機器學習
Multiple Linear Regression

授課老師：林彥廷

多元線性回歸
Multiple Linear Regression

- 多元線性回歸分析（又稱複回歸分析或多變項回歸分析）是建立因變量（Y）與自變量（X）的關係模型，利用對自變量的觀察，評估因變量的變化
- 多元線性回歸涉及一個因變量與兩個或以上的自變量

回歸 Regression

回歸 Regression

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

回歸 Regression

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Dependent variable (DV)

Multiple Linear Regression

回歸 Regression

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Dependent variable (DV)

Independent variable (IV)

Multiple Linear Regression

回歸 Regression

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

回歸 Regression

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

Dependent variable (DV)

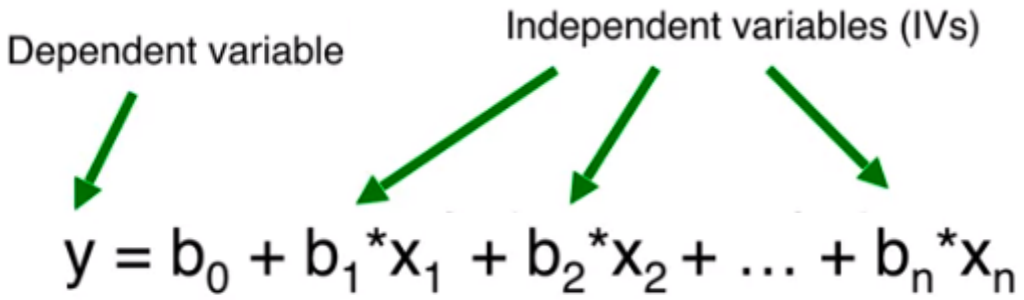
$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

回歸 Regression

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

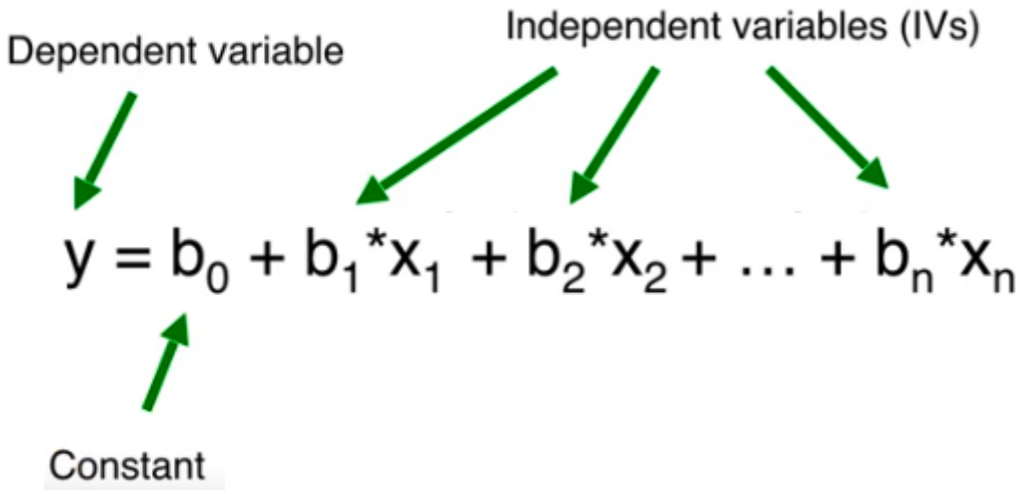


回歸 Regression

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

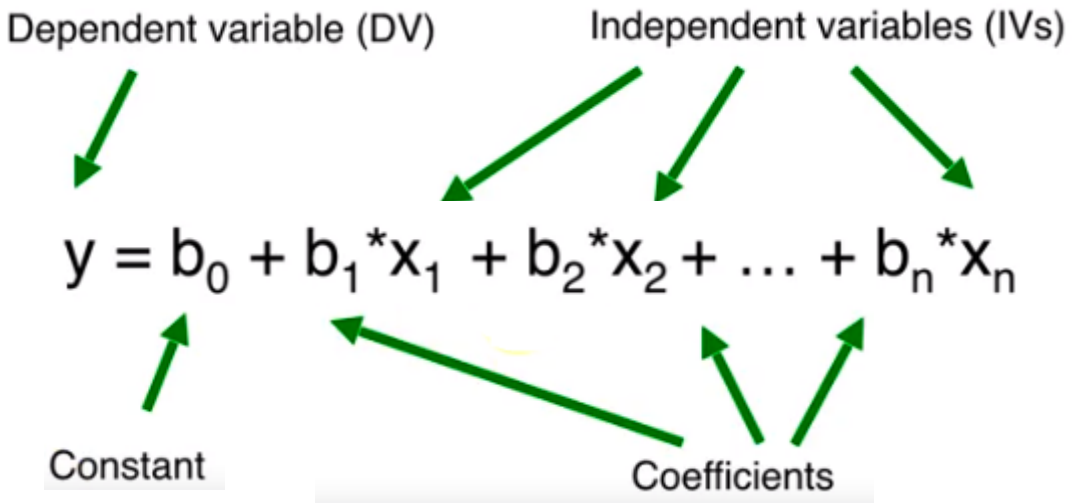


回歸 Regression

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression



多元線性回歸－注意事項

Assumptions of a Linear Regression:

1. Linearity	線性關係
2. Homoscedasticity	變項同質性
3. Multivariate normality	多元常態性
4. Independence of errors	誤差獨立性
5. Lack of multicollinearity	無多重共線性

Dummy Variables 虛擬變量

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

Dummy Variables 虛擬變量

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0$$

Dummy Variables 虛擬變量

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1$$

Dummy Variables 虛擬變量

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2$$

Dummy Variables 虛擬變量

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3$$

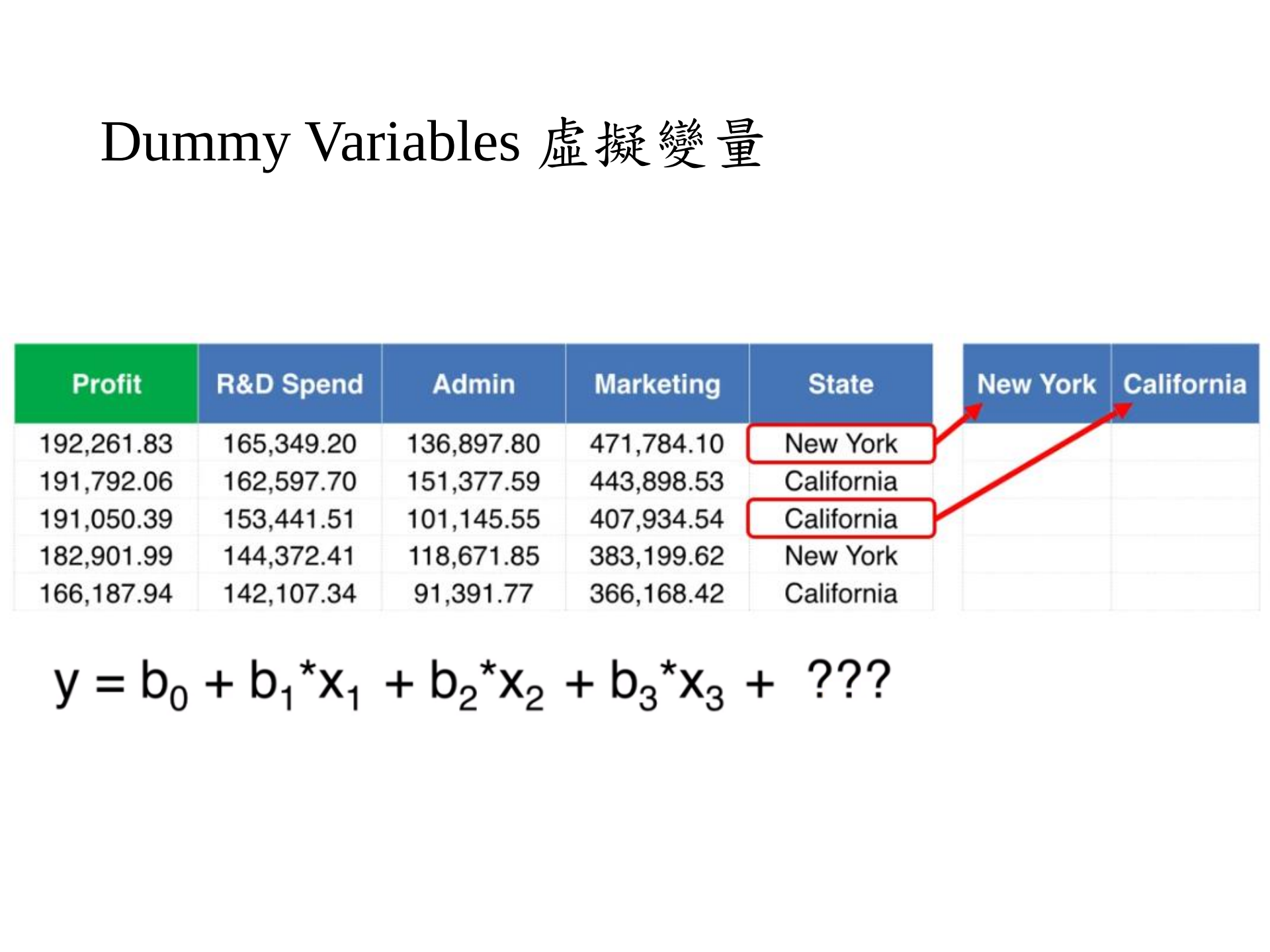
Dummy Variables 虛擬變量

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \; ???$$

Dummy Variables 虛擬變量

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York		
191,792.06	162,597.70	151,377.59	443,898.53	California		
191,050.39	153,441.51	101,145.55	407,934.54	California		
182,901.99	144,372.41	118,671.85	383,199.62	New York		
166,187.94	142,107.34	91,391.77	366,168.42	California		



$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ ???$$

Dummy Variables 虛擬變量

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	
191,792.06	162,597.70	151,377.59	443,898.53	California	0	
191,050.39	153,441.51	101,145.55	407,934.54	California	0	
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	
166,187.94	142,107.34	91,391.77	366,168.42	California	0	

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ ???$$

Dummy Variables 虛擬變量

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \; ???$$

Dummy Variables 虛擬變量

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

Dummy Variables

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1$$

Dummy Variables 虛擬變量

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

Dummy Variables

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1$$

Dummy Variables Trap 虛擬變量陷阱

Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + b_3{*}x_3 + b_4{*}D_1$$

Dummy Variables Trap 虛擬變量陷阱

Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + b_3{*}x_3 + b_4{*}D_1 + b_5{*}D_2$$

Dummy Variables Trap 虛擬變量陷阱

Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136	471	New York	1	0
191,792.06	162,597.70	151		rnia	0	1
191,050.39	153,441.51	101		rnia	0	1
182,901.99	144,372.41	118		York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$D_2 = 1 - D_1$

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + b_3{*}x_3 + b_4{*}D_1 + b_5{*}D_2$$

Dummy Variables Trap 虛擬變量陷阱

Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + b_3{*}x_3 + b_4{*}D_1 + b_5{*}D_2$$

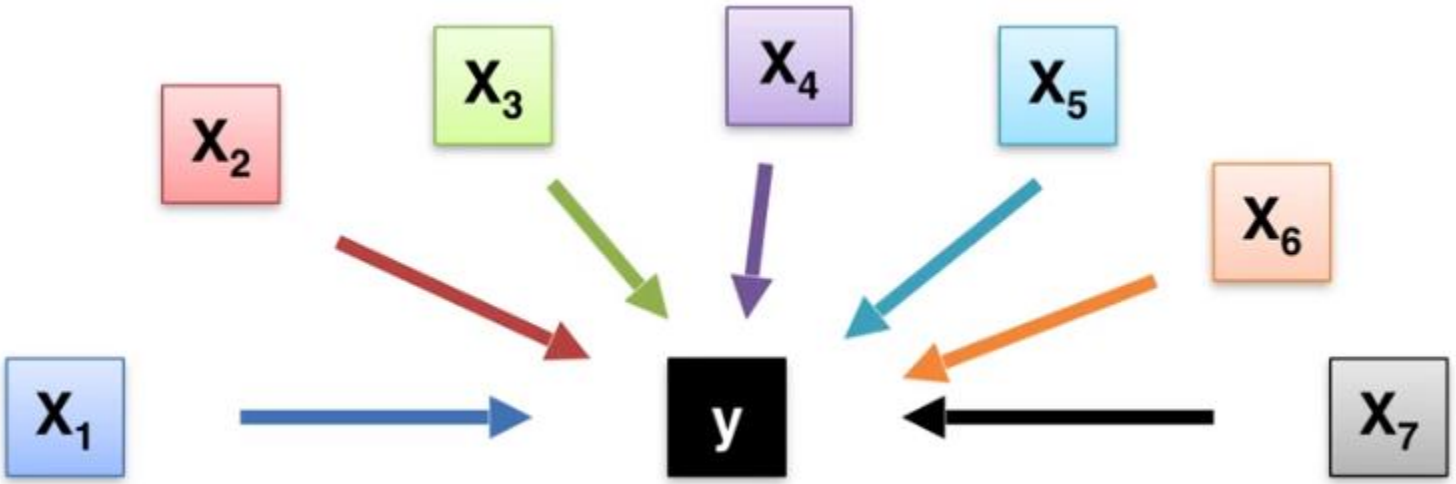
Dummy Variables Trap 虛擬變量陷阱

Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

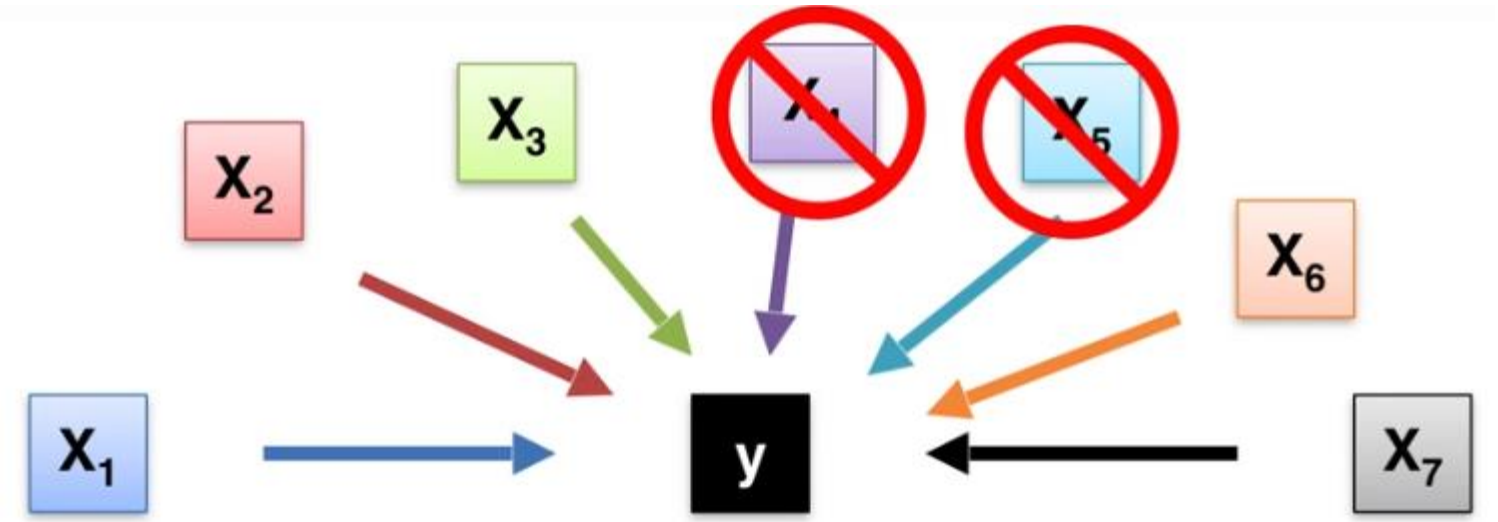
$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + b_3{*}x_3 + b_4{*}D_1 + b_5{*}D_2$$

Always omit one dummy variable

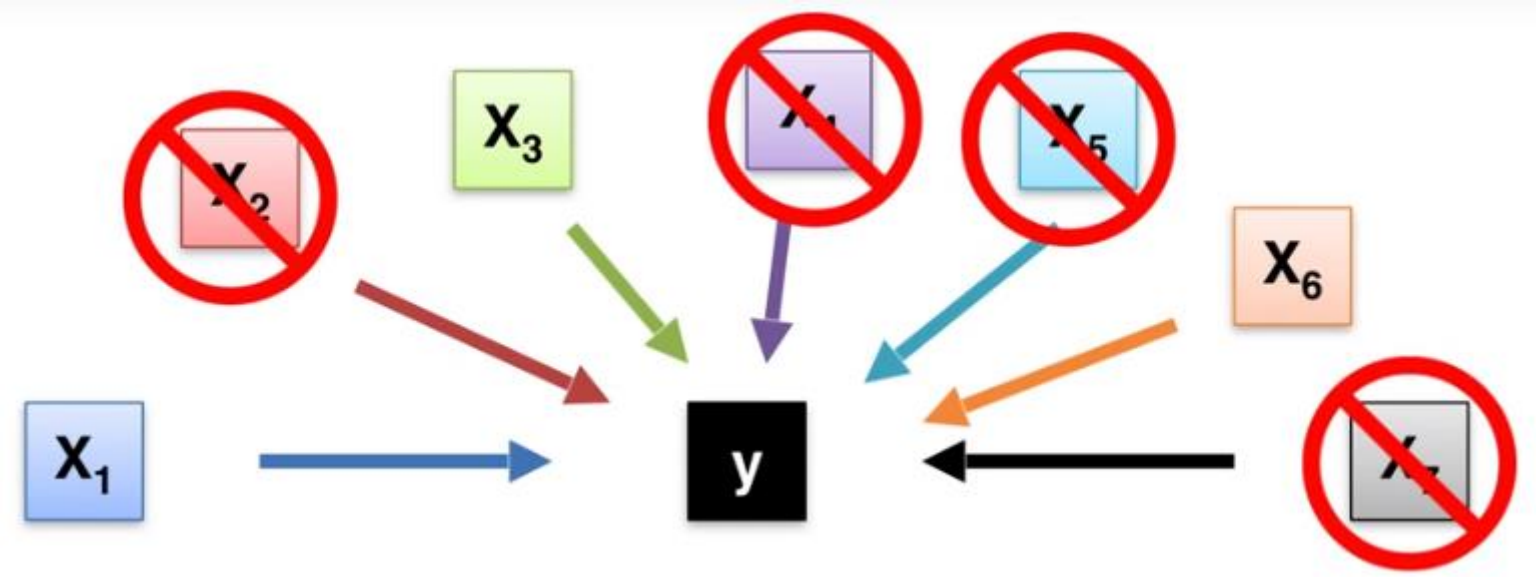
建立模型



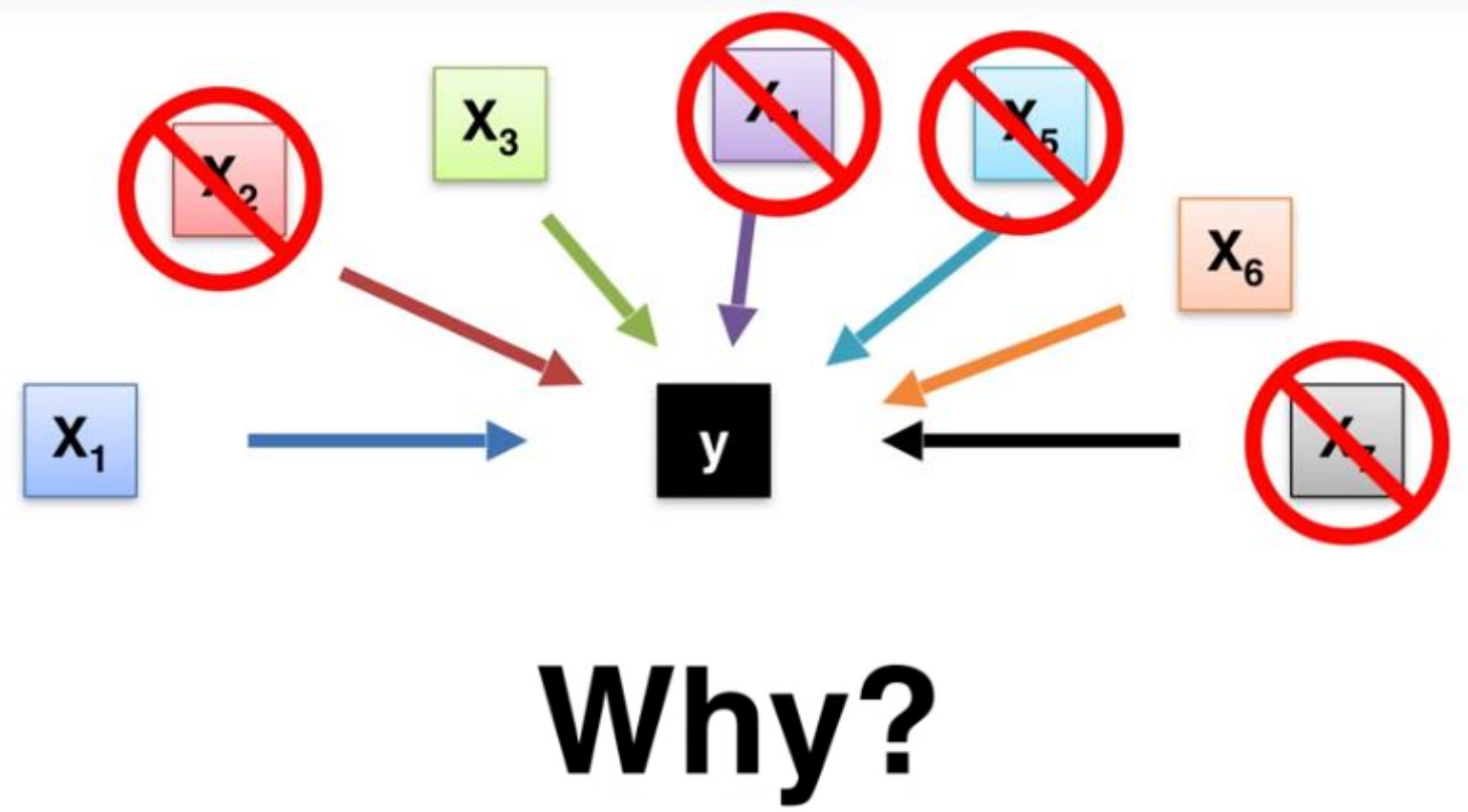
建立模型



建立模型

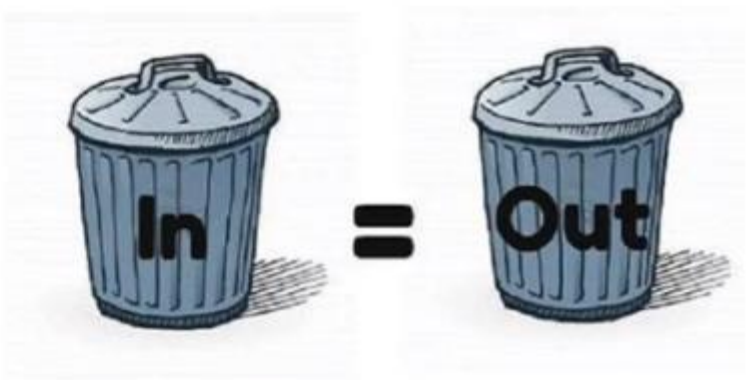


建立模型



建立模型

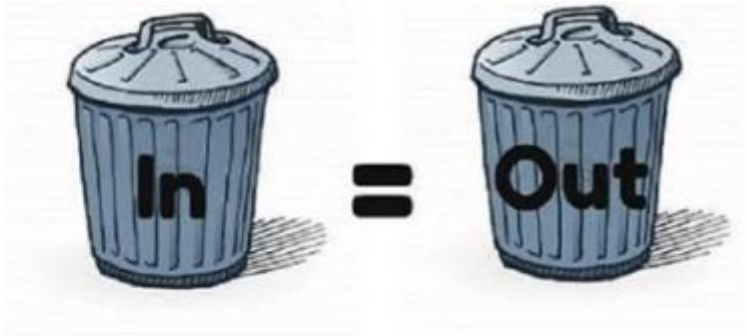
1)



2)

建立模型

1)



2)

建立模型

5 methods of building models:

1. All-in
2. Backward Elimination
3. Forward Selection
4. Bidirectional Elimination
5. Score Comparison

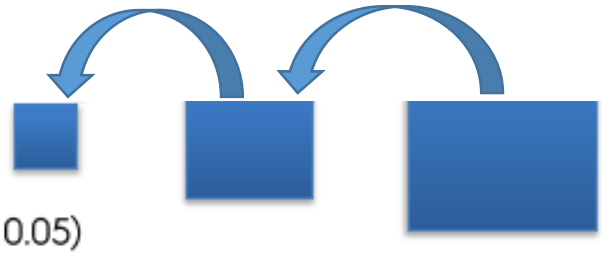
} Stepwise Regression

建立模型

"All-in" – cases:

- Prior knowledge; OR
- You have to; OR
- Preparing for Backward Elimination

建立模型



Backward Elimination

STEP 1: Select a significance level to stay in the model (e.g. SL = 0.05)

STEP 2: Fit the full model with all possible predictors

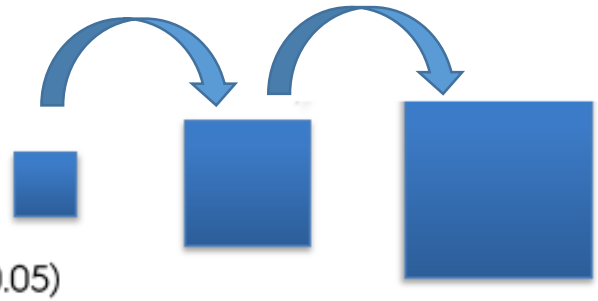
STEP 3: Consider the predictor with the highest P-value. If P > SL, go to STEP 4, otherwise go to FIN

STEP 4: Remove the predictor

STEP 5: Fit model without this variable*

FIN: Your Model Is Ready

建立模型



Forward Selection

STEP 1: Select a significance level to enter the model (e.g. SL = 0.05)

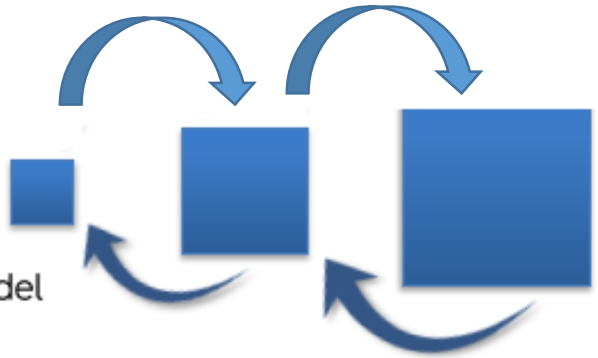
STEP 2: Fit all simple regression models $y \sim x_n$ Select the one with the lowest P-value

STEP 3: Keep this variable and fit all possible models with one extra predictor added to the one(s) you already have

STEP 4: Consider the predictor with the lowest P-value. If P < SL, go to STEP 3, otherwise go to FIN

FIN: Keep the previous model

建立模型



Bidirectional Elimination

STEP 1: Select a significance level to enter and to stay in the model
e.g.: SLENTER = 0.05, SLSTAY = 0.05

STEP 2: Perform the next step of Forward Selection (new variables must have: P < SLENTER to enter)

STEP 3: Perform ALL steps of Backward Elimination (old variables must have P < SLSTAY to stay)

STEP 4: No new variables can enter and no old variables can exit

FIN: Your Model Is Ready

建立模型

Score Comparison

STEP 1: Select a criterion of goodness of fit (e.g. Akaike criterion)

STEP 2: Construct All Possible Regression Models: 2^N-1 total combinations

STEP 3: Select the one with the best criterion

FIN: Your Model Is Ready

建立模型

Score Comparison

STEP 1: Select a criterion of goodness of fit (e.g. Akaike criterion)

STEP 2: Construct All Possible Regression Models: 2^N-1 total combinations

STEP 3: Select the one with the best criterion

FIN: Your Model Is Ready

Example:
10 columns means
1,023 models

THE END

ytlin@mail.nptu.edu.tw