THE CHINESE UNIVERSITY OF HONG KONG
Department of Mathematics

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

Exercise (Linear and Nonlinear Regression)

Last updated: 17/01/2025

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

1 Correlation Analysis

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:

	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

The correlation coefficient, denoted as r, measures the strength and direction of the linear relationship between two variables. Ranges from -1 to 1:

- $r = 1$: Perfect positive correlation.
- $r = -1$: Perfect negative correlation.
- $r = 0$: No linear correlation.

The formula for the correlation coefficient is given by the following:

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

where:

- n is the number of data points,
- x and y are the individual sample points,
- $\sum(xy)$ is the sum of the product of paired scores,
- $\sum x$ and $\sum y$ are the sums of the x and y scores respectively,
- $\sum x^2$ and $\sum y^2$ are the sums of the squares of x and y scores respectively.

To find the correlation coefficient for the electricity consumption and annual mean temperature data, we first calculate the required sums $\sum x, \sum y, \sum(xy), \sum x^2, \sum y^2$:

	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, $r =$

Conclusion for the relationship of the data:

2 Linear Regression

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. Also, we can use linear regression to make some predictions about the new data points in the future!

Best-fit Line

The average life expectancy at birth in the world is given as follows:

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? Also, using your definition, please try to predict the average life expectancy at birth in the world in years 2030, 2050, and 2077.

Linear Regression Model

In mathematics, we define a best-fit line as the line that minimizes the *residual sum of squares:*

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

By solving, we have the following result:

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

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

where:

- n is the number of data points,
- $\bar{x}$ is the mean of the independent variable,
- $\bar{y}$ is the mean of the dependent variable.

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)

	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:

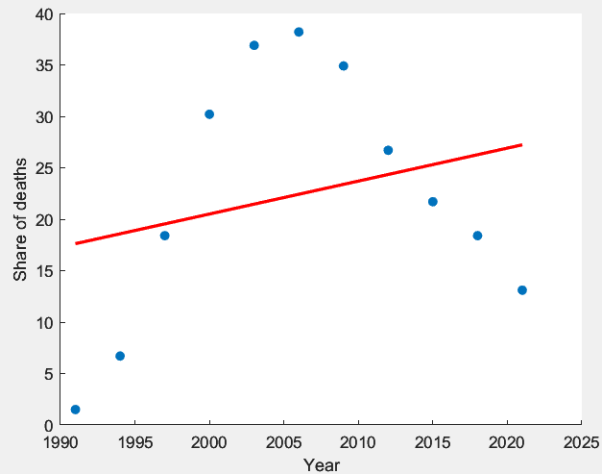
Isn't the calculation tedious? Luckily, we have tools for you. Try to use Linear Regression with R Shiny: `https://www.math.cuhk.edu.hk/app/mathmodel/tool.html` to check your calculations!

3 Non-linear Regression

Sometimes, linear regression cannot accurately approximate our data. See the example below:

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:



Of course this is not the function we want!
Go check the Non-Linear Regression R Shiny tool at `https://www.math.cuhk.edu.hk/app/mathmodel/tool.html` to see more functions that can be used to approximate the data.

More examples

We can put many different real-life situations into our analysis:
In the following, x = weekly study hours by different students, y = scores the students got in an exam:

	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:

Also, the correlation coefficient r is:

Extra question: Do you think there is a good correlation between studying hours and examination scores? Why/ why not?
Extra question 2: Do you think there are some more functions suitable for the analysis of the data? (i.e. non-linear ones?)

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)

1. House Price v.s. Age of the House
2. Rate of a chemical reaction v.s. Concentration of reactants.
3. Employee Salaries v.s. Years of experience
4. Heart Rate v.s. Duration of exercise

4 Verification

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:

Correlation Coefficient

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.

Residuals

Residual is the difference between predicted values of y (dependent variable) and observed values of y.
It is very intuitive to understand: the smaller the residuals, the better the model is. (Or, is it?)

The Sum of Squares/ Variance/ Standard Deviation

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

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

which is very similar to the variance/ standard deviation we studied in secondary school.