521 M7410 –Adjustment and Analysis of Spatial Information

Fall Semester 2021

Homework No. 0

handed out Thursday, September 23, 2021
due Thursday, September 30, 2021, 14:20 **Name:** 李睿莆

Data Adjustment - why/what/how to adjust?

You are given a data set showing the fatality rate of a disease (D) at a certain age (A).

1. Please describe (in your own words) the relation between these two factors.

2. Could you use a mathematical model to represent this relation? If positive, please give an explicit form of the model. Besides, give the parameter values of the model.

3. How confident are you on the above mentioned model? How can you evaluate it?

Age (A)	20	25	30	35	40	45	50	55	60	65	70	75	80
Disease Rate (D)	2%	6%	7%	13%	21%	25%	30%	37%	35%	31%	26%	36%	45%

Your (individual) final report should contain (use A4 papers):

- this page as the cover sheet
- source code(s) and outputs; do not forget to add your name and lots of comment cards to the source listing (% .....)
- input and output files from program [input/output values used and calculated], if any
- plots, including captions on axes, title, your name, LB#/HM#, course title, date (if any)
- derivation and description of formulas used, accompanied by figures where applicable
- evidence of computational accuracy
- discussion of results

1. Please describe (in your own words) the relation between these two factors.

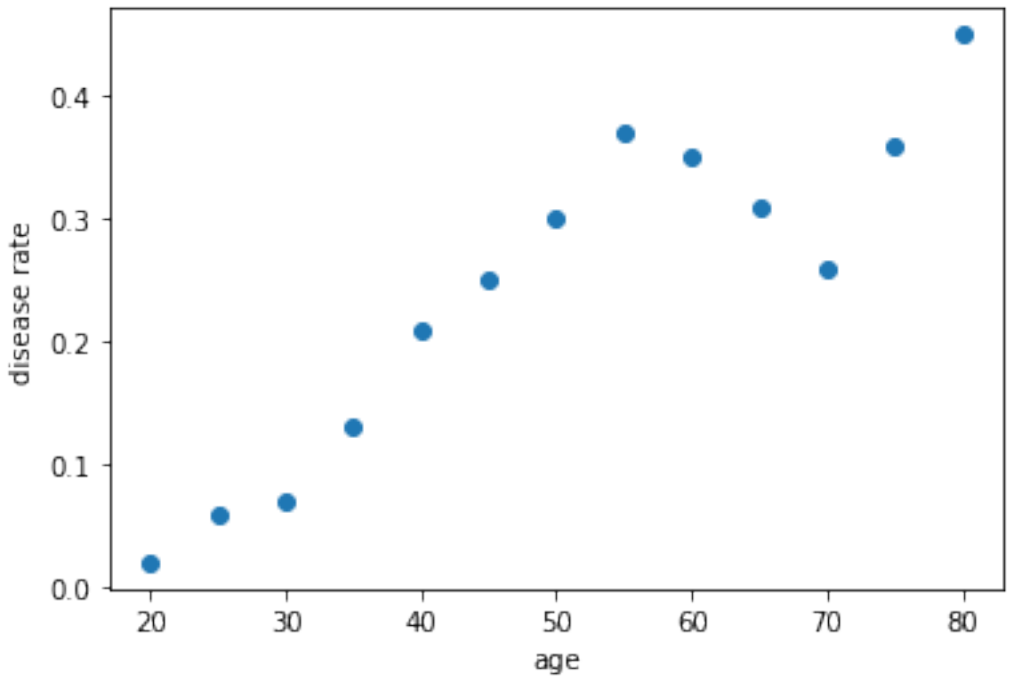

圖 1 年齡與發病率散布圖

為了找出年齡與發病率之間的關聯性，首先觀察這份資料是否存在簡單的線性關係，將年齡作為 x，發病率作為 y 繪製出散布圖(如圖 1)，可以得知 20 歲到 55 歲的區間和 70 歲到 80 歲的區間中，年齡與發病率呈現正相關；而 55 歲到 70 歲的區間中年齡與發病率呈現負相關。

2. Could you use a mathematical model to represent this relation? If positive, please give an explicit form of the model. Besides, give the parameter values of the model.

依據資料分布而言，兩個變量呈現非線性函數的關係，若要以一個簡單回歸模型配適這份資料效果可能不盡理想，因此以下我們將採用多項式回歸模型擬合資料。圖 2～圖 7 為二次函數到七次函數模型擬合資料的結果，包含 95% 信賴水準下，模型估計母體平均反應的信賴區間、模型對新觀測值的預測區間，以及模型指標如判定係數 R^2 與調整後判定係數R_a^2。

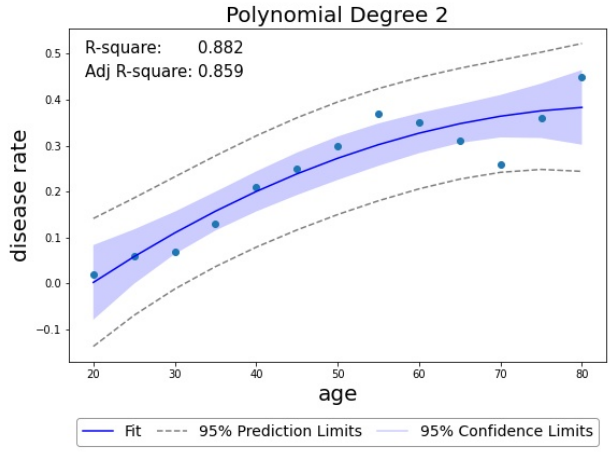

圖 2 二次函數模型

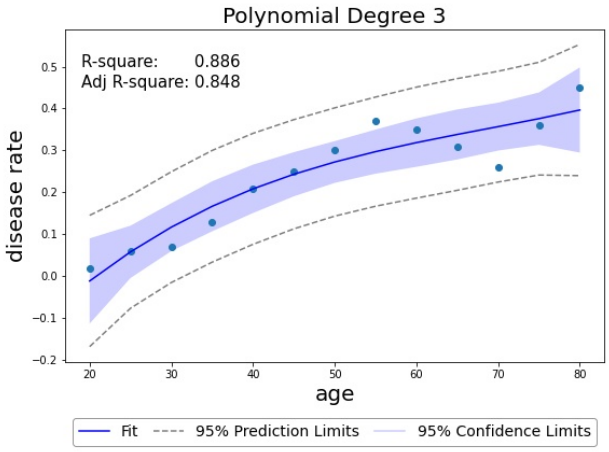

圖 3 三次函數模型

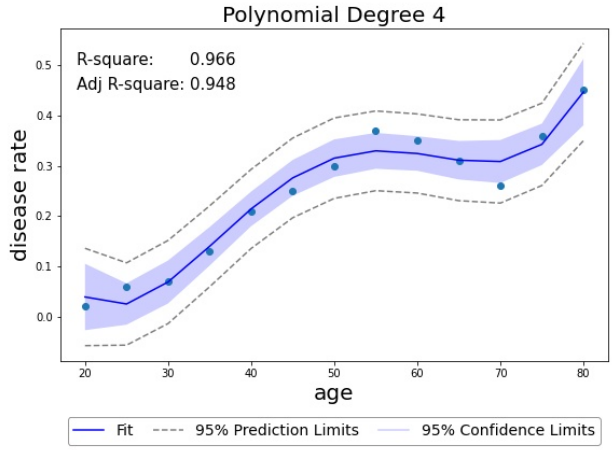

圖 4 四次函數模型

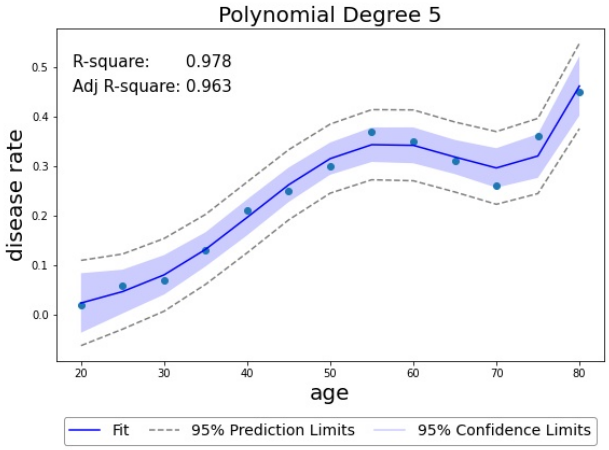

圖 5 五次函數模型

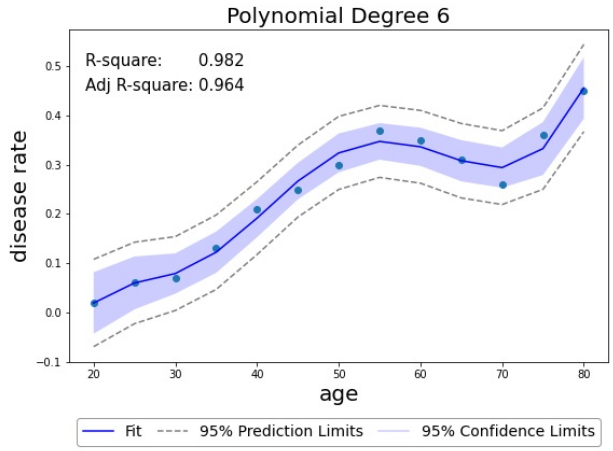

圖 6 六次函數模型

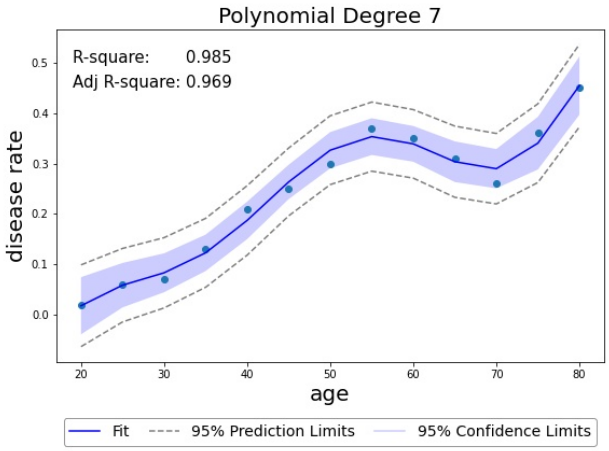

圖 7 七次函數模型

由上述列出的六種模型可見，多項式迴歸模型的函數冪次越高，R^2 值越大，而這並不代表該模型就是一個好的模型，僅代表模型越擬合這份資料。然而模型越擬合資料將會產生過擬合(over-fitting)的問題，代表模型僅能在目前的資料集上表現良好，在測試集上的表現準確度較低。理論上若想要嚴謹的建立良好並且能泛化的預測模型，可以透過模型預測測試集的結果來評估模型表現的好壞，如 accuracy、F-score 等；抑或是在訓練模型前將資料集切分成訓練集與驗證集，再做 K-fold 交叉驗證，以指標分數最高的模型做為最終採用模型，較能避免 over-fitting 的問題。

然而本次研究的資料量僅有 13 筆，若取 30%資料量做為驗證(測試)集，除了模型訓練資料僅剩 9 筆以外，還需考慮因資料量過少使得在抽樣時產生訓練集、驗證集資料分布不均勻的問題。因此本次研究便不將原始資料進行 K-fold 驗證，僅建立能擬合資料且解釋性較高的模型，最終採用四次多項式迴歸模型進行後續分析，模型選擇原因詳述如後，模型公式及指標資訊如圖 8。

$$Y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4$$

Y	x	b_0	b_1	b_2	b_3	b_4
$disease\ rate$	age	1.5333	-0.1680	0.0063	$-9.276 * 10^{-5}$	$4.718 * 10^{-7}$

OLS Regression Results

Dep. Variable:	y	R-squared:	0.966
Model:	OLS	Adj. R-squared:	0.948
Method:	Least Squares	F-statistic:	56.20
Date:	Wed, 29 Sep 2021	Prob (F-statistic):	6.78e-06
Time:	23:13:00	Log-Likelihood:	30.002
No. Observations:	13	AIC:	-50.00
Df Residuals:	8	BIC:	-47.18
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.5333	0.486	3.158	0.013	0.414	2.653
x1	-0.1680	0.047	-3.591	0.007	-0.276	-0.060
x2	0.0063	0.002	4.030	0.004	0.003	0.010
x3	-9.276e-05	2.2e-05	-4.212	0.003	-0.000	-4.2e-05
x4	4.718e-07	1.1e-07	4.298	0.003	2.19e-07	7.25e-07

圖 8 四次多項式迴歸模型資訊

3. How confident are you on the above mentioned model? How can you evaluate it?

(3.1) R^2與R_a^2

該模型之判定係數 R^2 為 96.6%，調整後判定係數R_a^2為 94.8%。R^2 在統計上的意義為，發病率有 96.6%的變異可以透過模型解釋，另一種解釋則為引進年齡變數至模型後，估計誤差平方和減少了 96.6%。然而這種推論較適合使用單一回歸模型時進行評估，本次採用的多項式迴歸模型，自變數不只一個，當模型採用的自變數越多，R^2 就會越大，將會產生高估的現象。R_a^2在考量自變數的數量進行調整後，就能避免 R^2 過度膨脹。因此由 94.8%的R_a^2可見，在經過修正後數值是略低於判定係數 R^2，判定係數及調整後判定係數公式如下:

$$R^2 = \frac{\text{SSR}}{\text{SSTO}} = 1 - \frac{\text{SSE}}{\text{SSTO}}$$

$$R_a^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

其中 SSR 用來衡量觀測值的變異中，可藉由模型加以解釋的部分，SSTO 用來衡量觀測值的總變異，SSE 用來衡量針對模型調整後，觀測值的變異情形，也就是估計誤差平方和。n 為樣本數、k 為自變數數量。理論上R_a^2越大僅代表模型越擬合現有資料，並不絕對代表這個簡單回歸模型就是一個很好的預測模型，因為R_a^2並無提供任何有關預測 Y_{new} 好壞的訊息。此外，R_a^2為 94.8%並不代表模型即可解釋發病率 94.8%的變異，因為在經過自由度的修正後，R_a^2 公式便難以推論，使得在統計上難以說明其實質意義。

在本次模型選擇時，觀察二次函數到七次函數模型的R^2與R_a^2(如表 1)，發現二次函數到四次函數模型中增加自變量的個數有助於R_a^2顯著提升，當多項式函數模型冪次為五以上時，添加新的自變量後R^2與R_a^2皆無顯著增加，表示加入新的變量並未有效地獲得更多應變數的變異。因此本次模型選擇四次多項式迴歸模型進行資料擬合。

表 1 六種模型的R^2與R_a^2

	二次函數	三次函數	四次函數	五次函數	六次函數	七次函數
R^2	0.882	0.886	0.966	0.978	0.982	0.985
R_a^2	0.859	0.848	0.948	0.963	0.964	0.969

(3.2)信賴區間與預測區間

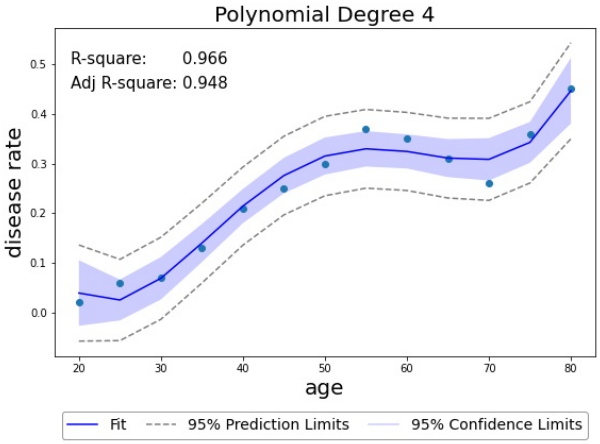

圖 9 95%信賴區間與預測區間

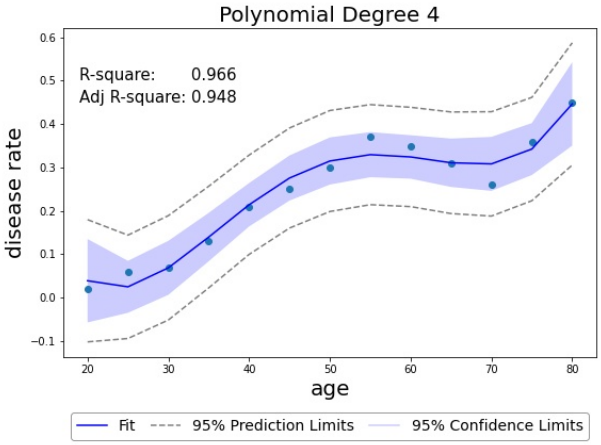

圖 10 99%信賴區間與預測區間

由圖 9 可見，在 95%信賴水準下，藍色區塊為模型估計 X(年齡)平均反應的信賴區間，黑色虛線為模型對觀測值的預測區間。值得一提的是，以估計的角度來看，年齡 20 歲的信賴區間較大，45 歲的信賴區間較小，這是因為當年齡愈靠近 50 歲(平均值)的區域，信賴區間會愈小，代表模型有較高的信心可以準確估計該年齡的平均確診率，預測區間也具有相同的情況。而在 55 歲與 70 歲時，本模型在 95%信賴水準下較無法準確估計其確診率，若是在 99%信賴水準下本模型即可準確的估計 20 歲至 80 歲的平均發病率。

(3.3)殘差圖與參數推論

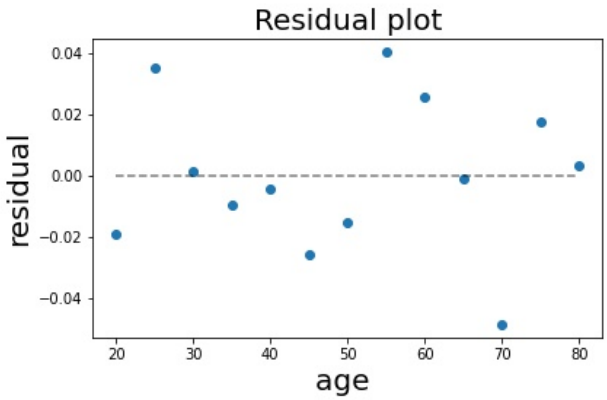

圖 11 殘差對 age 之散布圖

由圖 11 殘差圖可以清楚發現，模型殘差大約以 0 為中心正負 0.04 跳動，因此我們不拒絕殘差變異數為常數的假設，並接受此模型作為統計推論的配適模型。最後關於模型參數的統計推論，由於年齡為 0 時沒有實務上的意義，故推論截距β_0變不具意義。而在推論$\beta_1 \ldots \beta_4$時，由於各項自變數皆有共線性且高度相關，因此推論多項式迴歸係數時較難解釋其實質意義。

Code:

https://nbviewer.jupyter.org/github/106207411/Adjustment-Spatial-Information/blob/master/HW00/HW00.ipynb