



探討非線性關係

1

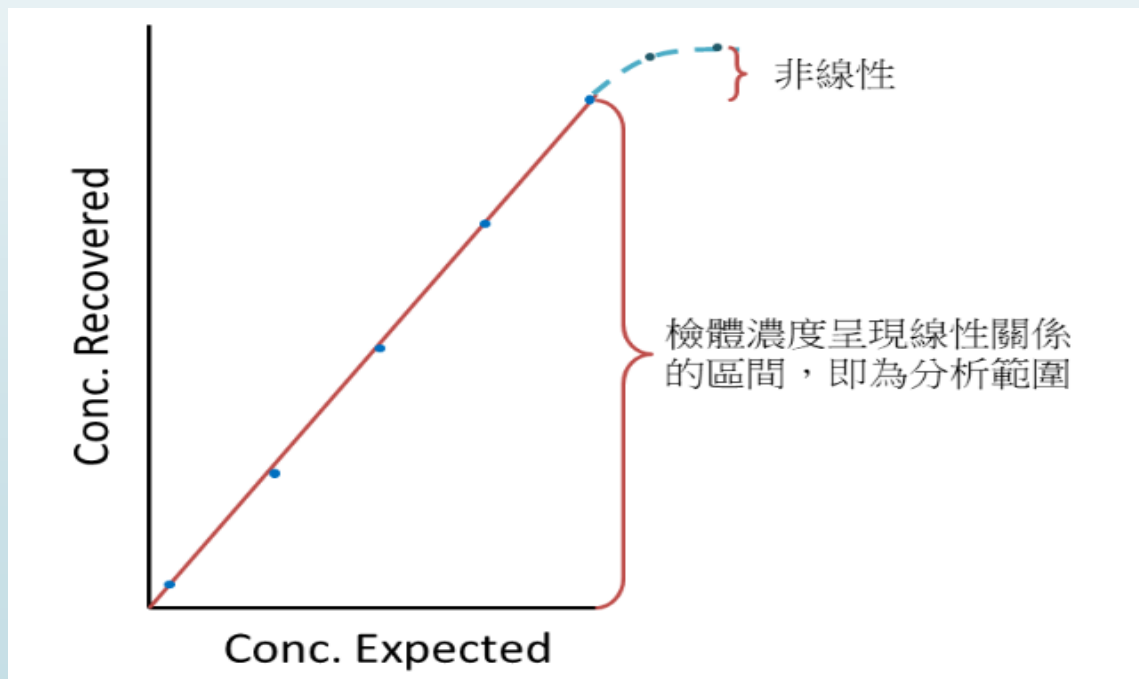


醫學研究部 基礎醫學科 生統小組：陳韻仔 博士

授課日期：115年6月9日

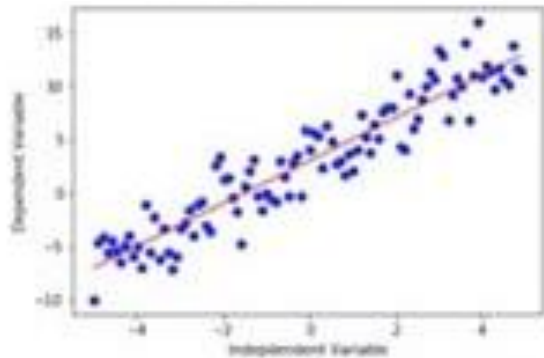
臨床試驗的效能評估項目： 線性關係 v.s. 非線性關係

- ◆ 以定量檢驗為例，**實際值與測量值呈線性關係**的區間為分析範圍。
- ◆ 臨床檢體經稀釋後量測，符合線性關係的區間則稱為**臨床可報告範圍**。
- ◆ **測量範圍 (Measuring Interval Range)** 通常以分析範圍 (Analytical Measurement Range, AMR) 或臨床可報告範圍 (Clinical Reportable Range, CRR) 表示。

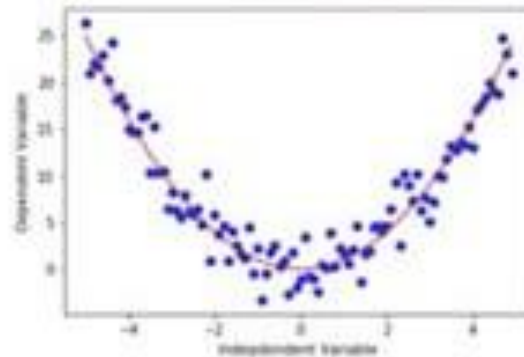


Different Types of Regression

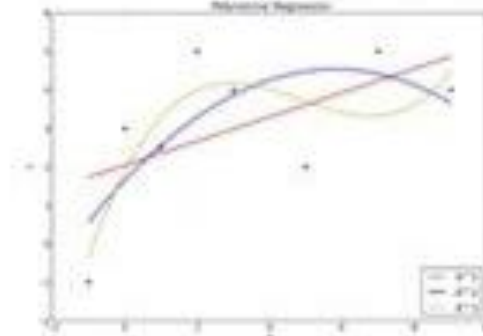
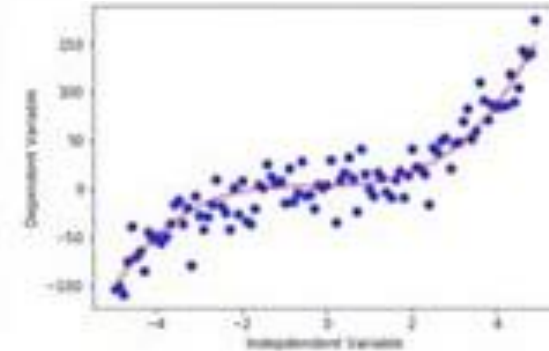
線性迴歸模型
Linear Regression



二次式迴歸模型
Quadratic Regression



三次 (立方) 迴歸模型
Cubic Regression



Linear Model v.s. Non-Linear (Smooth) Model

► 傳統迴歸分析模型的重要假設

- 自變項和依變項呈線性關係，然實際收案資料在分析時甚少能符合此假設條件
- 若能以模式模擬自變量與依變項之間的非線性關係，將可更忠實呈現自變項和依變項的相關

► 非線性模型

- **Restricted Cubic Spline (RCS) Model v.s. Fractional Polynomials (FP)**
v.s. Generalized Additive Model (GAM)
- 可清楚描述自變項與依變項之間的關係（無論線性或非線性關係）
- 適用於各分析模式（包括線性迴歸、邏輯斯迴歸、存活）

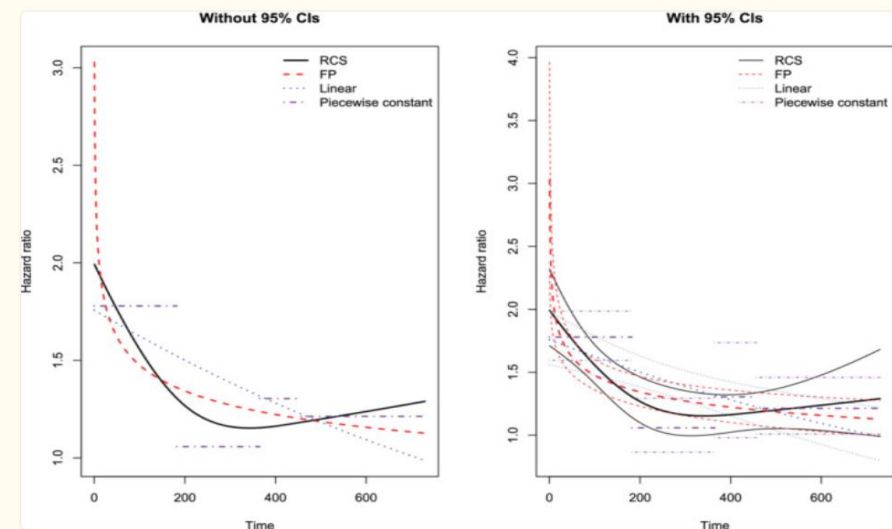
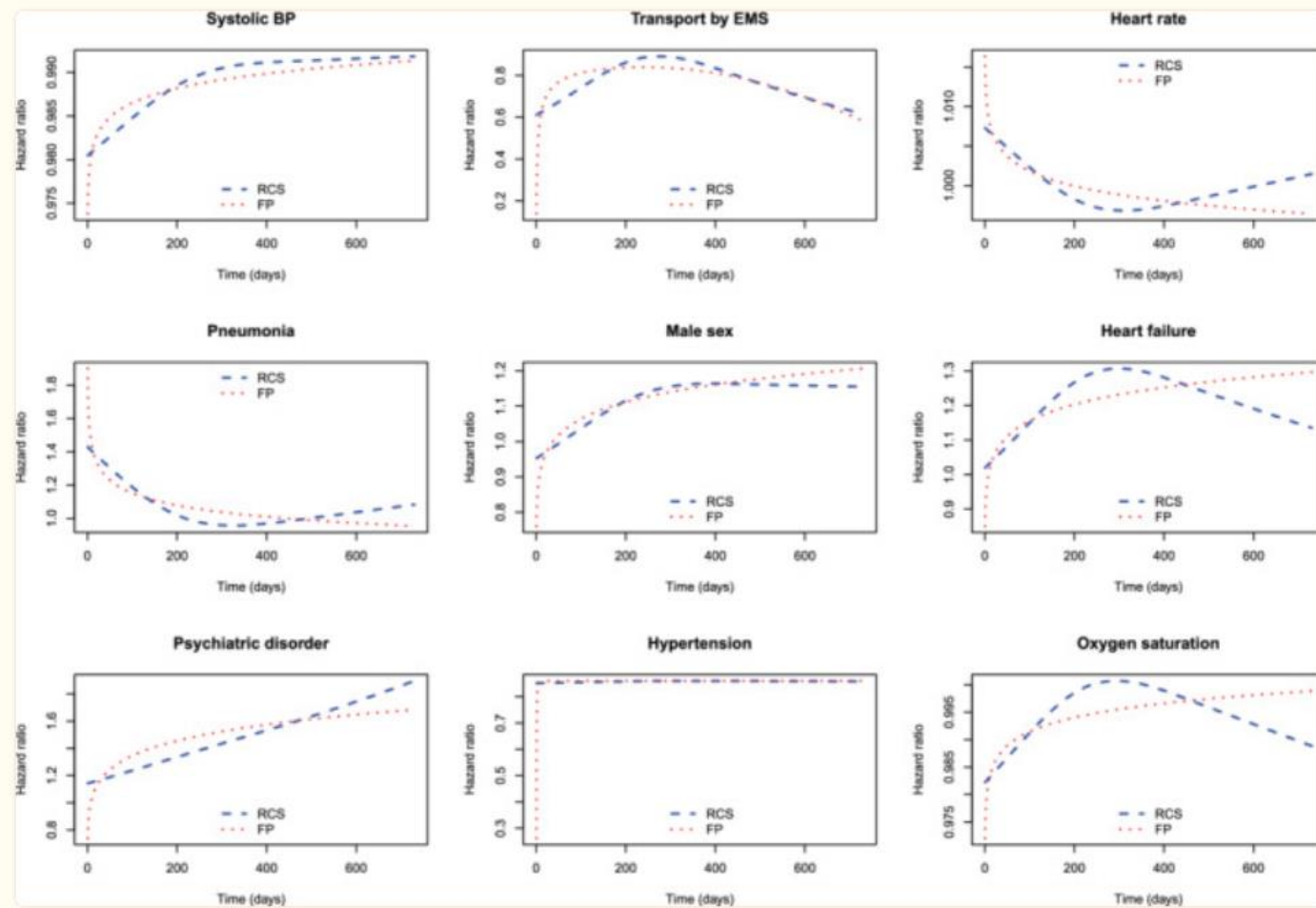
非線性模型的比較

5

特性	限制性立方樣條 (RCS)	分數多項式 (FP)	廣義加法模型 (GAM)
說明	使用立方平滑函數來近似數據，通過選擇適當的節點和平滑度來擬合數據。	使用多項式函數來近似數據，選擇最適合的多項式次數和係數來擬合數據。	使用平滑函數來建模關係，允許非線性關係的形式未被指定。
非線性建模方式	使用立方平滑函數	使用多項式函數	使用平滑函數
多項式次數選擇	不適用	根據模型配度選擇	不適用
節點數選擇	需要選擇適當的節點	不適用	不適用
適用性	在大型數據集上工作較好	在較小數據集上工作較好	在複雜數據集上工作較好
彈性	較有彈性，可以擬合不同形式的非線性關係	較有彈性，可以擬合不同形式的非線性關係	高度彈性，可以擬合複雜的非線性關係
選擇最佳模型	通過比較不同節點和平滑度來選擇最佳模型	通過比較不同多項式次數選擇最佳模型	通過比較不同平滑函數來選擇最佳模型
適合執行的統計軟體	R, SAS, Stata	R, Stata	R, Python, SAS
SAS 使用的模組	PROC TRANSREG	不適用	PROC GAM
R 使用的模組	splines	mfp	mgcv
Stata 使用的模組	mkspline	fp	不適用

當代主流的非線性關係

- Fractional polynomials (FP)
- Restricted cubic spline (RCS)



RESEARCH

Open Access



Body mass index at baseline directly predicts new-onset diabetes and to a lesser extent incident cardio-cerebrovascular events, but has a J-shaped relationship to all-cause mortality

Yoon-Jong Bae¹, Sang-Jun Shin¹ and Hee-Taik Kang^{2,3*}

Table 2 Cox-proportional hazards regression model for the incidence of diabetes, cardio-cerebrovascular diseases, or all-cause mortality according to body mass index category

		Underweight (BMI < 18.5 kg/m ²)	Normal (18.5–< 23 kg/m ²)	Overweight (23–< 25 kg/m ²)	Grade 1 Obesity (25–< 30 kg/m ²)	Grade 2 Obesity (30–< 35 kg/m ²)	Grade 2 Obesity (≥ 35 kg/m ²)
Primary Outcomes^a	Men	1.293 (1.224–1.365)	1	1.101 (1.073–1.129)	1.320 (1.288–1.353)	1.789 (1.689–1.897)	2.376 (2.019–2.857)
	Women	1.084 (1.010–1.163)	1	1.150 (1.116–1.185)	1.385 (1.346–1.425)	1.865 (1.725–2.019)	2.472 (2.025–3.028)
Diabetes	Men	0.843 (0.717–0.990)	1	1.603 (1.524–1.685)	2.431 (2.321–2.546)	4.784 (4.371–5.237)	5.592 (3.845–6.624)
	Women	0.744 (0.607–0.911)	1	1.624 (1.528–1.725)	2.567 (2.43–2.712)	4.372 (3.991–4.788)	4.808 (3.624–6.378)
CCVDs	Men	0.974 (0.894–1.060)	1	1.109 (1.073–1.146)	1.200 (1.162–1.241)	1.351 (1.226–1.489)	1.002 (0.551–1.821)
	Women	0.907 (0.827–0.995)	1	1.099 (1.060–1.140)	1.169 (1.128–1.211)	1.264 (1.158–1.380)	1.462 (1.101–2.215)
All-cause mortality	Men	1.710 (1.596–1.832)	1	0.780 (0.744–0.817)	0.730 (0.694–0.767)	0.948 (0.807–1.114)	1.029 (0.613–1.863)
	Women	1.698 (1.524–1.893)	1	0.801 (0.744–0.862)	0.830 (0.773–0.891)	1.123 (0.949–1.331)	1.368 (0.869–2.306)

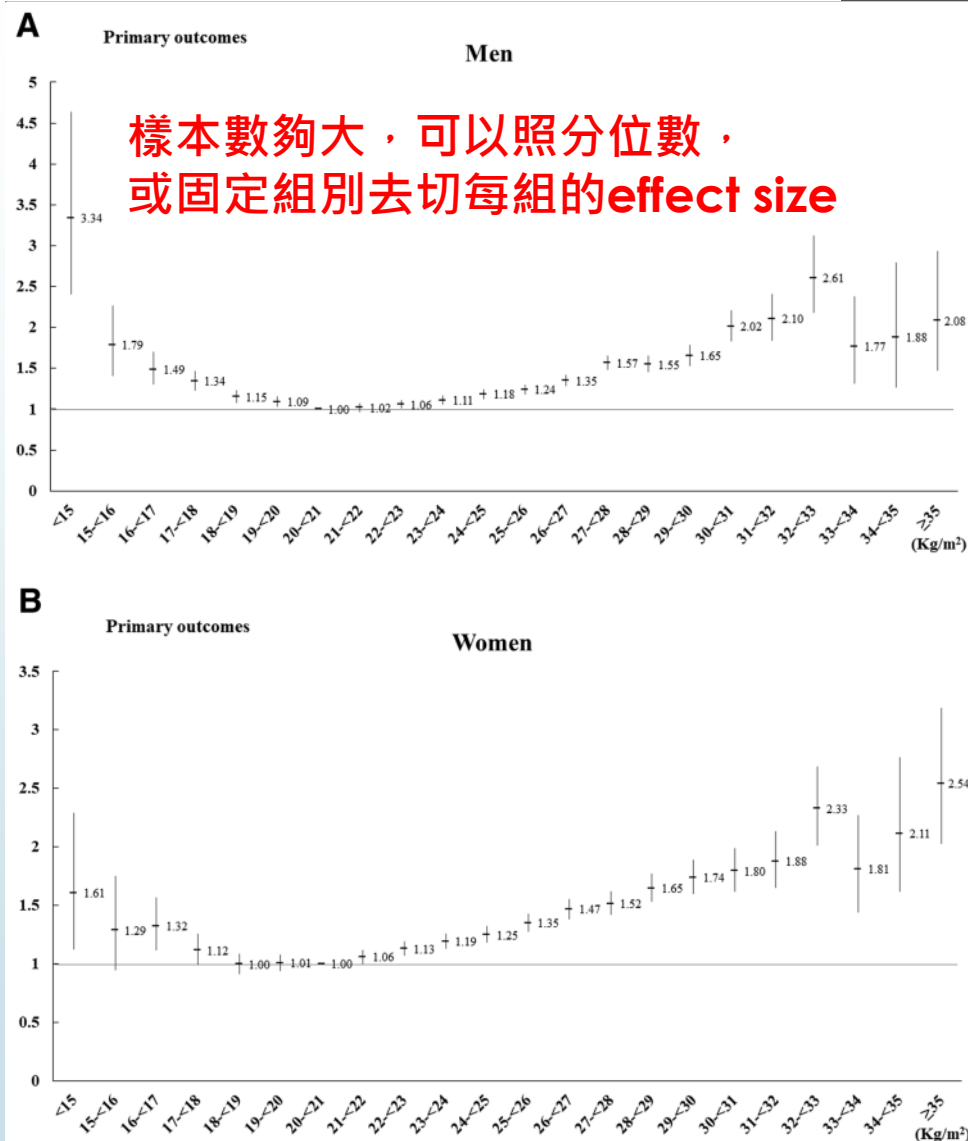
Adjusted for age, systolic blood pressure, glucose, total cholesterol, tobacco smoking, alcohol intake, physical activity, and economic status

Abbreviation: CCVDs Cardio-cerebrovascular diseases

^a Primary outcomes include diabetes mellitus, cardio-cerebrovascular diseases, or all-cause mortality

Table 1 Baseline characteristics according to sex

	Men	Women
Number (N)	167,500	143,916
Age, years	50.4 ± 8.6	51.4 ± 9.0
Body mass index, kg/m ²	23.8 ± 2.8	23.5 ± 2.9
Systolic blood pressure, mmHg	125.3 ± 15.6	120.5 ± 16.3
Glucose, mg/dL	91.3 ± 12.9	89.5 ± 11.9
Total cholesterol, mg/dL	197.0 ± 36.1	198.1 ± 36.5
Body mass index category, N (%)		
Underweight (< 18.5 kg/m ²)	4137 (2.5)	3861 (2.7)
Normal (18.5–< 23 kg/m ²)	61,916 (37.0)	61,708 (42.9)
Overweight (23–< 25 kg/m ²)	47,585 (28.4)	37,763 (26.2)
Grade 1 Obesity (25–< 30 kg/m ²)	51,220 (30.6)	37,137 (25.8)
Grade 2 obesity (30–< 35 kg/m ²)	2531 (1.5)	3220 (2.2)
Grade 3 obesity (≥ 35 kg/m ²)	111 (0.07)	227 (0.16)
Smoking status, N (%)		
Never smokers	66,415 (39.7)	138,671 (96.4)
Ever smokers	101,085 (60.3)	5245 (3.6)
Drinking status, N (%)		
Rare	55,825 (33.3)	116,095 (80.7)
Sometimes	79,507 (47.5)	25,058 (17.4)
Often	32,168 (19.2)	2763 (1.9)
Physical activity, N (%)		
Rare	83,575 (49.9)	95,767 (66.5)
Sometimes	69,568 (41.5)	36,091 (25.1)
Regular	14,357 (8.6)	12,058 (8.4)
Economic status, N (%)		
Low	27,488 (16.4)	39,624 (27.5)
Middle	55,054 (32.9)	46,155 (32.1)
High	84,958 (50.7)	58,137 (40.4)

**Fig. 2** Cox-proportional hazards regression models for the primary outcomes (diabetes mellitus, cardio-cerebrovascular diseases, or all-cause mortality) according to body mass index. Adjusted for age, systolic blood pressure, glucose, total cholesterol, tobacco smoking, alcohol intake, physical activity, and economic status

RESEARCH

Open Access



Body mass index, genetic susceptibility, and Alzheimer's disease: a longitudinal study based on 475,813 participants from the UK Biobank

Shiqi Yuan^{1†}, Wentao Wu^{2†}, Wen Ma^{2†}, Xiaxuan Huang¹, Tao Huang³, Min Peng¹, Anding Xu^{1*} and Jun Lyu^{3,4*} 

使用RCS方法

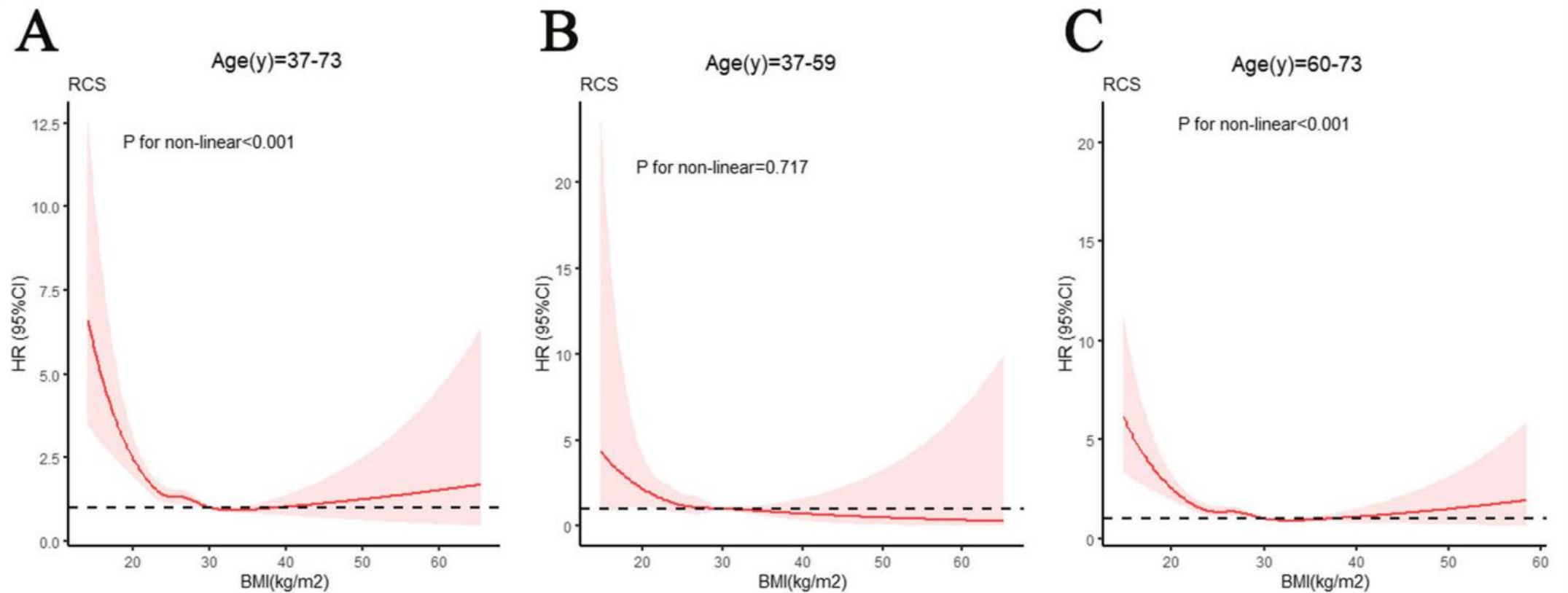


Fig. 2 The restricted cubic splines (RCS) for analysis of the relationship between BMI and incidence of AD. **A:** 37–73 (years), **B:** 37–59 (years), **C:** 60–73 (years). Adjusted for age, TDI, sex, smoking, ethnicity, education level, alcohol use, hypertension, stroke, myocardial infarction, and diabetes

當代主流的非線性關係

12

➤ Restricted cubic spline (RCS, R有完整套組)

- **Cubic spline** 是指在不同的節點之間，函數可以是三次方的關係
- 使用 **RCS** 時，使用「 $k - 1$ 」個自由度即可（不包含截距項）， k 為節點的數量
- 每個 **RCS** 只需要估計線性效果 X 以及 $k-2$ 個分段 **piecewise cubic**變項
- 需要決定兩件事情：第一是節點的數量，第二是節點的位置
- 節點數量的決定會直接跟樣本數有關係，通常樣本數越小（例如 <100 ）則建議採取三點節點就好，若樣本數很大（例如上千、上萬）或許最多可以考慮到五個節點
- 大多數的資料使用四個節點應該是合適
- 節點的位置，若有事先具有臨床意義的切點則直接採用，例如像是血壓的控制會就有國際學會制訂的標準（ <90 , $91-120$, $121-140$, $141-160$, >160 mmHg），然而大多數的臨床研究都是探索性的，因此極少一開始就已經知道節點要切在哪裡
- 「節點的位置」對 **RCS** 結果的影響比較小，反而是「節點的數量」的選擇比較重要

Number of knots K	Knot locations expressed in quantiles of the x variable							
3	0.1	0.5	0.9					
4	0.05	0.35	0.65	0.95				
5	0.05	0.275	0.5	0.725	0.95			
6	0.05	0.23	0.41	0.59	0.77	0.95		
7	0.025	0.1833	0.3417	0.5	0.6583	0.8167	0.975	

Table 2. Location of knots. From Harrell (2001), Regression Modeling Strategies.

使用 R 軟體的 **rms** 套件的預設節點所在分位數位置

R Code for RCS

#安裝讀檔軟體

```
install.packages("survival")
```

```
install.packages("haven")
```

```
install.packages("readxl")
```

```
library(survival)
```

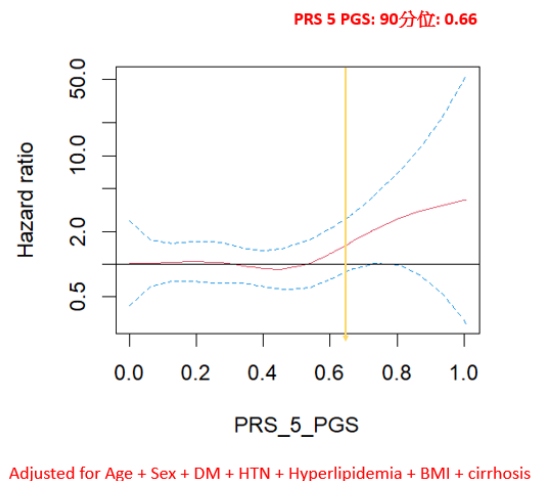
```
library(haven)
```

```
library(readxl)
```

Cox proportional hazards models were used to estimate hazard ratios for cardiovascular and mortality events.

Restricted cubic splines (RCS) under a Cox regression were applied when analyzing time to event data with continuous exposures to "smooth" continuous measurements with a specific outcome.

All analyses were performed using R (version 3.6.3)



Cox regression

Cox regression

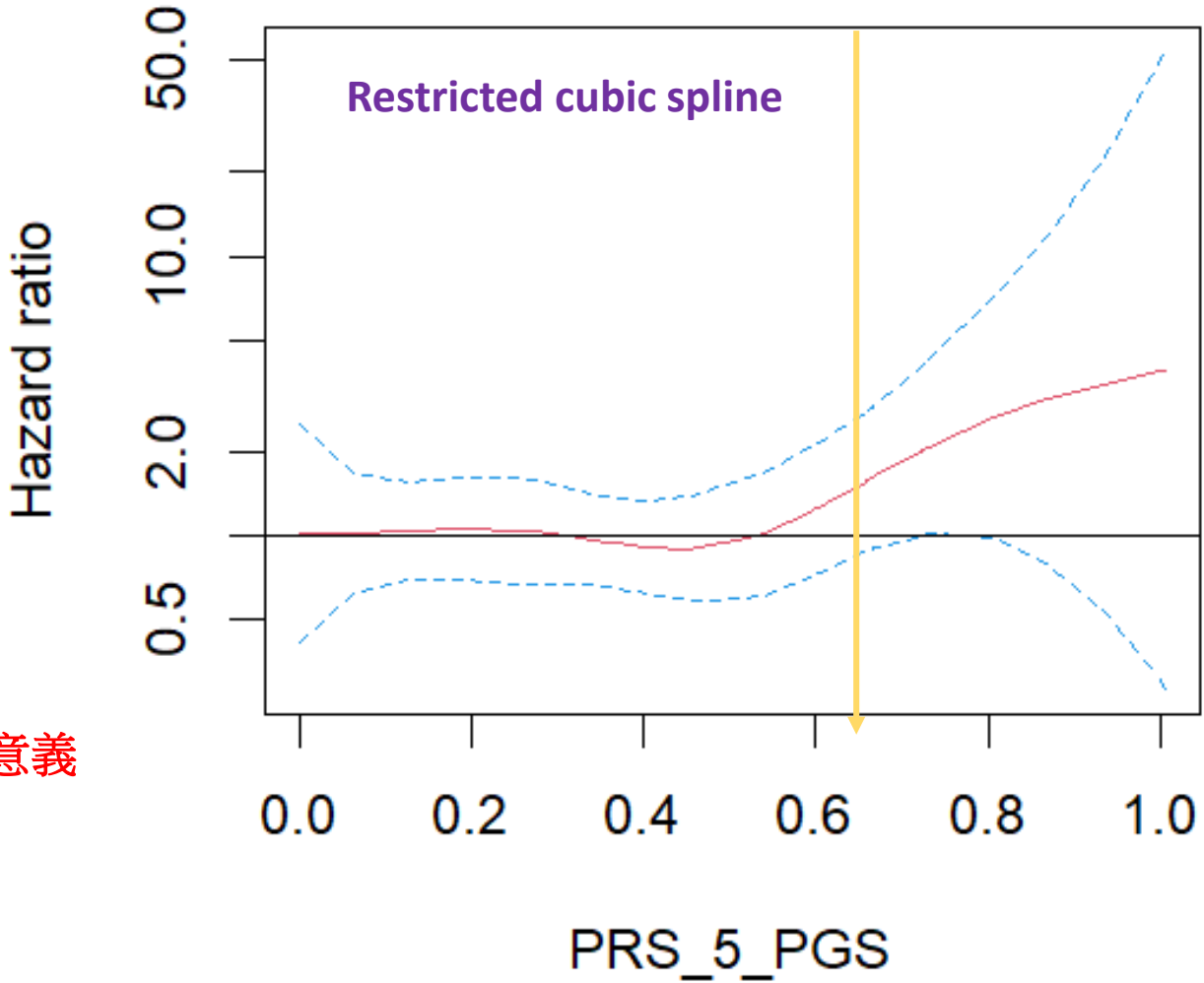
All patients	Model 0: crude effect		Model 1 : Adjusted for Age + Sex+ BMI + cirrhosis		Model 2 : Adjusted for Age + Sex + DM + HTN + Hyperlipidemia + BMI + cirrhosis	
	HR 95%CI	P-value	HR 95%CI	P-value	HR 95%CI	P-value
連續變數 PRS_5_PGS	2.04 (0.49-8.41)	0.32	1.94 (0.46-8.23)	0.37	2.15 (0.51-9.17)	0.30
分兩組 PRS_5_PGS_90 p = 1 0: reference)	3.13 (1.43-6.87)	0.004	3.02 (1.36-6.71)	0.007	2..88 (1.27-6.55)	0.02

參數		自由度	參數 估計值	標準 誤差	卡方	HR	95% CI		Pr > ChiSq	標籤
PRS_5_PGS_90p	1	1	1.05847	0.41908	6.3792	2.882	1.268	6.552	0.0115	PRS_5_PGS_90p 1
Age		1	0.02728	0.01642	2.7598	1.028	0.995	1.061	0.0967	Age
Sex	1	1	0.31682	0.34136	0.8614	1.373	0.703	2.680	0.3534	Sex 1
DM	1	1	0.13058	0.36384	0.1288	1.139	0.558	2.325	0.7197	DM 1
HTN	1	1	1.08608	0.50654	4.5972	2.963	1.098	7.996	0.0320	HTN 1
Hyperlipide mia	1	1	-0.85484	0.40973	4.3528	0.425	0.191	0.950	0.0369	Hyperlipide mia 1
BMI		1	0.08683	0.03483	6.2147	1.091	1.019	1.168	0.0127	BMI
cirrhosis	1	1	1.58076	0.36080	19.1950	4.859	2.396	9.854	<.0001	cirrhosis 1

根據RCS，要切到很高，才會有HR >1的顯著意義

PRS 5 PGS: 90分位: 0.66

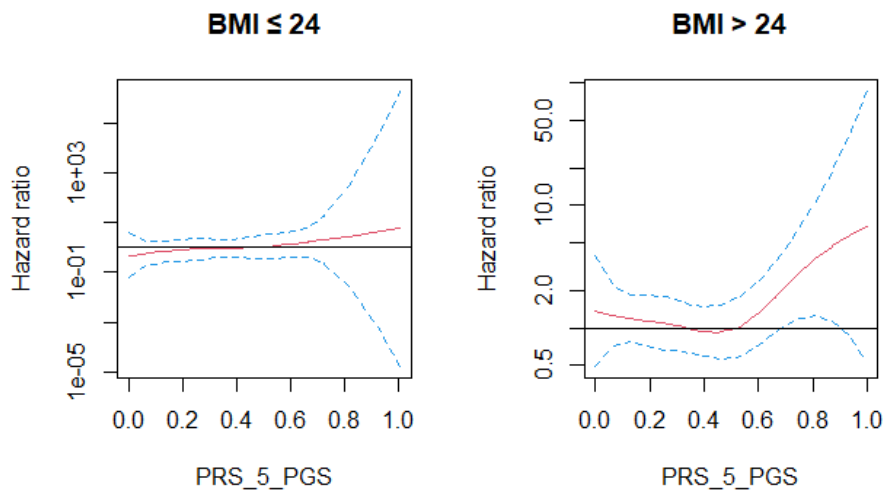
統計量			
PRS_5_PGS	N	有效	439
		遺漏	0
		平均數	.334380
		中位數	.330400
		標準差	.2232560
	百分位數	10	.064700
		20	.127600
		25	.127600
		30	.129400
		40	.265700
		50	.330400
		60	.393300
		70	.401400
		75	.458000
		80	.531400
		90	.659000



根據RCS，要切到很高，才會有HR >1的顯著意義

```
> # 描述性統計
>
> summary_stats <- list(
+   P10 = quantile(data_ori$PRS_5_PGS, probs = 0.1),
+   Q1 = quantile(data_ori$PRS_5_PGS, probs = 0.25),
+   Median = median(data_ori$PRS_5_PGS),
+   Q3 = quantile(data_ori$PRS_5_PGS, probs = 0.75),
+   Mean = mean(data_ori$PRS_5_PGS),
+   SD = sd(data_ori$PRS_5_PGS)
+ )
> print(summary_stats)
```

Adjusted for Age + Sex + DM + HTN + Hyperlipidemia + BMI + cirrhosis



Restricted cubic spline
依照BMI >= 24與否分組

Adjusted for Age + Sex + DM + HTN + Hyperlipidemia + BMI + cirrhosis

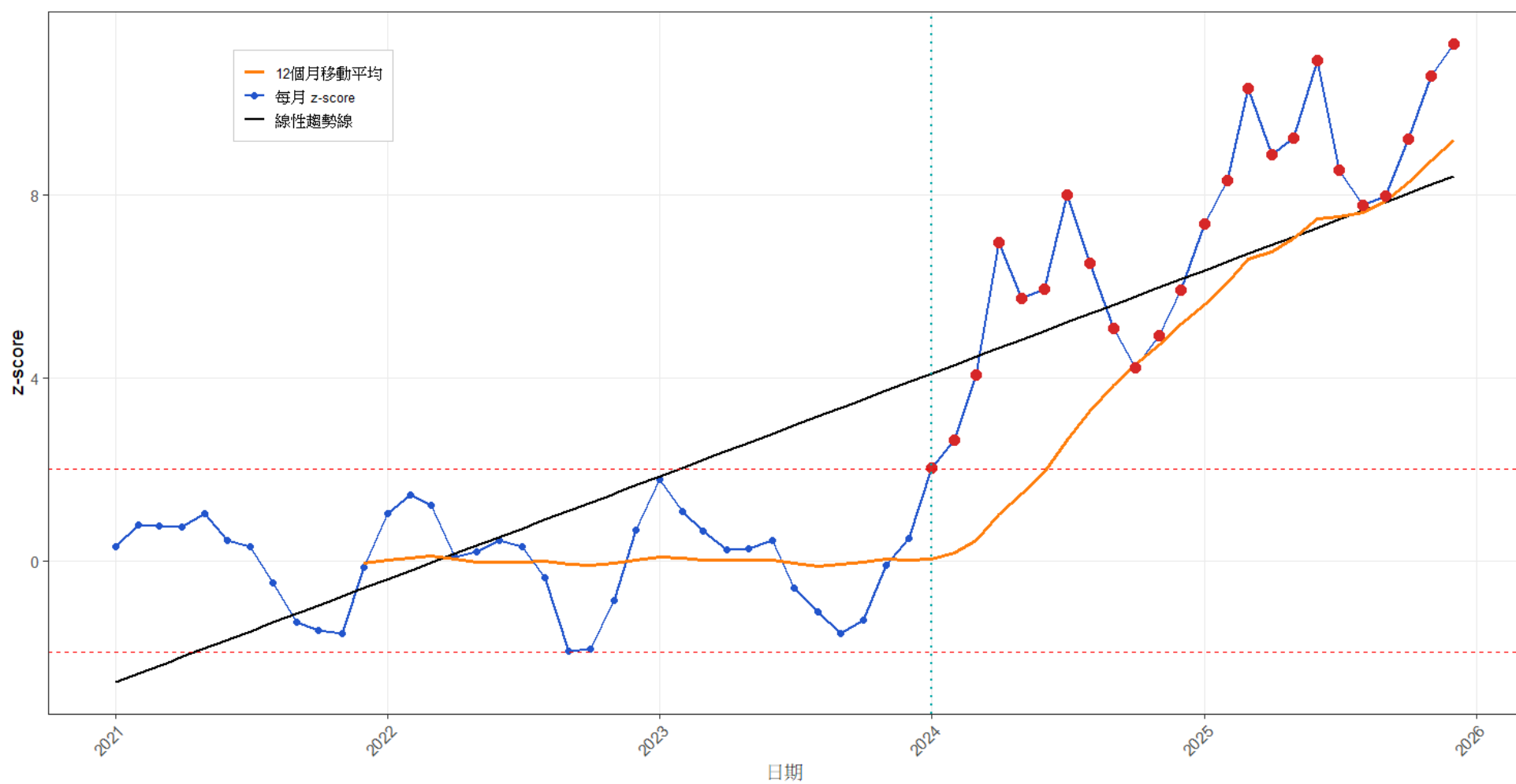
		Model 0: crude effect		Model 1 : Adjusted for Age + Sex + DM + HTN + Hyperlipidemia + cirrhosis	
Cox regression		HR 95%CI	P-value	HR 95%CI	P-value
BMI <24	連續變數PRS_5_PGS	2.29 (0.15-34.8)	0.55	5.44 (0.18-163.7)	0.33
BMI >= 24		1.89 (0.36-9.76)	0.45	2.24 (0.41-12.4)	0.35
BMI <24	分兩組PRS_5_PGS_90p = 1 0: reference)	1.17 (0.15-9.24)	0.88	2.78 (0.29-26.5)	0.37
BMI >= 24		4.15 (1.75-9.86)	0.001	4.97 (1.85-13.4)	0.005

時間序列資料

時間序列資料：像每個月追蹤同一指標，例如每月急診量。

時間序列的重點是「順序」：前後月份會互相影響。

因此時間序列不能只看平均，也要看趨勢、季節、變化點。



Z-score 異常點偵測



用途：快速找出偏離基準期的月份。

$z = (\text{觀測值} - \text{基準期平均}) / \text{基準期標準差}$ 。

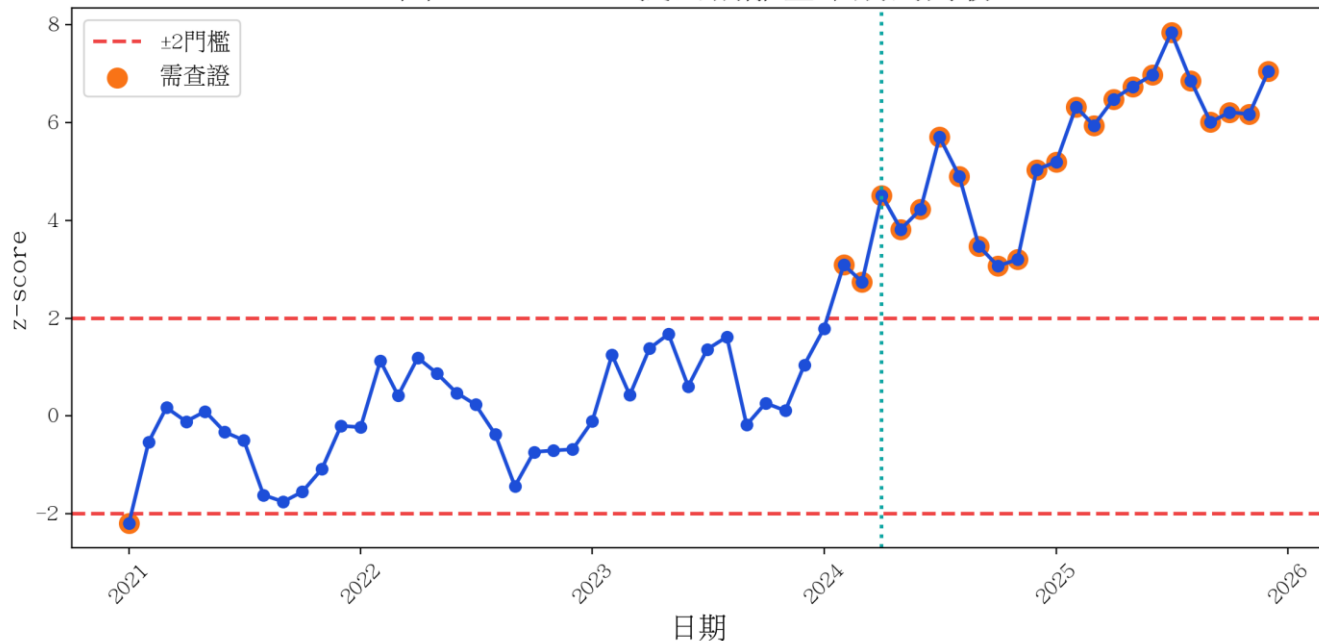
$|z| \geq 2$ 的月份可標示為「需查證」。

方法：以基準期平均與標準差計算 z-score

Z-score 找出需查證月份

用簡單標準化判斷「偏離很多」

圖4: z-score 找出偏離基準期的月份



方法：用基準期的平均與標準差計算 z-score。

判斷原則： $|z\text{-score}| \geq 2$ ，代表偏離基準期較明顯。

圖表解釋：橘色點是需查證月份。

提醒：顯著偏離不等於因果，只是提醒要查原因。

R code重點： $z = (\text{value} - \text{baseline_mean}) / \text{baseline_sd}$



Thank you for listening