได้เขียน Paper อธิบายเรื่องการทำ significance test ไว้



ถ้าเราทำการทดลอง 20 ครั้ง แล้วผลลัพธ์ออกมา reliable เหมือนเดิม 19 ครั้ง ผิดพลาดไป 1 ครั้งถือว่าเรายังรับได้กับผลลัพธ์นั้น

ผลลัพธ์ reliable (หรือ consistent) 19/20 ครั้ง = มั่นใจ 95% และเกิดข้อผิดพลาดได้ 1/20 ครั้ง = 5%

นักสถิติสาย Frequentist (Fisherian) เลยใช้ค่า 5% เหมือนประเพณีสืบต่อกันมาร้อยปีแล้ว จริงๆมันเปลี่ยนได้นะ 555+ จะใช้ 1%, 5%, 10% หรือ 22% ก็แล้วแต่งานเราเลยครับ

จริงๆระบบ NHST หรือ Null Hypothesis Significance Testing ได้รับการพัฒนาต่อจาก Fisher โดย Egon Pearson (ลูกชายของ Karl Pearson) และ Jerzy Neyman ในช่วงปี 1930s

จนกลายเป็นระบบ Sig Test ที่เราใช้กันมาถึงทุกวันนี้

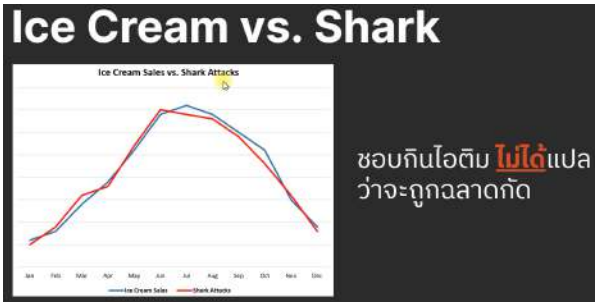
Ref: https://en.wikipedia.org/wiki/Ronald_Fisher

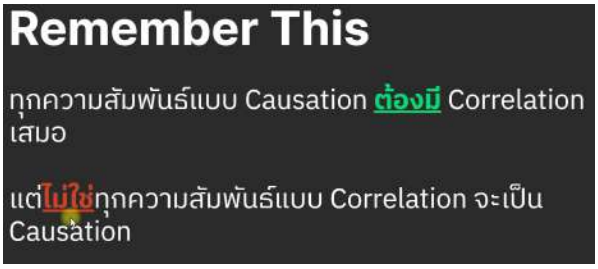
Correlation Does Not Imply Causation

ข้อควรระวังในการใช้โมเดลตระกูล Regression หรือ Correlation

Correlation does not imply causation

: ตัวแปรสองตัวมีความสัมพันธ์กัน(Correlation) ไม่ได้แปลว่าส่งผลต่อกันและกัน เช่น x ไม่ได้ก่อให้เกิด y (Causation : X cause Y or Y cause X)



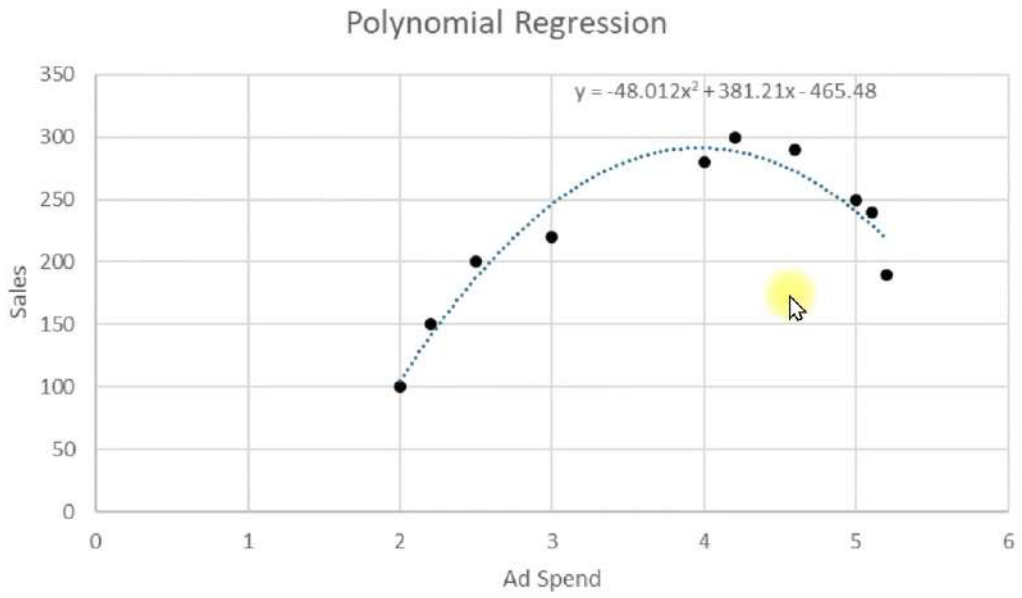


ถ้าเราอยากจะพิสูจน์ `Causation` วิธีมาตรฐานทางสถิติคือการทำ Experimentation เช่น RCT มี control และ test/ treatment ทดสอบและบันทึกผลอย่างจริงจัง

Other Types of Regression

สำหรับโมเดล Regression นอกจาก Linear Regression ที่เราเรียนในคอร์สนี้แล้ว โมเดลอื่นๆที่เรานิยมใช้กันจะมีอีกสองตัวคือ `Polynomial` Regression และ `Logistic` Regression

Polynomial Regression



เมื่อเราเจอกับ Non-Linear Data หรือข้อมูลที่มีลักษณะ Non-Linear Relationship เราทำการบิด Linear Regression เพื่อให้ fit กับ data

เช่น เมื่อเรายิงโฆษณาเยอะเกินไป เมื่อถึงจุดจุดหนึ่งกระแสตอบรับอาจลดลงได้

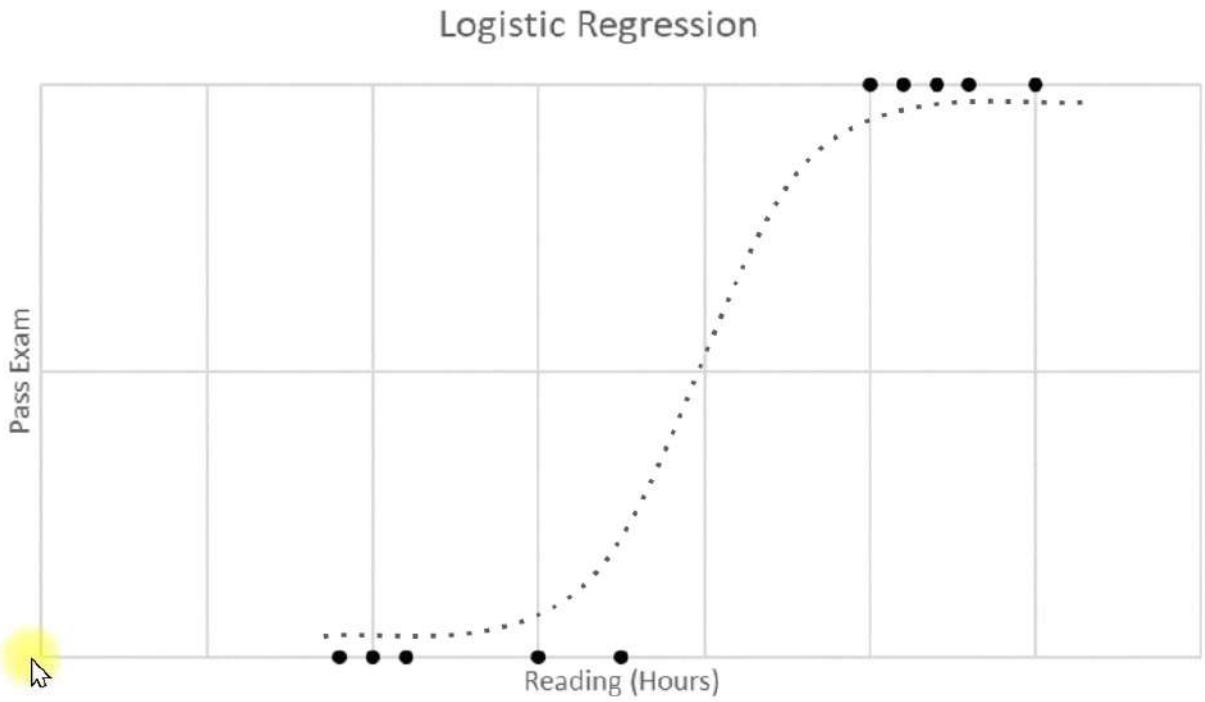
→ เราจึงบิด Linear Regression เป็น Polynomial Degree 2(x^2)

วิธีการบิด : นำ Ad Spend มายกกำลัง 2 สร้างเป็นคอลัมน์ใหม่ขึ้นมา → run linear regression กับตัวแปรใหม่ตัวนี้ด้วย เพื่อ estimate ค่า Co-efficient ขึ้นมาใหม่ → ได้โมเดลใหม่ขึ้นมาที่เริ่มเป็นเส้นโค้ง

**เราเรียกว่า Quadratic function : ใส่ตัวแปรใหม่ที่เป็น degree 2 เข้าไปใน Linear regression

Logistic Regression

เวลาเจอกับตัวแปรตาม y แบบ binary (0/1) เช่น สอบผ่านหรือไม่ผ่าน จะซื้อหรือไม่ซื้อ เป็นต้น ก็คือมีลักษณะเป็น yes/no Question



Take-Home Cheat Sheets

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 $1 - SS_{residual} / SS_{total}$