THE CHINESE UNIVERSITY OF HONG KONG
Department of Mathematics

Mathematical Modeling Project Team
mathmodel@math.cuhk.edu.hk

Exercise (Linear and Nonlinear Regression)

Last updated: March 11, 2026

> Information is the oil of the 21st century, and analytics is the combustion engine.
>
> – Peter Sondergaard

1 Correlation Analysis
相關性分析

Let's say we want to discuss the relation between residential electricity consumption and annual mean temperature in Hong Kong, we can try to determine the correlation coefficient for it. The data are as follows:
假設我們想探討香港住宅用電量與年平均溫度之間的關係，我們可以嘗試計算它們的相關係數。相關數據如下：

	Electricity consumption (TJ)	**Annual mean temperature (°C)**
2010	39344	23.2
2011	39872	23.0
2012	41189	23.4
2013	39941	23.3
2014	43415	23.5
2015	42368	24.2
2016	43120	23.6
2017	42127	23.9
2018	41965	23.9

Correlation Coefficient
相關係數

The correlation coefficient, denoted as r, measures the strength and direction of the linear relationship between two variables. Ranges from -1 to 1:
相關係數以r 表示，用於衡量兩個變數之間線性關係的強度與方向。其數值範圍介乎-1 至1 之間：

- $r = 1$: Perfect positive correlation.
  $r = 1$：完全正相關
- $r = -1$: Perfect negative correlation.
  $r = -1$：完全負相關
- $r = 0$: No linear correlation.
  $r = 0$：無線性相關

The formula for the correlation coefficient is given by the following:
相關係數的計算公式如下：

$$r = \frac{n\sum(xy) - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where:
其中：

- n is the number of data points,
  n 為數據點的數量
- x and y are the individual sample points,
  x 與y 為各個樣本點
- $\sum(xy)$ is the sum of the product of paired scores,
  $\sum(xy)$ 為配對數據的乘積之總和
- $\sum x$ and $\sum y$ are the sums of the x and y scores respectively,
  $\sum x$ 與$\sum y$ 分別為x 與y 的總和
- $\sum x^2$ and $\sum y^2$ are the sums of the squares of x and y scores respectively.
  $\sum x^2$ 與$\sum y^2$ 分別為x 與y 的平方和

To find the correlation coefficient for the electricity consumption and annual mean temperature data, we first calculate the required sums $\sum x, \sum y, \sum(xy), \sum x^2, \sum y^2$:
為了計算用電量與年平均溫度數據的相關係數，我們首先計算所需的總和值$\sum x$、$\sum y$、$\sum(xy)$、$\sum x^2$、$\sum y^2$：

	x	y	xy	x^2	y^2
2010	39344	23.2			
2011	39872	23.0			
2012	41189	23.4			
2013	39941	23.3			
2014	43415	23.5			
2015	42368	24.2			
2016	43120	23.6			
2017	42127	23.9			
2018	41965	23.9			
Sum					

Let's try to put our data into the formula, $r =$
現在我們將數據代入公式，$r =$

Conclusion for the relationship of the data:
數據關係的結論：

2 Linear Regression 線性迴歸

Linear regression is highly related to the correlation coefficient. If we have $r = 1$ or $r = -1$, then we can draw a line through all data points, since they have perfect correlation.
線性迴歸與相關係數密切相關。若$r = 1$ 或$r = -1$，我們就可以繪畫一條穿過所有數據點的直線，因為它們具有完全相關性。

Also, we can use linear regression to make some predictions about the new data points in the future!
此外，我們可以利用線性迴歸對未來的新數據點進行預測！

Best-fit Line 最佳擬合直線

The average life expectancy at birth in the world is given as follows:
全球出生時平均預期壽命數據如下：

Year	**life expectancy**
1950	46.4
1960	47.8
1970	56.3
1980	60.5
1990	64
2000	66.4
2010	70.1
2020	71.9

Can you try to propose some methods for finding the 'best-fit line' for the data?
你能嘗試提出一些尋找數據「最佳擬合直線」的方法嗎？

Also, using your definition, please try to predict the average life expectancy at birth in the world in years 2030, 2050, and 2077.
另外，請根據你定義的方法，嘗試預測2030 年、2050 年及2077 年全球出生時平均預期壽命。

Linear Regression Model
線性迴歸模型

In mathematics, we define a best-fit line as the line that minimizes the *residual sum of squares:*
在數學中，我們將最佳擬合直線定義為能使殘差平方和最小化的直線：

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

By solving, we have the following result:
通過求解，我們得到以下結果：

$$c = \frac{n\sum(xy) - \sum x \sum y}{n\sum(x^2) - (\sum x)^2}$$

$$m = \bar{y} - c\bar{x}$$

where:
其中：

- n is the number of data points,
  n 為數據點的數量
- $\bar{x}$ is the mean of the independent variable,
  $\bar{x}$ 為自變數的平均值
- $\bar{y}$ is the mean of the dependent variable.
  $\bar{y}$ 為應變數的平均值

Using the life expectancy data, again try to find the linear regression of the data.
(Hint: first determine which is the independent/ dependent variable, then apply the formula)
利用預期壽命數據，再次嘗試找出數據的線性迴歸方程。
（提示：首先確定哪個是自變數／應變數，然後應用公式）

	Year (x)	**life expectancy** (y)	xy	x^2
	1950	46.4		
	1960	47.8		
	1970	56.3		
	1980	60.5		
	1990	64		
	2000	66.4		
	2010	70.1		
	2020	71.9		
Sum				

Therefore, the equation of the straight line is the following:
因此，直線方程如下：

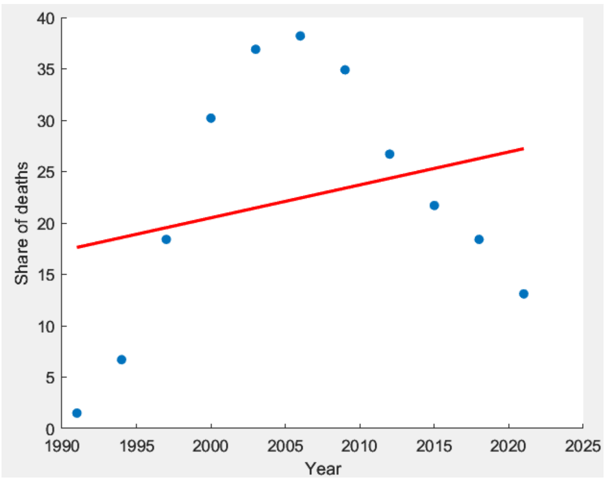
Isn't the calculation tedious? Luckily, we have tools for you. Try to use our Linear Regression R Shiny Tool (https://www.math.cuhk.edu.hk/app/mathmodel/tool.html) to check your calculations!
計算過程是否相當繁複？幸好，我們為大家準備了工具。嘗試使用我們的線性迴歸R Shiny 工具(https://www.math.cuhk.edu.hk/app/mathmodel/tool.html) 來驗證你的計算結果！

3 Non-linear Regression
非線性迴歸

Sometimes, linear regression cannot accurately approximate our data. See the example below:
有時候，線性迴歸無法準確地近似我們的數據。請看以下例子：

Year	Share of all deaths from HIV/AIDS in South Africa (in %)
1991	1.5
1994	6.7
1997	18.4
2000	30.2
2003	36.9
2006	38.2
2009	34.9
2012	26.7
2015	21.7
2018	18.4
2021	13.1

If we proceed to use linear regression, the graph will look like the following:
如果我們使用線性迴歸，圖像將會如下所示：



Of course, this is not the function we want!
顯然，這並非我們想要的函數！

Check out our Non-Linear Regression R Shiny tool (https://www.math.cuhk.edu.hk/app/mathmodel/tool.html) to see more functions that can be used to approximate the data.
歡迎查閱我們的非線性迴歸R Shiny 工具(https://www.math.cuhk.edu.hk/app/mathmodel/tool.html)，了解可以用於近似數據的更多函數。

More examples
更多例子

We can put many different real-life situations into our analysis
我們可以將許多不同的現實情境納入分析

In the following, x = weekly study hours by different students, y = scores the students got in an exam:
以下數據中，x = 不同學生的每週温習時數，y = 學生在考試中獲得的分數：

	x	y	xy	x^2	y^2
	4.8	54			
	2.5	21			
	5.1	47			
	3.2	27			
	8.5	75			
	3.5	30			
	1.5	20			
	9.2	88			
	5.5	60			
	8.3	81			
	2.7	25			
Sum					

Therefore, the equation of the linear regression line is:
因此，線性迴歸線的方程為：

Also, the correlation coefficient r is:
同時，相關係數r 為：

Extra question 1: Do you think there is a good correlation between studying hours and examination scores? Why/ why not?
額外問題1：你認為温習時數與考試成績之間是否存在良好的相關性？為甚麼？

Extra question 2: Do you think there are some more functions suitable for the analysis of the data? (i.e. non-linear ones?)
額外問題2：你認為是否有其他函數更適合分析這些數據？（例如非線性函數？）

The following are some even more examples that you can try out. Look for information online, and start your analysis. (Also, you can try to determine linear or non-linear model is more suitable)
以下是一些大家可以嘗試的更多例子。請在網上尋找相關資訊，並開始你的分析。（同時，你可以嘗試判斷線性或非線性模型哪個更為合適）

1. House Price v.s. Age of the House
   樓價與樓齡的關係

2. Rate of a chemical reaction v.s. Concentration of reactants.
   化學反應速率與反應物濃度的關係

3. Employee Salaries v.s. Years of experience
   員工薪資與年資的關係

4. Heart Rate v.s. Duration of exercise
   心率與運動時長的關係

4 Verification
模型驗證

After knowing how to approximate data using different functions, you will need to know which function is the best. Therefore, verification of models is essential. The following are some parameters that you can look into:
在了解如何使用不同函數近似數據後，我們需要判斷哪個函數最為合適。因此，模型驗證至關重要。以下是一些可供參考的參數：

Correlation Coefficient
相關係數

As we mentioned before, a correlation coefficient is a constant from -1 to 1, the closer it is to $1/-1$, the higher the approximating power our linear regression has.
如前所述，相關係數是介乎-1 與1 之間的常數，其數值越接近1 或-1，線性迴歸的近似能力越强。

Residuals
殘差

Residual is the difference between predicted values of y (dependent variable) and observed values of y.
殘差是指y（應變數）的預測值與實際觀測值之間的差異。
It is very intuitive to understand: the smaller the residuals, the better the model is. (Or, is it?)
這很容易理解：殘差越小，模型越佳。（或者，真的是這樣嗎？）

The Sum of Squares/ Variance/ Standard Deviation
平方和／方差／標準差

The sum of the squares measures the deviation of a set of data from the mean. The more diverse the data is, the harder it is to model it well.
Remark: the formula for the sum of squares is
平方和用於衡量一組數據與其平均值的偏離程度。數據越分散，建立良好模型的難度就越大。註：平方和的計算公式為：

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

which is very similar to the variance/ standard deviation we studied in secondary school.
這與我們在中學學習的方差／標準差公式非常相似。

5 More Exercises
更多練習

1. Modelling Advertising Expenditures
廣告支出建模

Go to https://www.kaggle.com/datasets/brsahan/advertising-spend-vs-sales/data to obtain the data about advertising expenditures across various channels (TV, radio, and newspaper) and sales performance. Try to answer the following:
前往https://www.kaggle.com/datasets/brsahan/advertising-spend-vs-sales/data 獲取有關不同媒體渠道（電視、電台、報紙）的廣告支出與銷售表現的數據。嘗試回答以下問題：

(a) What are the linear regression lines and the corresponding correlation coefficients?
線性迴歸線及其對應的相關係數分別為何？

(b) Which way of advertisement shows a stronger relationship with sales?
哪一種廣告方式與銷售額的關係最為密切？

(c) If you were the advertisement team, how would you adjust the advertisement expenditure?
如果你是廣告團隊成員，你會如何調整廣告支出？

(d) How can we further improve the model?
我們可以如何進一步改進模型？

2. Modelling Cybercrime
網絡犯罪建模

Go to https://www.kaggle.com/datasets/huzpsb/cybersecurity-incidents-dataset/data to obtain the data on cybersecurity incidents. Consider cases in Hong Kong (HK), America (US), China (CN), and Japan (JP), and try to answer the following:
前往https://www.kaggle.com/datasets/huzpsb/cybersecurity-incidents-dataset/data 獲取網絡安全事件的數據。考慮香港（HK）、美國（US）、中國（CN）及日本（JP）的情況，嘗試回答以下問題：

(a) What are the linear regression lines for these places respectively?
這些地方的線性迴歸線分別為何？

(b) Which places do you think are in greater danger of cybersecurity?
你認為哪些地方的網絡安全風險較高？

(c) Can you think of any factors affecting the loss cost by cybercrimes? How are they related?
你能想到哪些影響網絡犯罪損失成本的因素？它們之間有何關係？

(d) Try to further create models on loss cost by cybercrime against the factor that you proposed.
嘗試根據你提出的因素，建立網絡犯罪損失成本的模型。

3. Modelling House Prices
樓價建模

Go to https://www.kaggle.com/datasets/harlfoxem/housesalesprediction and obtain the information about houses sold in King County, Washington, USA. Answer the following:
前往https://www.kaggle.com/datasets/harlfoxem/housesalesprediction 獲取美國華盛頓州金縣出售房屋的資訊。回答以下問題：

(a) Plot house prices against a single independent variable, such as the size of the house (square feet).
繪畫樓價與單一自變數（例如房屋面積（平方英尺））的關係圖。

(b) Fit a linear regression model to predict house price based on size.
建立一個基於房屋面積預測樓價的線性迴歸模型。

(c) Calculate the regression line equation and interpret the slope and intercept.
計算迴歸線方程，並解釋斜率和截距的意義。

(d) Evaluate the model's performance using metrics like residual.
使用殘差等指標評估模型的表現。

(e) Discuss whether a linear regression is appropriate or whether there exists another model that might fit better.
討論線性迴歸是否合適，以及是否存在其他更合適的模型。