VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY
INTERNATIONAL UNIVERSITY
DEPARTMENT OF MATHEMATICS

STATISTICS, PROJECT REPORT:
STUDYING TEMPERATURE INCREMENT IN VIETNAM

Lecturer: Dr. Nguyen Minh Quan

Group Members: Le Nguyen Dang Khoa - MAMAIU19008
Nguyen Minh Quan - MAMAIU19036
Nguyen Ha Uyen - MAMAIU19017

Ho Chi Minh City, Vietnam
December, 2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays, global warming has been rising as one of the most important concerns of human society. First mentioned by James Hansen in 1988 [1], this scientific term refers to the increment of global surface temperature, driven by rising greenhouse gas levels in the atmosphere. Statistical data show that over the last century, average temperature has risen by approximately 2.3°C due to industrial development [2]. It induces numerous affects on our ecosystem, including extreme weather conditions like droughts, hurricanes and floods. In Vietnam, such disasters brought catastrophic impacts on people's livelihoods and the economy: floods in 2020 caused electricity shortage, infrastructure destruction, 280 deaths and an economic losses of around 35 billion VND [3].

Understanding that global warming has the potential to cause serious damage to the planet, scientists have been conducting multiple studies attempting to derive solutions. For most of them, investigating recorded temperature data plays an important role: not only can we recognize temperature increment, but we also can verify the effectiveness of applied environmental strategies. A decreasing pattern in recorded temperature of a country may signal that the government policies on greenhouse gas emission of factories are working well, while a converse pattern may suggest that another solution should be applied.

For the above purpose, in this project, our group attempt to investigate whether there is an increment in global temperature by analysing recorded data. Besides, we also provide a statistical summary of the data and a regression model for predicting future temperatures. The project results can either be used as evidences supporting policies against global warming, or statistics in studies concerning relationships among climate aspects.

By submitting this project report, we declare that the entirety of the work contained therein is our own original work, that we are the sole author. We really appreciate the assistance of our Instructor, Dr. Nguyen Minh Quan, for his much-valued supports and guidance throughout the

lab sessions as well as his encouragement and optimism during the online lectures.

Chapter 2

Data Analysis

2.1 General Information

For the purpose of this study, we used the dataset [4] provided by Berkeley Earth, a U.S. non-profit organization affiliated with Lawrence Berkeley National Laboratory. More information about Berkeley Earth can be found on their official website [5]. A screenshot of the dataset after implemented into Python by the `pandas` library is shown below.

	dt	AverageTemperature	AverageTemperatureUncertainty	Country	year	month
0	1743-11-01	4.384	2.294	Åland	1743	11
1	1744-04-01	1.530	4.680	Åland	1744	4
2	1744-05-01	6.702	1.789	Åland	1744	5
3	1744-06-01	11.609	1.577	Åland	1744	6
4	1744-07-01	15.342	1.410	Åland	1744	7
...	...	...	...	...	...	...
544806	2013-04-01	21.142	0.495	Zimbabwe	2013	4
544807	2013-05-01	19.059	1.022	Zimbabwe	2013	5
544808	2013-06-01	17.613	0.473	Zimbabwe	2013	6
544809	2013-07-01	17.000	0.453	Zimbabwe	2013	7
544810	2013-08-01	19.759	0.717	Zimbabwe	2013	8

544811 rows × 6 columns

Figure 2.1: The Dataset

The dataset contains monthly average temperature of 243 countries across all continents, from November 1743 upon December 2013. It has 6 distinct attributes, namely:

- dt: the date associated with the recorded temperature;
- AverageTemperature: the temperature recorded, in celsius;
- AverageTemperatureUncertainty: the 95% margin of error around the average;

- Country: the country associated with the recorded temperature;
- year, month: the year and month from the date in the first column.

2.2 Sample Selection

A screenshot of the Vietnamese portion of the original dataset is given below, along with a summary from the `pandas` library and a scatterplot from the `matplotlib` library.

	dt	AverageTemperature	AverageTemperatureUncertainty	Country	year	month
0	1825-01-01	19.539	2.026	Vietnam	1825	1
1	1825-02-01	19.797	1.533	Vietnam	1825	2
2	1825-03-01	22.176	1.955	Vietnam	1825	3
3	1825-04-01	25.197	2.123	Vietnam	1825	4
4	1825-05-01	26.235	1.358	Vietnam	1825	5
...	...	...	...	...	...	...
2088	2013-04-01	25.887	0.515	Vietnam	2013	4
2089	2013-05-01	27.443	0.389	Vietnam	2013	5
2090	2013-06-01	27.623	0.299	Vietnam	2013	6
2091	2013-07-01	27.109	0.545	Vietnam	2013	7
2092	2013-08-01	27.026	0.281	Vietnam	2013	8

2093 rows × 6 columns

Figure 2.2: Derived Vietnam's Temperature Data

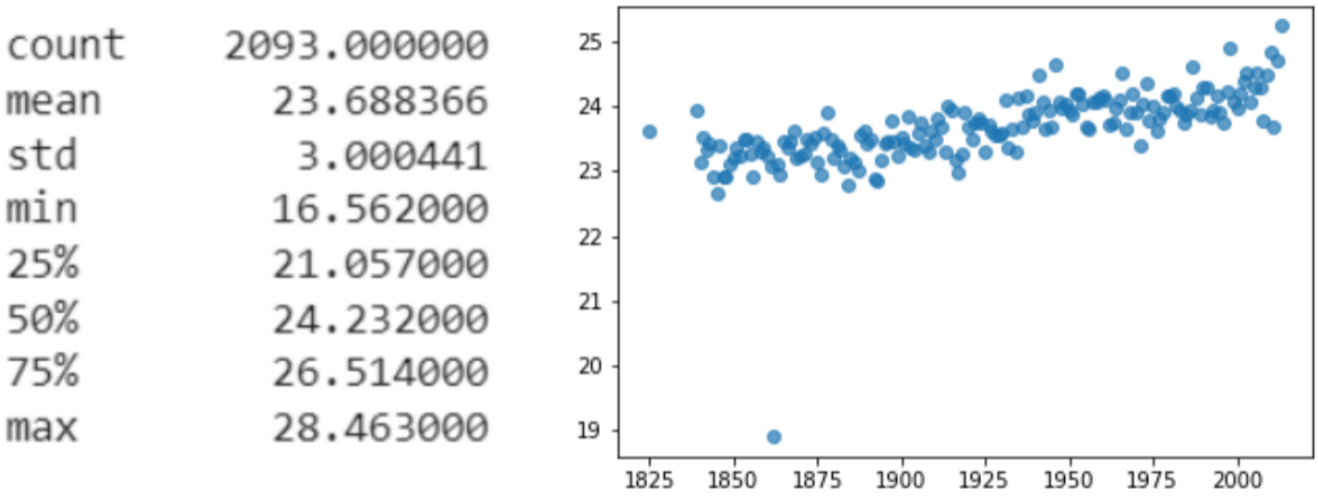

Figure 2.3: Summary & Scatterplot of Vietnam's Temperature Data

As seen from the above results, the data points and the quantiles of the portion distribute quite evenly around the mean (approximately 23.69). Also, there is potentially a linear relationship between average temperature and year: calculations in Python yield a correlation of 0.119.

For testing the increment of Vietnam's average temperature, we derive two sub-portions of the original portion, corresponding to the time intervals 1930 - 1960 and 1985 - 2013, as two samples. The summaries and scatterplots of these samples are given below.

count	360.000000
mean	23.945347
std	2.936334
min	17.228000
25%	21.259750
50%	24.373000
75%	26.757250
max	28.158000

count	344.000000
mean	24.220663
std	2.808183
min	17.726000
25%	21.793500
50%	24.954000
75%	26.775750
max	28.304000

Figure 2.4: Summary of Samples

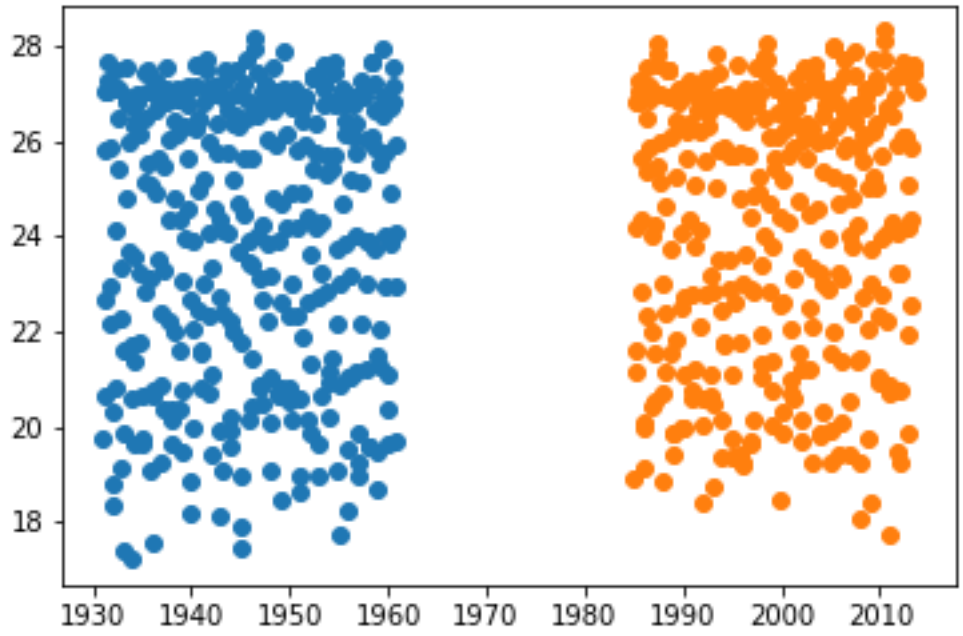

Figure 2.5: Scatterplot of Samples

Observing the summaries, it can be seen that the average of the first sample is slightly smaller than that of the second one. Hence, it is reasonable to propose that **the (actual) average temperature during 1930 - 1960 is smaller than the average temperature during 1985 - 2013.**

2.3 Hypothesis Testing

We now apply statistical methods on the derived samples to verify the acceptability of the claim proposed above, up to some level of confidence.

2.3.1 The $t-$test

First introduced by its author William S. Gosset in his 1908 publication [?], the $t-$ test is a statistical hypothesis test usually used when a scaling term is unknown and replaced by an estimation from a data sample. The $t-$ test are most commonly used for the purpose of comparing the mean of a population to a specified value, or comparing the means of two populations. Two tables describing the $t-$ test procedure for these two usages are given below.

Prompt	We are given a sample S of a population X with mean μ and a specified value μ_0. We attempt to verify a claim H_0 (called the **null hypothesis**) on μ and μ_0.
Calculation Step	Step 1: Calculate the sample mean $\overline{X}$, the sample size n and the sample variance s^2; Step 2: Calculate the test statistic $TS = (\overline{X} - \mu_0)/(s/\sqrt{n})$; Step 3: Calculate the appropriate $p-$value p for H_0.
Conclusion	If p is smaller than a given **level of confidence** α, we reject H_0. If conversely $p \geq \alpha$, we do not have enough evidence to reject H_0.

Table 2.1: The $t-$ test Procedure - One Sample

Prompt	We are given two samples S_1, S_2 of two populations X_1, X_2 with means μ_1, μ_2. We attempt to verify a claim H_0 on μ_1 and μ_2.
Calculation Step	Step 1: Calculate the sample means $\overline{X}_1, \overline{X}_2$, the sample sizes n_1, n_2 and the sample variances s_1^2, s_2^2; Step 2: Calculate the pooled sample variance $s_p^2 = ((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2)$; Step 3: Calculate the test statistic $TS = (\overline{X}_1 - \overline{X}_2)/\sqrt{s_p^2(1/n_1 + 1/n_2)}$; Step 4: Calculate the appropriate $p-$value p for H_0.
Conclusion	If p is smaller than a given **level of confidence** α, we reject H_0. If conversely $p \geq \alpha$, we do not have enough evidence to reject H_0.

Table 2.2: The $t-$ test Procedure - Two Samples

The $p-$value formula varies depending on the claim H_0, as described below. Note that T_i is a random variable of $t-$distribution with i degrees of freedom.

H_0	$\mu = \mu_0$	$\mu \leq \mu_0$	$\mu \geq \mu_0$
$p-$value	$2 \cdot \mathbb{P}(T_{n-1} \geq \lvert TS \rvert)$	$\mathbb{P}(T_{n-1} \geq TS)$	$\mathbb{P}(T_{n-1} \leq TS)$

Table 2.3: $p-$value Formula - One Sample

H_0	$\mu_1 = \mu_2$	$\mu_1 \leq \mu_2$	$\mu_1 \geq \mu_2$
$p-$value	$2 \cdot \mathbb{P}(T_{n_1+n_2-2} \geq \lvert TS \rvert)$	$\mathbb{P}(T_{n_1+n_2-2} \geq TS)$	$\mathbb{P}(T_{n_1+n_2-2} \leq TS)$

Table 2.4: $p-$value Formula - Two Samples

2.3.2 Test Results

Corresponding to our claim, the null and alternative hypothesis is

$$\begin{aligned} H_0 : &\quad \mu_1 \geq \mu_2 \\ H_a : &\quad \mu_1 < \mu_2 \end{aligned}$$

and the formula for $p-$value is

$$p = \mathbb{P}(T_{n_1+n_2-2} \leq TS).$$

Calculations in Python yield a $p-$value of 0.088. It means that we can reject H_0 and support our claim H_0 with a confidence level of at least 0.008. If otherwise we use the time intervals 1825 - 1925 and 1926 - 2014 for comparison, the corresponding $p-$value is $2.019 \cdot 10^{-7}$, and thus we can reject H_0 at almost any level of confidence.

2.4 Regression & Prediction

We now apply the least square method to derive a linear approximation of the dataset, which can be used for future predictions.

2.4.1 Least Square Regression

Independently discovered by Legendre [6] and Gauss [7], the least squares regression is one of the most standard and widely used statistical methods in data-fitting. Suppose that a researcher believes that there is a linear relationship between two variables, say x and y, that he/she is concerned about. The researcher then attempts to describe this relationship as accurate as possible, by the *regression line*

$$y = \alpha x + \beta$$

where α and β are unknown parameters. In particular, the researcher makes n observations to obtain n data points $(x_1, y_1), ..., (x_n, y_n)$ and set

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i, \forall\, i = \overline{1, n}.$$

Here the $\hat{y}_i$s are called *fitted values* from the regression line, while $\hat{\alpha}$ and $\hat{\beta}$ are the *estimators* (for α and β) to be determined. The *residual* of each data point is then given by

$$\hat{u_i} = y_i - \hat{y}_i, \forall\, i = \overline{1, n}.$$

The least squares method, also known in this case as the Ordinary Least Squares (OLS) method, chooses $\hat{\alpha}$ and $\hat{\beta}$ that minimize the *residual sum of squares* (RSS) defined as

$$L := \sum_{i=1}^{n} \hat{u_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

Setting the first order partial derivatives of L with respect to $\hat{\alpha}$ and $\hat{\beta}$ equal to zero

$$\frac{\partial L}{\partial \hat{\alpha}} = -2 \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \quad \text{and} \quad \frac{\partial L}{\partial \hat{\beta}} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

yields the OLS estimators for $\hat{\alpha}$ and $\hat{\beta}$ as

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \quad \text{and} \quad \hat{\alpha} = \overline{y} - \hat{\beta}\, \overline{x},$$

where $\overline{x}, \overline{y}$ are the sample means of the observations. The *estimated regression line* is given by

$$y = \hat{\alpha} x + \hat{\beta}$$

which can be used as an approximation of the original sample.

2.4.2 Regression Results

Applying the least square regression on the provided dataset, using the `polyfit` function from the `numpy` library yields the estimated regression line as

$$y = 0.007x + 9.244.$$

A plot of the dataset and the estimated regression line is given below.

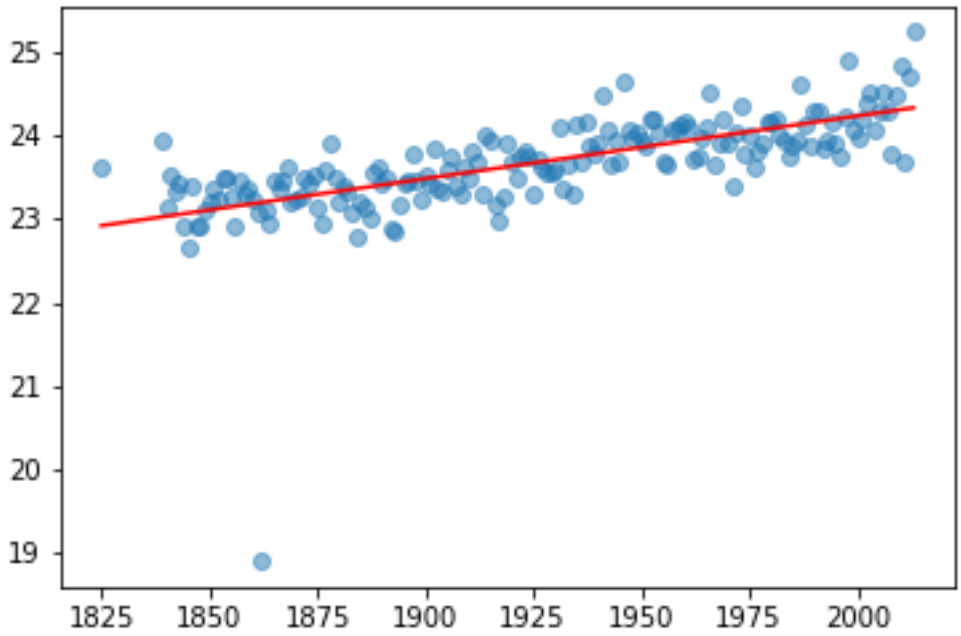

Figure 2.6: Data Scatterplot (Blue) and Regression Line (Red)

Using the estimated parameters, we can predict the average global surface temperature for the next decade conveniently. The results are provided below.

Year	2014	2016	2018	2020	2022	2024
Temperature	24.334	24.349	24.364	24.379	24.394	24.409

2.5 Summary

Summarizing on the results and visualizations yields the following key takeaways:

- The hypothesis testing result on temperature increment is much more significant with the larger time intervals. It may signals two phenomenons: first, there is an increment in global temperature, and larger sample size overwhelmes the data noises and provides a more accurate result; second, the increment of temperature with respect to year is rather slow, thus the pattern is more easily seen by observing longer time intervals;
- The estimated regression line fits appropriately among the dataset, indicating that there exists weak linear relationship between average temperature and year. However, in reality the data varies around the approximation line, so the predicted temperature should be consider as the center of a confidence interval rather than an estimation. For instance, the average temperature in 2022 is predicted to lies in the interval, say, $(23.894, 24.894)$.

Bibliography

[1] S. Weart, *The Public and Climate Change: The Summer of 1988*, The Discovery of Global Warming, American Institute of Physics, January 2020.
`https://history.aip.org/climate/public2.htm#S1988`

[2] V. Masson-Delmotte, P. Zhai, A. Pirani and S. L. Connors, *Climate Change 2021: The Physical Science Basis*, Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 2021.
`https://shorturl.at/zJPZ1`

[3] Vietnam Disaster Management Authority, *Losses on Weather Disaster in 2020.*
`https://shorturl.at/pqxIS`

[4] Berkeley Earth, *Climate Change: Earth Surface Temperature Data.*
`https://shorturl.at/lprNR`

[5] Berkeley Earth Official Website.
`http://berkeleyearth.org/about/`

[6] A. M. Legendre, *Nouvelles methodes pour la determination des orbites des cometes*, F. Didot, Paris, 1805.

[7] C. F. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, F. Perthes & I. H. Besser, Hamburg, 1809.

[8] Google Colaboratory session, with code & execution result available.
`https://colab.research.google.com/drive/1dQyilyjkujVRNe7xet3bSY908o12frv_`