École polytechnique fédérale de Lausanne

Institute of Mathematics

Course Project

Autumn Semester 2021

Statistical Analysis and Visualization of Meteorological Data

Carried out for the Statistical Computation and Visualization course

Under the supervision and direction of
Professor Mehdi Gholam

Authored by
Luca Bracone, Blerton Rashiti,
Luca Nyckees, Kieran Vaudaux

Contents

1 Introduction

Statistical and visual analysis of data are major components of the general data science domain. In this project, we look at various datasets revolving around meteorological recordings from various stations within Switzerland. Statistical analysis of meteorological data plays an important role in understanding and modeling key features in climate change, as well as making short-term predictions on certain meteorological elements. Here, we are interested in providing an efficient pipeline aiming at analysing meteorological data through basic statistical methods such as linear regression and time series analysis. Moreover, we concentrate in providing a significant amount of visualization tools to combine with the statistical results.

The main question we try to answer revolves around the mean temperature element, recorded across Switzerland. We formulate it as follows.

Question. Is there a significant increase in the average temperature trend in Switzerland from 1901 to the present day ?

To try to answer this question, we will first focus on the evolution of the average temperature in Geneva. This will allow us to refine and improve our statistical study on the Geneva observatory data, before extending it to the rest of the weather stations.

2 Evolution of the Mean Temperature at Geneva Observatory

In this section, we will focus on the modelling of our data, which we will see as Times Series.

We have the daily average temperatures at the Geneva observatory from 1^{er} January 1901 to August 2, 2021. That is to say 44044 average temperature record, spread over a period of more than 120 years. As this amount of data is very large, we have chosen to proceed in stages. To do this, we will first look for the presence of a significant increase in the trend in the Time Series of annual mean temperatures at the Geneva observatory, which we have calculated from the daily data. We can then make our study more complex by looking at the time series of monthly, weekly and daily mean temperatures.

2.1 Annual Mean Temperature at Geneva Observatory

It seems natural to ask whether transforming our data by averaging the annual temperature is relevant. Indeed, knowing that during a year the temperature can vary from $-10°C$ in winter to more than $30°C$ in summer, does it really make sense to consider the average of these values? How do we correctly interpret these values and what would it really mean if there was a significant increase in the trend from 1901 to the present? While we have more refined data than annual average temperatures, looking at this one could be debated. However, as many studies also look at annual mean temperatures, we will accept, for the purposes of this project, that the presence of a significant increase in the trend of annual mean temperatures in Geneva would be an additional indication of the presence of climate warming (in Geneva). To confirm this idea, Figure 1 below allows us to see that the global behaviour of the Time Series of annual averages is similar to that of the Time Series of annual median temperatures, as well as the standard deviation of the annual average temperatures seems to be homoskedastic. This supports the idea that the annual mean temperatures are not overly affected by the presence of extreme temperatures or by the increase in temperature variability during a year.

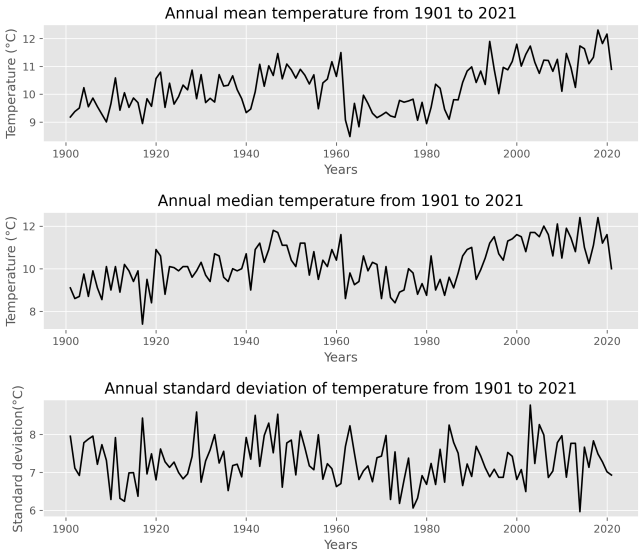

Figure 1: Time series plot of mean, median and standard deviation annual temperatures at the Geneva observatory

The Time Series of average temperatures appears to have an increasing trend over the years, but with a sudden cooling from 1962.

In order to distinguish more clearly the periods that correspond to a warming or not, we present on Figure 2 the histogram of the annual anomalies that we have standardised. We recall that the temperature anomaly is the difference between the temperature measured in a place (here Geneva), compared to the normal average temperature observed in this same place.

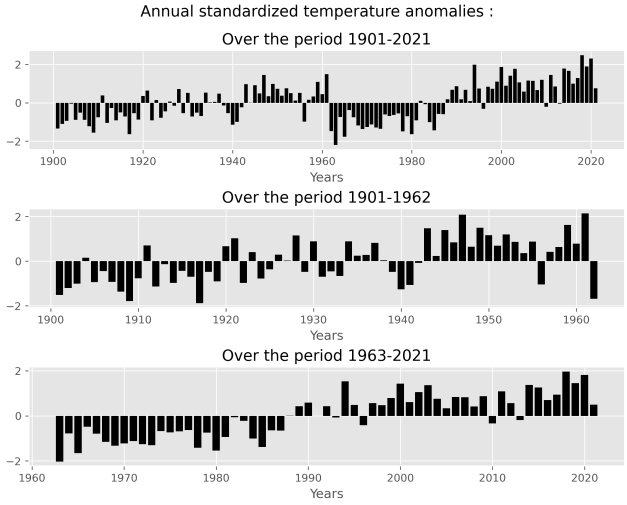

Figure 2: Histogram of annual temperature anomalies standardised over several periods

The visual analysis of these histograms allows us to distinguish four periods 1901-1942, 1943-1961, 1962-1987 and 1988-2021. During the first period, the anomalies tend to be negative, then positive during the second. From the third period, we again observe a cycle of negative and then positive anomalies, but this time more pronounced. This seems to be consistent with the Time Series of average temperatures, in which we had observed an increasing trend but a significant decrease in temperature at the beginning of this third period.

In order to model our data, we will try to follow the principles of parsimony as much as possible,

in order to choose the simplest model that effectively explains our data.

If we denote the Time Series of annual averages by$\{\mathbf{A}_t\}_t$ $t = 1901, ...2021$, one of the simplest models we could propose is that our observations $\{\mathbf{A}_t\}_t$ are from a normal distribution, $\mathbf{A}_t \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. To test this we will first compare the empirical distribution of our data with the distribution of a normal distribution of mean $\bar{\mathbf{A}}$ and variance S^2.

In Figures 3 and 4, we see that our empirical distribution is quite close to that of a normal distribution, despite the fact that we only have 121 observations.

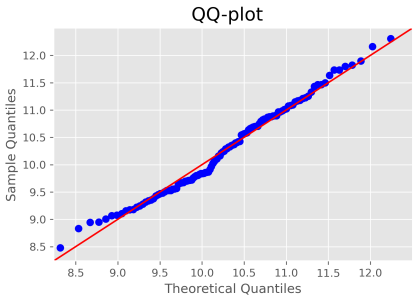

Figure 3: QQ-plot of average annual temperatures

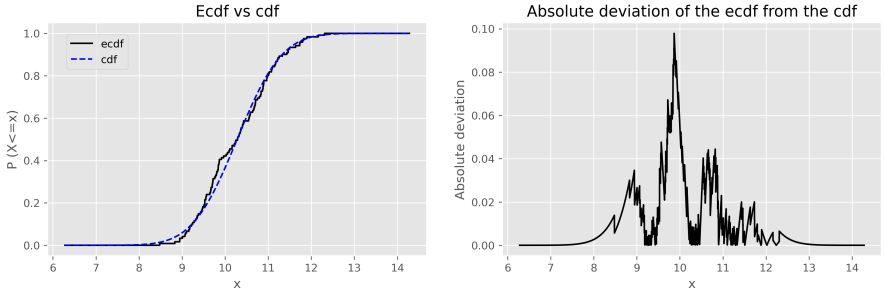

Figure 4: Empirical distribution vs theoretical distribution

Nevertheless, Figure 5 highlights a significant correlation between the annual average temperatures. This leads us to question the independence of the $\mathbf{A}_t$ observations. Indeed, if $\{\mathbf{A}_t\}_{t=1901}^{2021}$ were independent and identically distributed, we should have that the auto-correlations as well as the partial auto-correlations on Figure 5 are approximately in the red zone, which corresponds to an approximate confidence interval for the auto-correlations in the case of an iid sequence.

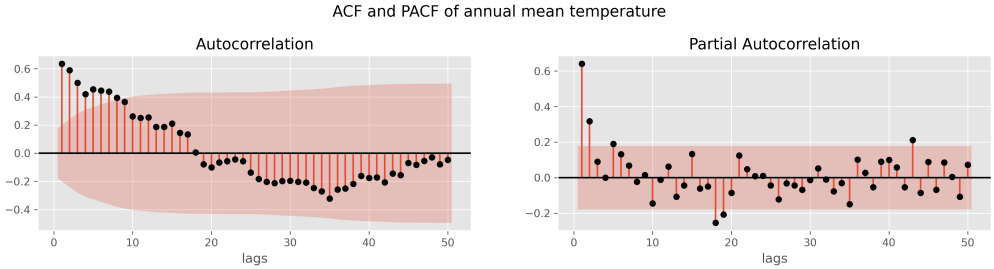

Figure 5: Auto-correlation and partial auto-correlation function of the mean temperature

In order to test the independence of our observations, we will use the portmanteau test and several of its variations, namely the Ljung-Box, Box-Pierce and McLeod and Li tests. Figure 1 shows the

p-values of these three tests for different numbers in the auto-correlation sequence considered in the test statistics. These tests all have the null hypothesis that the sequence $\{\mathbf{A}_t\}_{t=1901}^{2021}$ is iid.

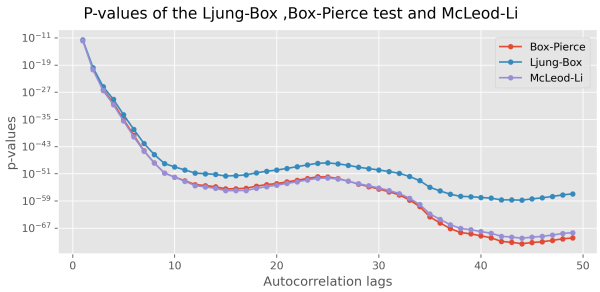

Figure 6: Statistical tests on the mean temperature

Since the p-values of these tests are all less than 10^{-11}, we can conclude that at any significance level greater than 10^{-11}, the sequence of mean annual temperatures is not from the model $\mathbf{A}_t \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.

To take these dependencies into account, we will consider a more general model from the Time Series study. First of all, we will test the stationarity of our Time Series in order to know which Time Series model could be applied to our data. Without going into formal definitions, a Time Series is stationary if it has a constant mean over time and if its variance is also time invariant. The test we use to test the stationarity of our Time Series is the Augmented Dickey-Fuller test (ADF), which tests the null hypothesis that a unit root is present in the Time Series, which would make the Time Series non-stationary. This test gives us a p-value of $p = 0.8711$, which is far from significant. This result tends to make us think that the Time Series is not stationary, which can certainly be explained by the presence of an increasing trend that we had already noticed visually on Figure 1. To test this we use another version of the Augmented Dickey-Fuller test to test the trend-stationarity of the Time Series. With this test we obtain a p-value of $p = 0.0162$ which is significant, at the standard significance level of $\alpha = 0.05$ for example.

Following this result, we are therefore led to first model the trend of our Time Series before trying to model our data with a stationary Time Series model. To model our trend we use a generalized linear regression, i.e. a linear regression in which we do not assume the independence of our errors. We chose to model the trend as an affine function of time $\mathbf{A}_t = \beta_0 + \beta_1 t + \epsilon_t$ with $\epsilon = (\epsilon_{1901}, ..., \epsilon_{2021})^T \sim \mathcal{N}(0, \mathbf{\Sigma})$, so as to keep the model simple and to be able to easily infer the sign of β_1, which will allow us to detect or not a significant growth of the mean temperature trend. To do this, we estimated the covariance matrix of ϵ by the Toeplitz matrix generated by the sequence of auto-correlations of our Time Series. Figure 7, shows us the fit of the trend estimate by a line. We notice visually that the fit is quite good overall, but that the line has difficulty in approximating the period 1960-1990 correctly. This is due to the fact that, as we saw in Figure 1, the mean annual temperature drops sharply in 1962 before resuming a "normal" behaviour in relation to the rest of the Time Series.

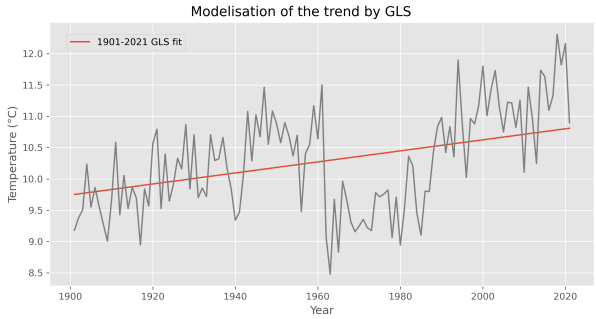

Figure 7: GLS estimate of the trend

Initially, we wondered whether this sudden change was due to a change in equipment or to the automation of the temperature recordings, which could have affected the average daily temperatures and, by transitivity, the average annual temperatures. But, given the very large temperature difference, this seems unlikely and we have not found any information that would support this hypothesis. However, by doing some additional research we were able to find articles and websites on which the years 1962-1964 were described as particularly cold years with harsh winters. This sudden cooling could then be simply due to a temporary local cooling. We then considered this sudden cooling as a rare event, and in order not to affect our study too much, we chose to estimate the trend of the annual mean temperatures over the periods 1901-1961 and 1961-2021 separately. This led to the estimate in Figure 2 below.

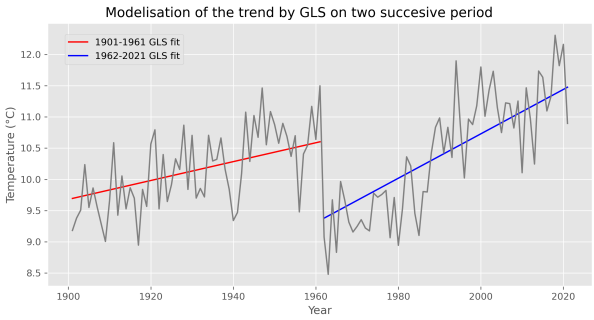

Figure 8: GLS estimate of the trend on two period

The results of the two GLS estimates are summarised in the tables 9 and 10 which represent the results for the period 1901-1961 and the period 1962-2021 respectively. In the table 9, we see that the p-value of the t-test is $p = 0.1$, on the first period we can only reject the null hypothesis, that $\beta_1 = 0$, at a significance level of $\alpha = 0.1$. However, the 95% confidence interval which is given contains 0, so we cannot conclude anything about the positivity of β_1 at the threshold of $\alpha = 0.05$ on this period.

Dep. Variable:	Mean	**R-squared:**	0.045
Model:	GLS	**Adj. R-squared:**	0.029
Method:	Least Squares	**F-statistic:**	2.790
Df Model:	1	**Prob (F-statistic):**	0.100
Df Residuals:	59	**Log-Likelihood:**	-27.276
No. Observations:	61	**AIC:**	58.55
		BIC:	62.77

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
β_0	-19.0425	17.477	-1.090	0.280	-54.013	15.928
β_1	0.0151	0.009	1.670	0.100	-0.003	0.033

Figure 9: GLS Regression Result on the period 1901-1961

On the other hand, the table 10 allows us to reject the null hypothesis, which is that $\beta_1 = 0$ over the period 1962-2021, at the significance level $\alpha = 0.05$ because we obtain a p-value associated with the t-test of β_1 of $p = 0.042$. Moreover, the 95% confidence interval of β_1 is [0.001,0.070], which contains only strictly positive values. If our model subsequently proves to be consistent with our data, then we would have a significant indication of an increasing trend in mean temperatures over the period 1962-2021.

Dep. Variable:	Mean	**R-squared:**	0.070
Model:	GLS	**Adj. R-squared:**	0.053
Method:	Least Squares	**F-statistic:**	4.333
Df Model:	1	**Prob (F-statistic):**	0.0418
Df Residuals:	58	**Log-Likelihood:**	-34.044
No. Observations:	60	**AIC:**	72.09
		BIC:	76.28

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
β_0	-60.3638	34.008	-1.775	0.081	-128.439	7.711
β_1	0.0355	0.017	2.082	0.042	0.001	0.070

Figure 10: GLS Regression Result on the period 1962-2021

We will now consider the Time Series $\{\mathbf{A}_t\}_{t=1901}^{2021}$ which is the Time Series $\{\mathbf{A}_t\}_{t=1901}^{2021}$ from which we have subtracted the trend calculated above, over two disjoint periods. Figure 11 allows us to visualise this new Time Series.

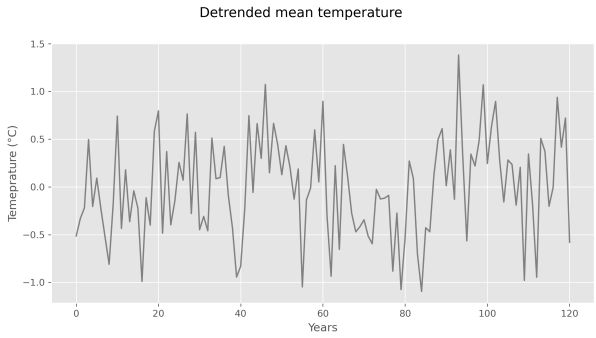

Figure 11: Time Series of the detrended annual mean temperature

Therefore, it is interesting to test again the independence of the elements of the new Time Series. For this we proceed as before, we have calculated the auto-correlation and partial auto-correlation sequences of the new Time Series, which can be seen in Figure 12. Almost all terms of the sequence are within the approximate confidence interval in red, but one can see that some terms are still outside. Moreover, for a number of lags greater than 20, we see in Figure 13 that the p-values associated with the Ljung-Box and Box-Pierce tests allow us to reject the null hypothesis of independence of the Time Series elements.

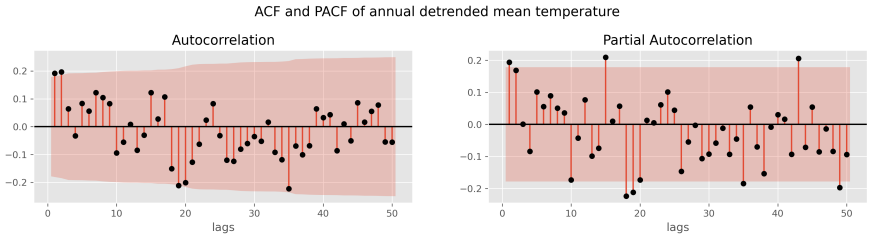

Figure 12: Auto-correlation and partial auto-correlation function of the detrended mean temperature

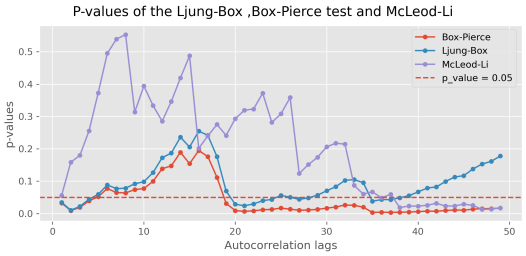

Figure 13: Statistical tests on the detrended temperature

Even though removing the trend from the Time Series of mean annual temperatures has reduced the dependence between observations, there is still enough dependence between observations to use a Times Series model to model $\{\tilde{\mathbf{A}}_t\}_{t=1901}^{2021}$. Especially since this new Time Series is now stationary. Indeed, the Augmented Dickey-Fuller test on this one gives us a p-value of $p = 2.0748e-08$ and thus allows us to reject the null hypothesis of non-stationarity at a significance level of $\alpha = 0.05$.

Figure 14 allows us to compare several models for the Time Series $\{\tilde{\mathbf{A}}_t\}_{t=1901}^{2021}$ thanks to the Akaike information criterion (AIC). We have tried to model the Time Series as arising from an $ARMA(p, q)$ for $p \in \{0, ..., 3\}$ and $q \in \{0, ..., 9\}$. Our choice of restricting p and q is mainly due to the fact that our Time Series is not very large and that we wanted to try to keep the model as simple as possible to model our data.

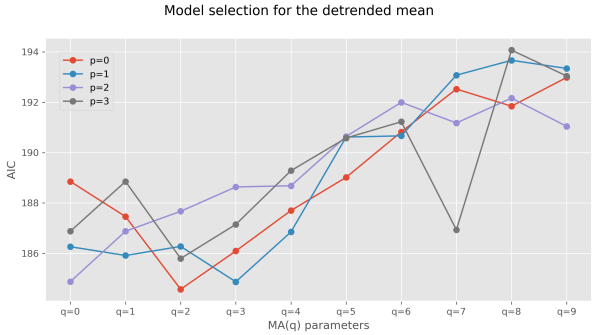

Figure 14: Plot of the AIC for different models

By choosing the model which minimises the AIC, we are led to consider the model $ARMA(0, 2)$ to model the Time Series $\{\tilde{\mathbf{A}}_t\}_{t=1901}^{2021}$, which amounts to considering a model $MA(2)$.

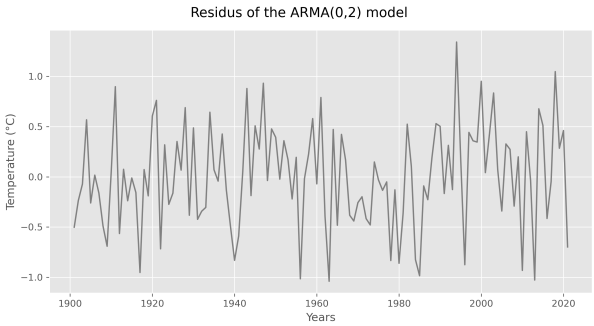

Figure 15: Plot of the residues of the $MA(2)$ model for the detrended mean temperature Time Series

In order to confirm the consistency of our model, we tested the independence and distribution of the residuals obtained. Figure 16 allows us to see that the residuals of our model do not allow us to reject the null hypothesis of independence of the Ljung-Box, Box-Pierce and McLeod-Li tests, at a significance level of $\alpha = 0.05$, except for the McLeod-Li test which rejects the null hypothesis for a value of the auto-correlation lags.

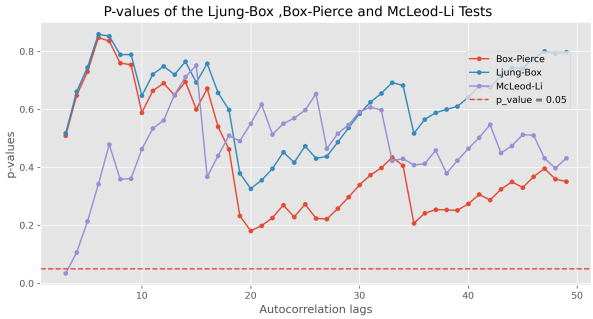

Figure 16: Statistical tests on the residues of the $MA(2)$ model for the detrended mean temperature Time Series

Now that our data appear to be independent we can use the Mann-Whitney U test, which compares the distribution of the first n of data with the distribution of the rest of the data. This test has the null hypothesis that the probability that a variable generated by the first distribution is greater than a variable generated by the second distribution is equal to the probability that a variable generated by the second distribution is greater than a variable generated by the first distribution. In Figure 17, we have the p-values obtained by this test for different values of $n = 1, ..., 119$.

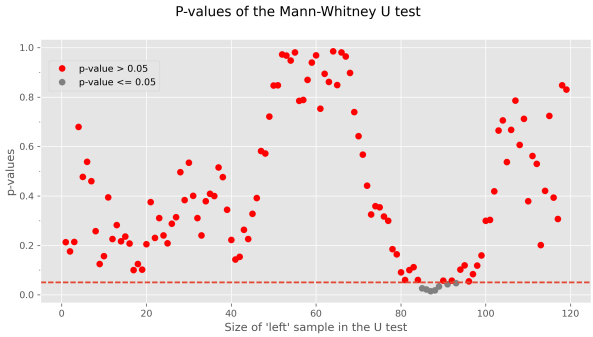

Figure 17: P-values of the Mann-Whitney U test for different sizes of the left sample

We therefore seem to have obtained residuals that are independent of each other. Moreover, after performing the Jacques-Bera test, which has the null hypothesis that our data are from a normal distribution, and the Goldfeld-Quandt test, which has the null hypothesis that the data are homoskedastic, we get the p-values :

- Normality test : p-value = 0.724945701756591
- Heteroskedasticity test : p-value = 0.15657168727289075

Both p-values are significant at a threshold of $\alpha = 0.05$, so we cannot reject the null hypotheses of normality and homoscedasticity.

2.2 Analysis using multiple linear regression

In this subsection we make use of basic linear models to capture the trend of the average temperature over the course of time. In order to make use of the ordinary least squares estimator, we separate our observations into chunks that we will assume are i. i. d.̇What we do is that we group each observation based on its date. As such we obtain 365 data chunks of the form

$$D_1 = \{TG(\text{1st Jan } 1901), TG(\text{1st Jan } 1902), TG(\text{1st Jan } 1903), \dots\}$$
$$D_2 = \{TG(\text{2nd Jan } 1901), TG(\text{2nd Jan } 1902), TG(\text{2nd Jan } 1903), \dots\}$$
$$\dots$$

where TG is the average temperature for that day. We continue with the assumption that the variables within a certain chunk are independent and identically distributed. For each chunk we fit a simple linear model:

$$TG = \beta_0 + \beta_1 \text{year} + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I)$. If we can show that for a sufficiently large number of chunks, that $\beta_1 > 0$ with a significant degree of certainty, then we will have that as years go by, the average temperature increases.

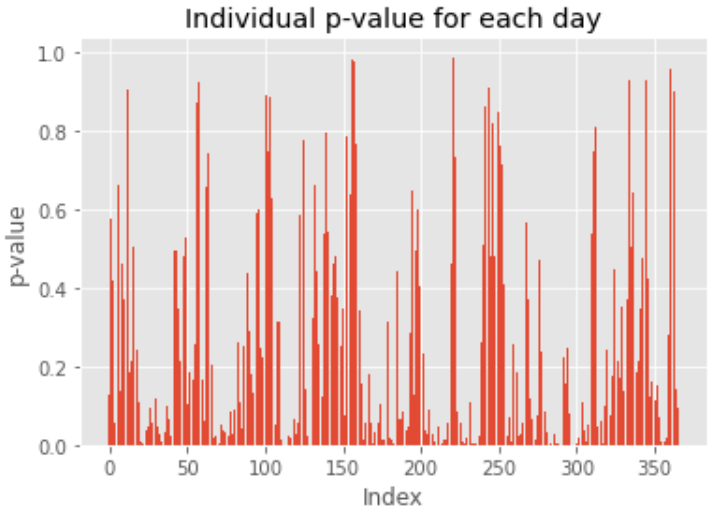

Figure 18: The p-values for the hypothesis H_0: $\beta_1 = 0$ for each chunk (day of a year). Only about a third are below 0.05

In Figure 18 we show the result of this operation. We find that there is only a little number of slopes β_1, therefore it is unclear if it is possible by using this unorthodox method to conclude that the average temperature increases as time increases. For completeness sake, we plot in Figure 19 the estimated coefficients. Though most of them are not statistically significant, a large majority are positive. The insignificance of the results we found in this section are likely due to having too little data after separating the dataset into chunks, though it was necessary to do it in order to try this method in the first place.

Figure 19: The estimated β_1 coefficients for each chunk. Note that the unit in the y-axis are divided by 10. So, a β_1 of 0.1 means each year the temperature for that given day increases by 0.01 on average

3 Result

In this preliminary part of our study on the evolution of the mean temperature trend, we have restricted our analyses to the mean annual temperatures at the Geneva observatory. This allowed us to obtain initial results regarding the increase in the annual trend of mean temperatures in Geneva. We found the presence of an increase in the annual trend, but that this was significant at a significance level of $\alpha = 0.05$ only over the period 1962-2021.

4 Discussion on the future objectives of the project

In the next part of our study, we will try to refine our results by applying a similar approach to the Time Series of monthly, then weekly and finally daily averages. Then, we will try to extend our results on the evolution of average temperatures to data from other meteorological stations in Switzerland. In parallel to this, we will try to apply other methods to answer our problem, such as the study of extremes, which could eventually allow us to model the sudden drop in temperature that we observed in 1962. In addition, we are also developing an "application" to allow a simplified and more interactive visualisation of our results.

5 Developping an App

We are working on developping an application to complement the visual and interactive aspect of the data analysis. We make use of the *Streamlit* library to implement various user-friendly features appearing on a local host platform. A version of the application can already found on the GitHub repository and is ready to use by following the simple commands here-under.

First, open a shell/terminal and go to the directory in which you saved the project. Then, go to the extension "/notebooks". This could look like this:

```
cd Desktop/SCV/project1/notebooks
```

Next, run the following command line to run the app.

```
streamlit run streamlit_app.py
```

NB. One should install *Streamlit* and *Pyviz* before doing this.

Although the application offers an easy way to interact with the data and is a fundamental aspect of the visualization pipeline, the contents we present are temporary as we may decide to include other plots and remove some results from it. The *Pyviz* library is not used yet, but we may use it later on for building sequences of graphs.