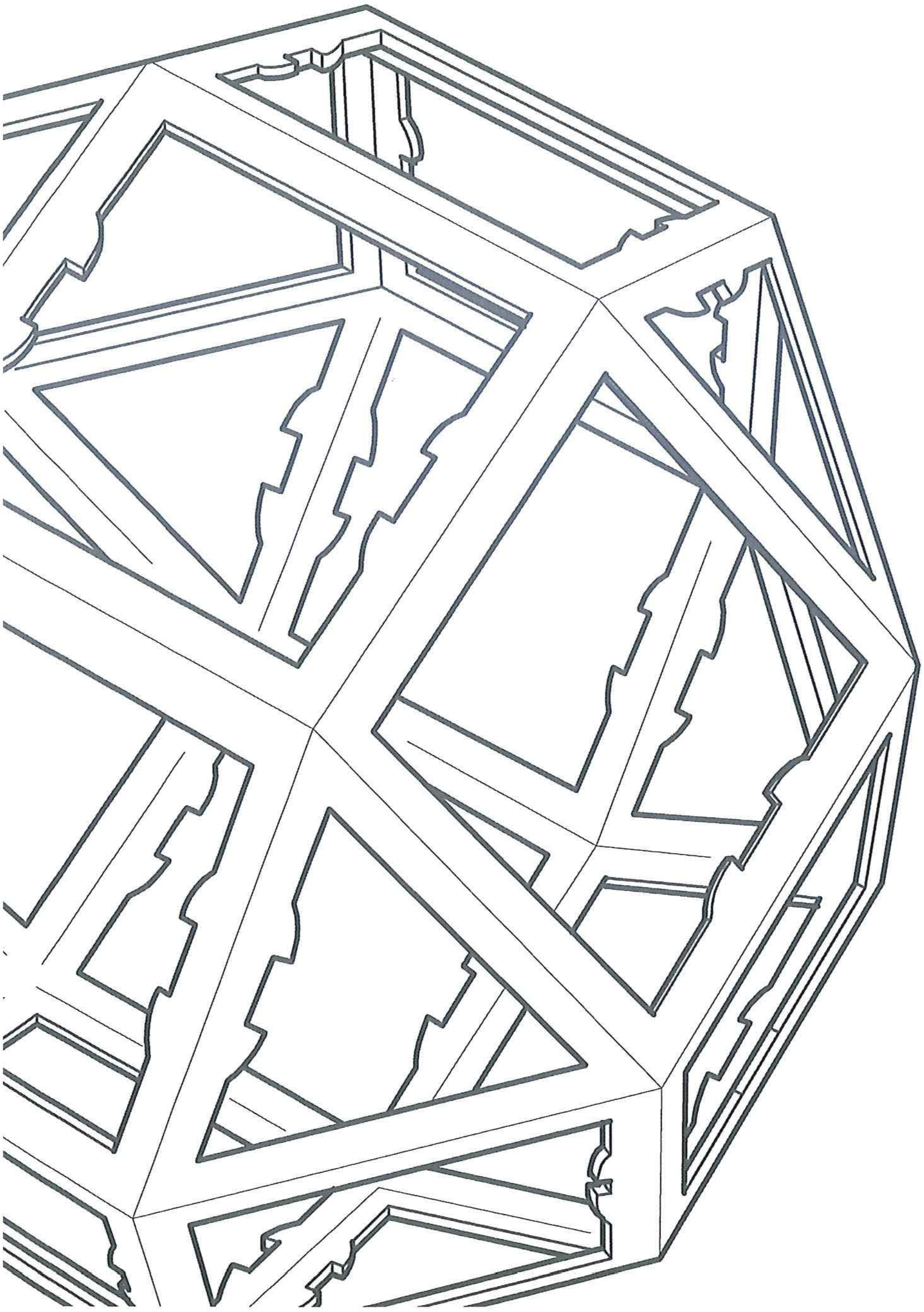
Department of Mathematics and Statistics
College of Engineering

UC
UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Summer Research Project

Weather Data Analysis using R

Ahmed Omar Yusuf



11/12

Weather Data Analysis using R

Ahmed Omar Yusuf
University of Canterbury
aoy12@uclive.ac.nz

February 13, 2012

Abstract

My primary purpose of undertaking this project during the summer was to develop computational skills in R as well as apply the techniques learnt in class to weather data which is still a relatively unexplored research area in Pakistan utilising the capabilities of R

Contents

1. Introduction

I undertook this project to expose myself to establish an understanding behind the working of R as well as apply the statistical knowledge acquired and developed through coursework completed till date. R being a predecessor to S-Plus is intensively utilised nowadays in industries ranging from management consulting, engineering to banking and finance in a wide range of applications. I decided to take up the challenge of analysing weather data through R, which I found is still a relatively un-researched area specifically using its capabilities for example in comparison to econometrics. The topic research is of a highly controversial nature which continues to baffle many scientists and ordinary citizens alike and supports lucrative multi-billion dollar industries like hybrid vehicles' production and wind mill Krohn (2000).

During the course of this project the main objectives I achieved were running statistical analysis and building linear models of different forms, variables, locations in order to establish proper relationship primarily between time and temperature as well as the compare the two locations studied. In this report I will carry out a short, brief study of weather data of two locations i.e. Lahore and Karachi in Pakistan and discuss the results as well as their implication in a wide range of areas

1.1 Dataset

Acquiring a comprehensive dataset which had a record of almost four to six decades was one of the most complex task that was accomplished through the NOAA servers in the USA. The 'factory fresh' data itself required a lot of manipulation as it had great length of patches missing during times of war, the segregation of the Indo-Pak subcontinent. To be precise 8335 of datapoints were missing for Lahore and 25443 for Karachi. Both datasets included the following parameters: Station Number, pressure, visibility, wind gust, maximum and minimum temperature for the day in question. For the purposes of my project, I manipulated the data in order to include only the dates and the corresponding temperatures. The following amendments were made:

- The data being imperial (i.e. in Fahrenheit) had to be converted in to Celsius to provide a better intuition and perspective for the analysis and for the reader.

- The dates in the data itself were present as raw numbers *eg.* 19841008 which had to be turned into proper dates such as 08/10/1984 as well as the number of years from January 1, 1970 *eg.* 14.75 in order for R to easily analyse the numbers as well as do something meaningful which in this case was building models. The last step mentioned was established by dividing the number of days from January 1, 1970 such as 5387.04 by 365.224 as explained by

- The major data sets of the respective locations were subsetted further by winters and summers. The selection of seasons itself was based on summer and winter solstices web (2009)

- Most importantly, despite my careful selection of time period due to missing values as mentioned above the dataset still had a string of values missing. Technically, such cases are also known as "Missing Completely at Random" (MCAR) as explained by Ms. J.Scheffer Scheffer (2002). There are many ways of dealing with such a scenario as listed below.

  1. Case Deletion: observations are deleted $X_1, Y_1, \ldots\ X_n, Y_n$ either pairwise or listwise,

2. Single Imputation: mean, median values are imputed in place of the missing data values,
3. Multiple Imputation: the data is replicated, imputed and analyzed separately followed by a recombination of all the data sets together. [1].

However, I took a an approach of imputing mean values to missing cases

1.2 Usage of R

R itself is a versatile tool being utilised almost in every imaginable way with extensive discussions online at the CRAN servers, as well as at sites like `http://www.r-bloggers.com/about`R Q & A blog. Also, a wide range of available 'packages' makes it a specially interesting tool kit to have for the future. However, this vast variety of packages itself is a double edged sword. As a relatively new user of R, I found a wide range of packages as well as a lack of extensive documentation such as the one's like MATLAB to be at times cumbersome although packages such as ggplot2 by Hadley Wickham made task such as making plots a lot easier.

[1] The reader should to refer to Ms. J.Scheffer's paper for a complete discussion of the above given methods is beyond the scope of this report. Scheffer (2002)

2. Methods

This section highlights the different methods used to analyse the data and their related assumptions. These are the areas that will be covered:

- Linear Models and their assumptions
- Hypothesis testing
- Checking of assumptions
- Comparison of models

2.1 What is a linear model?

The simplest linear model involves only one **independent variable** *eg.* time (in years) denoted by X and states that a **dependent variable** such as Temperature denoted by Y is related by a means of an equation. Modelling in this context refers to the development of a mathematical relationship that aims to establish and describe an equation between the X and Y variables. More specifically it aims to find out how the mean of Y, *E(Y)*. In this case, *E(Temperature)* changes with changing conditions with the assumption that the variance of temperature remains unchanged. Also, any other variable independent variable such as time measured in years '*contributing*' information, enters the model as a predictor variable. By assumption, since X's are known they have individual have unknown multiplicative constants called **parameters** which affect the performance of the model itself. Normally these are denoted by $\beta_0, \beta_1, \dots, \beta_n$. For example,$Y = \beta_0 + \beta_1 X$. Thus, the model is said to be *linear in parameters* as explained by Rawlings et al. (1998).Lacey states that the objective is to find the best estimates for β_0 and β_1 by minimizing the residual error between the observed and fitted value .

Mathematically, $E(Y_i) = \beta_0 + \beta_1(X_i) \Rightarrow E(Temperature) = \beta_0 + \beta_1(Y_i)$ also known as the fitted regression line β_0 = y-intercept at time t=0, β_1 = the slope of line, with a unit of time as measured in years from 1970. But, what about the absolute value of Y_i?. It is defined as :

$$Y_i = \beta_0 + \beta_1(X_i) + \varepsilon \Rightarrow Temperature_i = \beta_0 + \beta_1(T_i) + \varepsilon_i$$

These are the implications of the assumptions mentioned above:

1. The ε_i of an observation i which also forms the deviation between observed and fitted observation for x_i have mean, $\mu = 0$ and have the same standard deviation σ.
2. All the ε_i's are independently & identically distributed.
3. Thus, $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ as explained by Rawlings et al. (1998)

Violation of the above stated assumptions can distort the results of the hypothesis testing conducted on linear models but not the parametric estimates themselves as explained by Rawlings et al. (1998). All the tests conducted including t-tests, F-tests require the data set's (i.e. the error term's) distribution to be normally distributed. In such a scenario F-tests are considered to be rather robust to changes in normality Hill & Lewicki (2006). The method I use to investigate normality are described in later sections as well as a technique known as transformation of variables is explained in the following section. The bulk of both location's data started from 1973 and so the linear model output give intercept at time t=0 which is thus calculated as the January 1, 1970.

2.1.1 Harmonic Models

Often data modeled exhibits periodic behavior which repeats itself every s time periods and the frequency, $\omega_i = 2\pi/s_i$,

1. s = 4 quarterly data
2. s = 12 monthly data
3. s = 365.224 daily data (applicable to this scenario as the available data's frequency was daily)
4. i = 1,2,3 ... n is the explanatory variable (i=t= number of days since 1 January, 1970)

The basic idea being that however the raw data is constructed be it hourly, daily, bi-monthly are converted to a to a fraction of the interval of periodicity *eg.* a year in this case, then multiplying the result by 2π. The general equation then becomes:

$$Y_i = \beta_0 + \beta_1(X) + \beta_2 \sin(\omega_i) + \beta_3 \cos(\omega_i) + \varepsilon_i$$

as explained by Cox (2006),Rawlings et al. (1998). As for the form applicable to the weather data:

$$Y_t = \beta_0 + \beta_1(Years) + \beta_2 \sin(\frac{2\pi t}{365.224}) + \beta_3 \cos(\frac{2\pi t}{365.224}) + \varepsilon_t$$

Like any other model, harmonic models have their own theoretical shortcomings *eg.* years of unequal lengths i.e. leap year may create problems over extremely lengthy time periods such as 1000 years.

2.1.2 Transformation of variables

Linear modelling is based on the assumptions discussed in section 2.1. One of the main problems with real-life data is that often the data in general, its error terms violate the assumptions. Therefore, in the related linear models there is a prevalence of non-normality associated with the distribution of the error terms. Following the consequences of non-normality of a model's error terms I transform the predictor variables in hopes of achieving a better result.

2.2 Hypothesis Tests

In this section I will briefly discuss the main tests conducted to check the significance of results of the linear models including F-tests, t-tests their associated p-values studentized Breusch-Pagan test, [1]

1. F-test indicates in a simple linear model whether X_i (time in years) explains a significant proportion of the variance observed in $Y_i(temperature) \Rightarrow H_a : \beta_1 \neq 0$ compared to the null hypothesis that the fitted model is largely dependent only on the intercept i.e. β_0, where $H_0 : \beta_1$=0 Technically what it is aims to explain is whether

$$F = \frac{(\text{TSS} - \text{RSS})/(k-1)}{\text{RSS}/(n-k)}$$

, where, TSS= total sum of squares RSS = residual sum of squares k= number of parameters in the model n= number of observations in the model An F-test statistic for a multiple regression model such as the one involving harmonic regression is testing to check whether **all** predictors are i.e. $\beta_1, \beta_2, \ldots \beta_n$ are significantly contributing to the independent variable. Empirically, if the calculated F-value has a p-value which is less than the a given value of α=0.05, then, the null hypothesis is rejected. In this case, it has the following general interpretation for the liner models I constructed:

[1] Basic knowledge of the above-mentioned tests' underlying theory is assumed, otherwise the reader is advised to consult any guide on mathematical statistics

- Simple Linear models: time measured in years has contributed to X degrees per year changes in temperature over a period of 1 year in comparison to the null hypothesis where β_1=0
- Harmonic model: Time as well as its transformation as mentioned in section 2.1.1 have contributed to changes in temperature over a cyclical period of a year in comparison to model where all the parameters are zero.

2. T-tests in this case investigate whether individual parameters calculated in the model are not significantly different from 0 versus the alternative that they are (on an individual basis) significant i.e. $\beta_0 \neq 0, \beta_1 \neq 0, \ldots \beta_n \neq 0$. They use the p-values and $\alpha = 0.05$ in a fashion similar to the F-tests described above.

2.2.1 Test for constant variance

The test that I use is known as Breusch Pagan(B-P) test and it tests whether a model's residuals have the same variance as defined by the linear model assumption versus the alternative hypothesis that they are not.

Mathematically,

$$H_0 = \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$$

, which are homoscedastic models

$$H_a = \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2 \cdots \neq \sigma_n^2$$

which are heteroscedastic models

- Again a p-value, calculated if less than the normal value of $\alpha = 0.05$. The null hypothesis is rejected.
- By construction if the test is being carried out on a basic model with just β_0 i.e. the intercept at time t=0 then the test results are invalid.

2.3 Checking for underlying assumptions for normality

I use normal Q-Q plot for each model constructed, to 'judge' the normality of the underlying models. The interpretation is as follows: the closer the data points are to the 45° line, the closer is the model's underlying distribution to normal distribution

2.4 Comparing models

Akaike's An Information Criterion (AIC) is a common measure of comparing models, it is calculated in the following method: AIC = 2K - 2 ln (L), where K is the number of parameters in the model and L is the maximum likelihood value of the model. Lowest value of AIC amongst the models gives the best answer as explained by R Development Core Team (2010).

3. Results

In this section of the report I will discuss the findings of the exploratory data analysis as well as the technical output related to both the locations using a combination of plots as well as statistical tests conducted on the data sets.

3.1 Exploratory Data Analysis

3.1.1 Five number summary

Location	Min	Max	1st Q	Median	3rd Q	Mean	Standard Deviation
Lahore	7.00	41.00	20.00	24.02	29.00	24.04	6.56
Karachi	10.00	39.00	24.00	27.00	29.00	26.08	4.17

Lahore has a larger range than Karachi, as it is hotter in summers and colder in winters than Karachi on average. In my opinion, the empirical results tend to underestimate the mean of Lahore primarily due to the imputation of mean for a good proportion of the time studied.

3.1.2 Average Plots

In order to give a general view the following plots were constructed whereby the means of both the seasons i.e. winter and summers were calculated as well degree 2, polynomial splines were added to the plots for individual seasons.

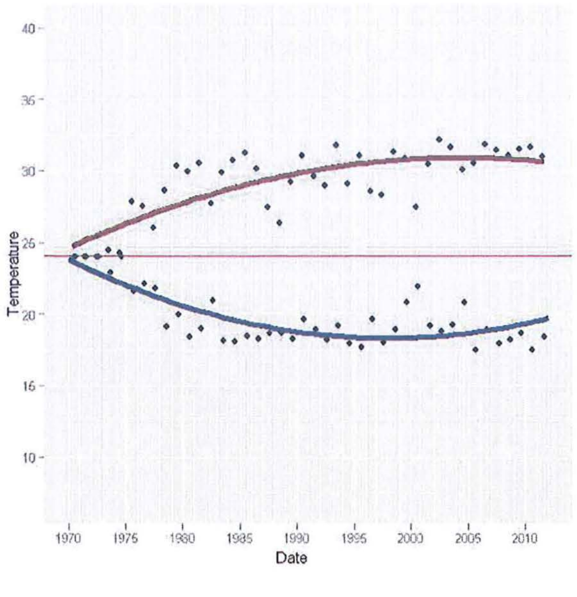


Figure 3.1: Lahore

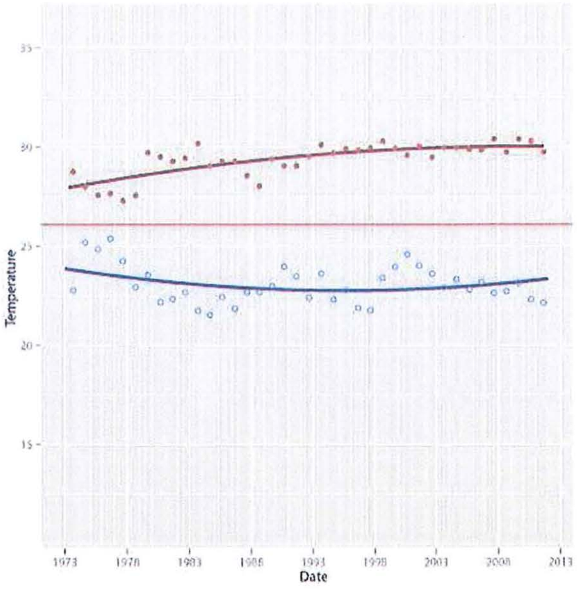


Figure 3.2: Karachi

3.1.3 Density Plots

Density plots of absolute temperature data for both the locations are shown. The aim in this section is to show the plots visually to demonstrate the underlying distribution of the largest data set for each location, however, each respective location's underlying model distribution is dependent on the distribution of its error terms of the specific model being studied as discussed in section 2.1. Lahore's plot 3.3 has a head and shoulder pattern due to three factors: Lahore had about three years long imputed data utilising the mean value of 24.02° which forms the head, the shoulders form the other two peaks on either ends occurring at 14.23°, 31.2° respectively corresponding to the hot and cold seasons. In comparison, 3.4 Karachi's density plot is a rather smooth and has two distinct peaks occurring at twenty degrees, twenty -seven degrees for summers and winters respectively. The aforementioned reasoning will become clearer in the next section.

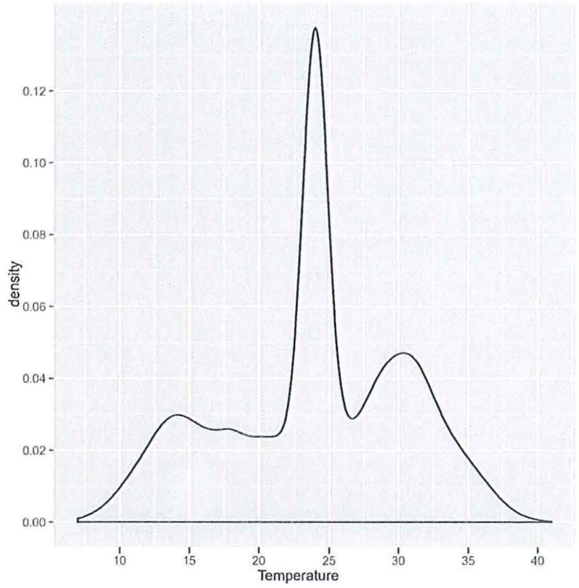


Figure 3.3: Lahore

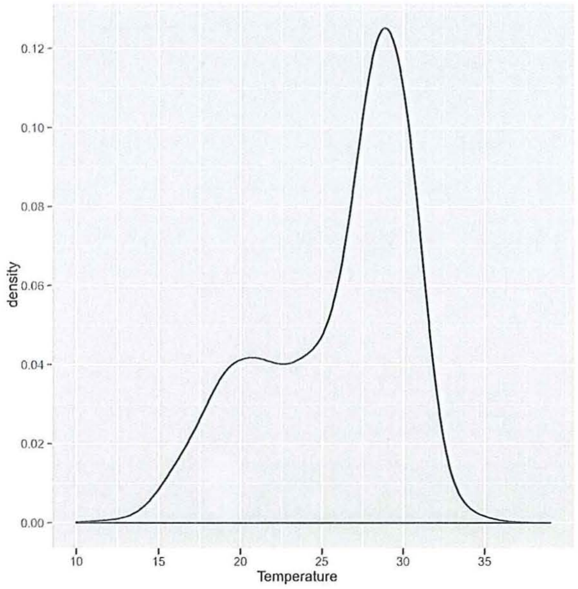


Figure 3.4: Karachi

3.1.4 Summer and Winter Comparative plots

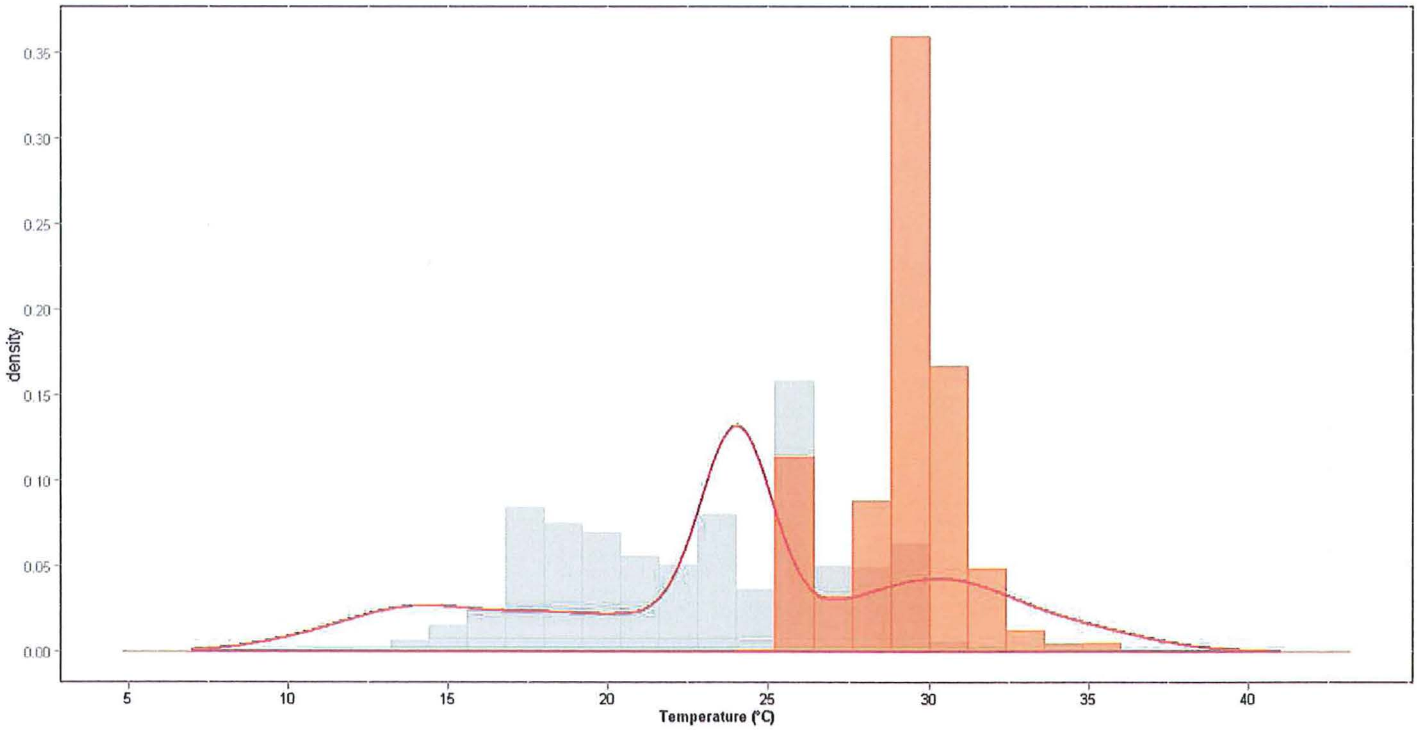


Figure 3.5: Summer-Winter histograms overlaid with the density plot for Lahore

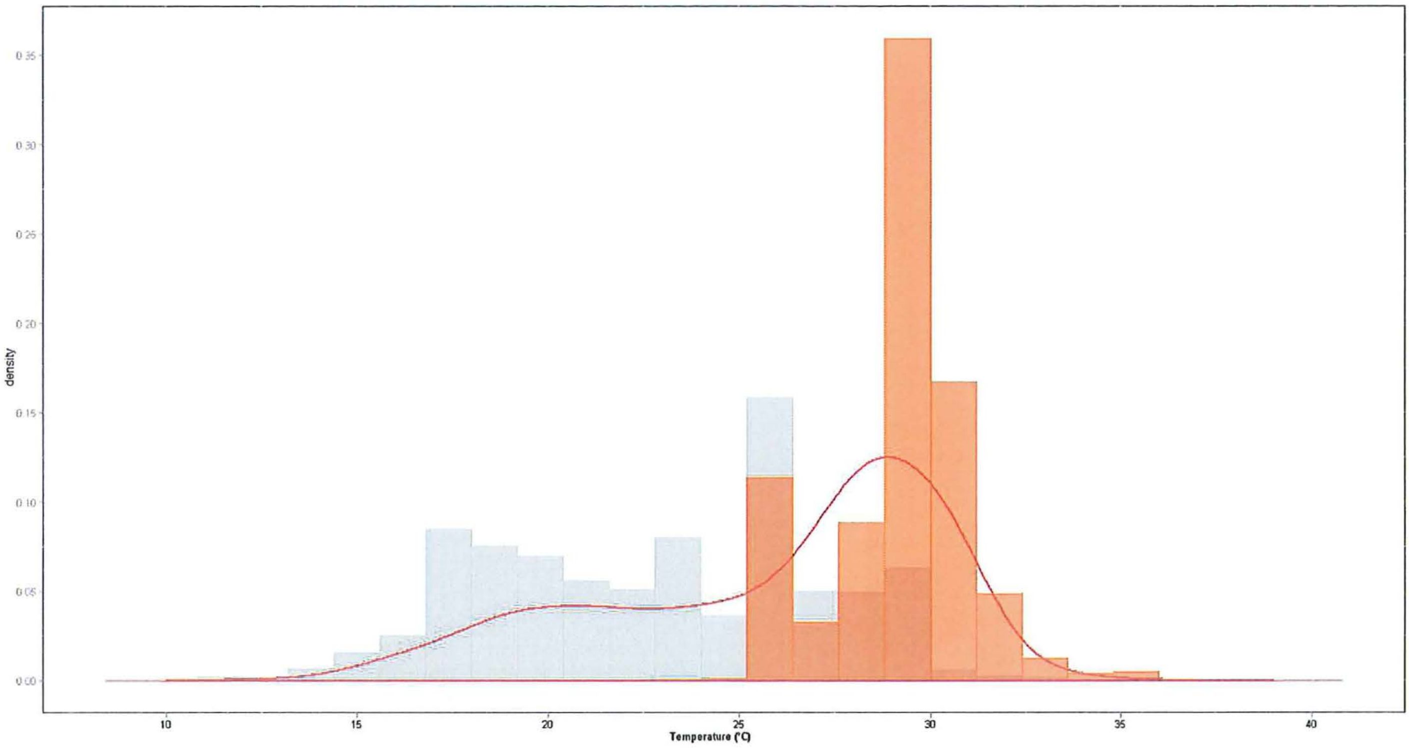


Figure 3.6: Summer-Winter histograms overlaid with the density plot for Karachi

3.1.5 Normal Q-Q plots

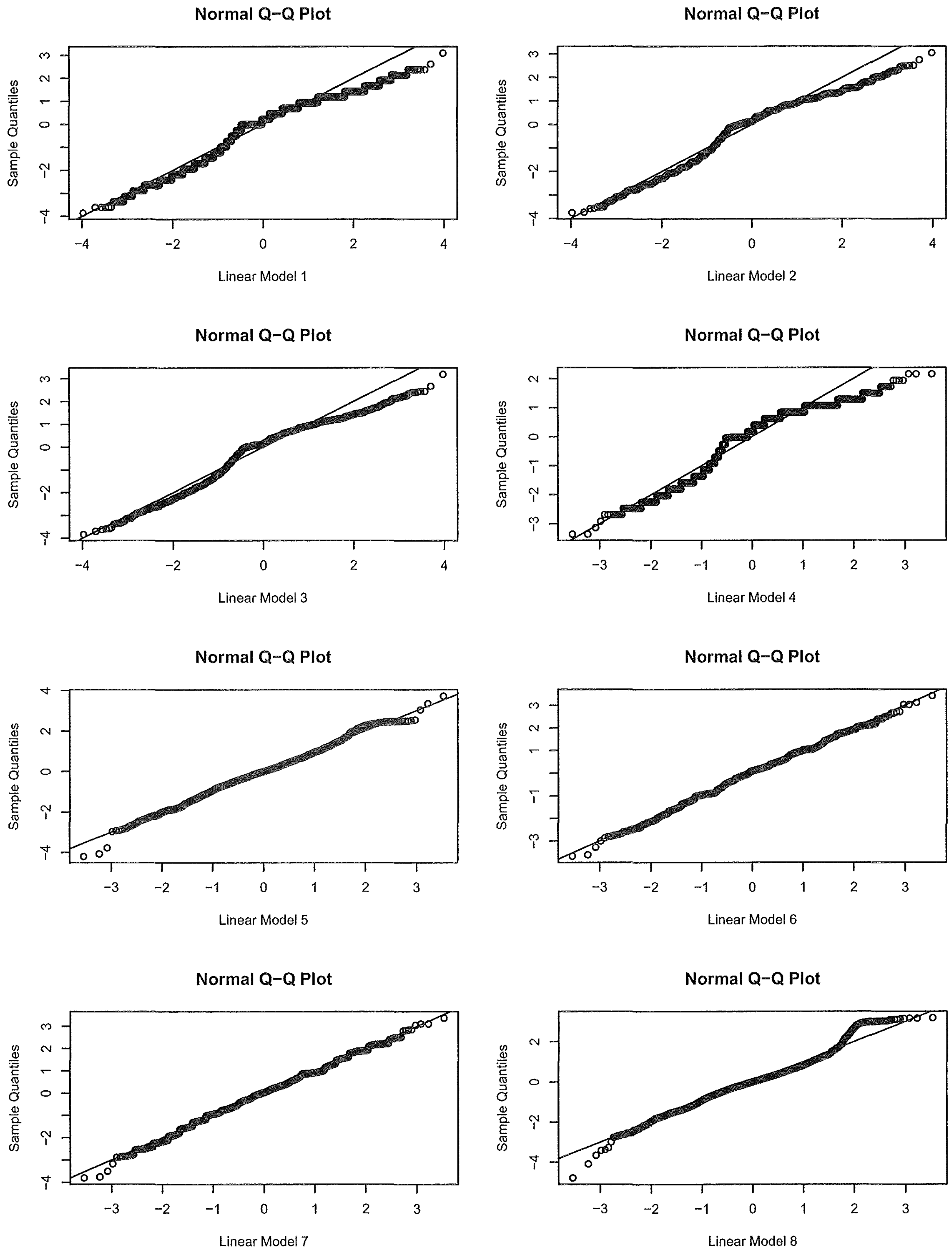


Figure 3.7: Karachi

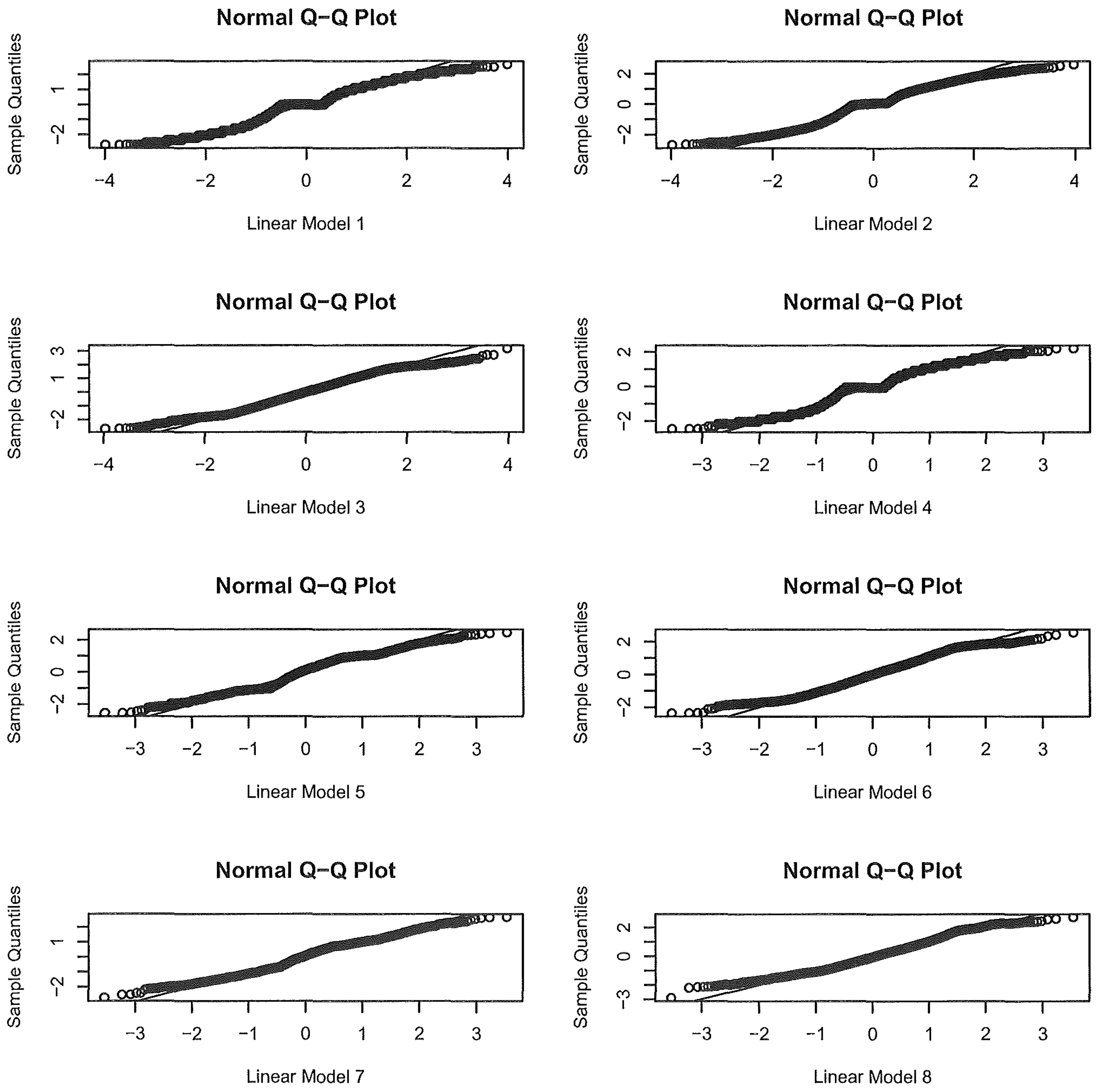


Figure 3.8: Lahore

3.2 Linear Models

There are in total 16 models with 8 models for each location. They are discussed in the following order:

1. Complete data model with no predictors

2. Complete data model only with years as a predictor - simple model

3. Complete data model with years, sine and cosine terms as predictors- harmonic model

4. Extreme comparison models with no predictors

5. Extreme comparison model only with years as predictor and a dummy indicator variable (1 for summers and 0 for winters) - (additive terms)

6. Extreme comparison model with sine and cosine terms in addition to the previously defined variables - (additive terms)

7. Extreme comparison model only with years as predictor and a a dummy indicator variable (1 for summers and 0 for winters) - (multiplicative terms)

8. Extreme comparison model with sine and cosine terms in addition to the previously defined variables - (multiplicative terms)

3.2.1 Karachi's results

```
Residuals:
     Min       1Q   Median       3Q      Max
-16.0777  -2.0777   0.9223   2.9223  12.9223
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.07772    0.03498   745.5   <2e-16 ***
Residual standard error: 4.175 on 14244 degrees of freedom
```

This model's coefficient is significant i.e. β_0 is significantly different from zero as indicated by the p-value and at time, t=0 there is a temperature of 26.08°. The 'residual standard error' is a studentized version of the above given residuals from the model. These are used to check the equal variance assumption in B-P tests in all the 16 models constructed.

```
Residuals:
     Min       1Q   Median       3Q      Max
-15.9551  -2.3509   0.6047   3.0286  13.3207
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.579970   0.078030 327.824  < 2e-16 ***
Years        0.022121   0.003101   7.133 1.03e-12 ***
Residual standard error: 4.167 on 14243 degrees of freedom
Multiple R-squared: 0.00356,Adjusted R-squared: 0.00349
F-statistic: 50.88 on 1 and 14243 DF,  p-value: 1.029e-12
```

This model's coefficient is significant and i.e. β_0 is significantly different from zero as indicated by the p-value of the t-test and at time, t=0 in this model there is a temperature of 25.58°. In addition, the slope of the years component is contributing on average a 0.02 degree per year to model. The model on the whole is considered to be significant as indicated by the p-value (for the F-test) in the last line. In this case only a mere 0.36% of the model's variation can be attributed to the time (i.e. year variable).

```
Residuals:
     Min       1Q   Median       3Q      Max
-15.5871  -2.5276   0.5216   3.1691  12.6806
Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  135.1261    14.2494   9.483  < 2e-16 ***
Years         -4.4624     0.6267  -7.121 1.12e-12 ***
SINE         240.7968    34.1218   7.057 1.78e-12 ***
COSINE      -107.6734    14.0412  -7.668 1.85e-14 ***
Residual standard error: 4.158 on 14241 degrees of freedom
Multiple R-squared: 0.008358,Adjusted R-squared: 0.008149
F-statistic: 40.01 on 3 and 14241 DF,  p-value: < 2.2e-16
```

According to harmonic regression this model's coefficient is significant and has a β_0 is significantly different from zero as indicated by the p-value of the t-test and at time, t=0 in this model there is a temperature of 135.13°. In addition, the slope of the years component is contributing on average a -4.46 degree per year to model. The model on the whole is considered to be significant in comparison to a model where there are no predictors as indicated by the p-value in the last line. The sine and cosine terms have coefficients of 240.8 and -107.67 respectively. This contribution's significance can be '*judged* ' by a major increase in R-squared statistics in comparison to say the last model. In this case 0.84% of the model's variation can be attributed to the range of predictors involved in this harmonic regression model.

```
Residuals:
     Min       1Q   Median       3Q      Max
-15.1961  -3.1961   0.8039   3.8039   9.8039
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.1961     0.0927   282.6   <2e-16 ***
Residual standard error: 4.54 on 2398 degrees of freedom
```

This model's coefficient is significant and i.e. β_0 is significantly different from zero as indicated by the p-value and at time, t=0 there is a temperature of 26.20°.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              18.777897   0.247570  75.849  < 2e-16 ***
as.factor(Indicator)[T.1] 10.248404   0.201234  50.928  < 2e-16 ***
Years                     0.024273   0.008935   2.717  0.00664 **
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 4.928 on 2396 degrees of freedom
Multiple R-squared: 0.5203,Adjusted R-squared: 0.5199
F-statistic:  1300 on 2 and 2396 DF,  p-value: < 2.2e-16
```

Basic interpretation: All the coefficients are significant as indicated by the model, however, this time round the interpretation is that at that the years component does contribute around 0.02 degrees per year however, the indicator function i.e. contributes 10 degrees in the summers. The model as a whole is significant as shown by the p-value. 52.03% of the variation in temperature is explained by the above given model's predictors.

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.0916  -1.4522  -0.0148   1.4445   8.9476
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               29.109111   0.228865 127.189  < 2e-16 ***
as.factor(Indicator)[T.1] -7.713643   0.400254 -19.272  < 2e-16 ***
Years                      0.020098   0.004381   4.588 4.72e-06 ***
SINE                       0.065938   0.110808   0.595    0.552
COSINE                    -9.114813   0.235788 -38.657  < 2e-16 ***
Residual standard error: 2.416 on 2394 degrees of freedom
Multiple R-squared: 0.7173,Adjusted R-squared: 0.7169
F-statistic:  1519 on 4 and 2394 DF,  p-value: < 2.2e-16
```

Basic interpretation: All the coefficients except the sine term's are significant as indicated by their individual p-values corresponding to their t-statistics. The sine term is insignificant as it has a p-value greater than α of 0.05. At time t=0 the intercept become 29.12 degrees. The year component's interpretation is around 0.02 degrees per year however, the indicator variable i.e. contributes -7.7 degrees in the summers on average. The model as a whole is significant as shown by the p-value. 71.73% of the variation in temperature is explained by the above given model's predictors.

```
Residuals:
    Min      1Q  Median      3Q     Max
-12.239  -2.227   0.086   2.189  10.871
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     23.329834   0.209827 111.186  < 2e-16 ***
as.factor(Indicator)[T.1]        4.808953   0.296031  16.245  < 2e-16 ***
Years                           -0.012777   0.008287  -1.542    0.123
as.factor(Indicator)[T.1]:Years  0.066809   0.011728   5.696 1.37e-08 ***
Residual standard error: 3.234 on 2395 degrees of freedom
Multiple R-squared: 0.4932,Adjusted R-squared: 0.4926
F-statistic:   777 on 3 and 2395 DF,  p-value: < 2.2e-16
```

The model itself is significant so we reject the null hypothesis. The coefficients of the model are all significant as indicated by their p-values, except the year's component as it has a p-value greater than 0.05. The R^2 is now 49.32% so approximately 50 percent of the model's variation is explained by its predictor variables. Temperature at t=0 is at 23.33 degrees.

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.1691  -1.1997  -0.0045   1.1458   6.7885
Coefficients:
```

```
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      27.464888   0.356743  76.988  < 2e-16 ***
as.factor(Indicator)[T.1]        -1.858218   0.505581  -3.675 0.000243 ***
Years                            -0.015147   0.005485  -2.762 0.005795 **
SINE                             -2.646601   0.223691 -11.831  < 2e-16 ***
COSINE                           -7.257198   0.351940 -20.621  < 2e-16 ***
as.factor(Indicator)[T.1]:Years   0.070226   0.007763   9.047  < 2e-16 ***
as.factor(Indicator)[T.1]:SINE    3.973578   0.250801  15.844  < 2e-16 ***
as.factor(Indicator)[T.1]:COSINE 4.232484   0.533725   7.930 3.33e-15 ***
Residual standard error: 2.14 on 2391 degrees of freedom
Multiple R-squared: 0.7784,Adjusted R-squared: 0.7778
F-statistic:  1200 on 7 and 2391 DF,  p-value: < 2.2e-16
```

The model itself is significant so the null hypothesis is rejected. The coefficients of the model are all significant as indicated by their p-values, the R squared value is now 77.84% so approximately 78 percent of the model's variation is explained by its predictor variables. Temperature at t=0 is at 27.46 degrees.

3.2.2 F - test for linear model comparison

```
Model 1: Temperature ~ Years
Model 2: Temperature ~ Years + SINE + COSINE
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  14243 247372
2  14241 246180  2    1191.2 34.453 1.183e-15 ***

Model 1: Temperature ~ as.factor(Indicator) * Years
Model 2: Temperature ~ as.factor(Indicator) * (Years + SINE + COSINE)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   2395 25053
2   2391 10953  4     14100 769.48 < 2.2e-16 ***

Model 1: Temperature ~ as.factor(Indicator) + Years + SINE + COSINE
Model 2: Temperature ~ as.factor(Indicator) * (Years + SINE + COSINE)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   2394 13974
2   2391 10953  3    3020.6 219.79 < 2.2e-16 ***

Model 1: Temperature ~ 1
Model 2: Temperature ~ as.factor(Indicator) * (Years + SINE + COSINE)
 Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2398 49436
2   2391 10953  7     38483 1200.1 < 2.2e-16 ***

Model 1: Temperature ~ as.factor(Indicator) + Years
Model 2: Temperature ~ as.factor(Indicator) * Years
```

```
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   2396 25392
2   2395 25053  1    339.44 32.450 1.373e-08 ****

Model 1: Temperature ~ 1
Model 2: Temperature ~ as.factor(Indicator) * Years
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2398 49436
2   2395 25053  3     24383 776.97 < 2.2e-16 ***

Model 1: Temperature ~ as.factor(Indicator) + Years
Model 2: Temperature ~ as.factor(Indicator) + Years + SINE + COSINE
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   2396 25392
2   2394 13974  2     11419 978.14 < 2.2e-16 ***

Model 1: Temperature ~ 1
Model 2: Temperature ~ as.factor(Indicator) + Years + SINE + COSINE
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2398 121307
2   2394  42837  4     78470 1096.4 < 2.2e-16 ***
```

All the above model tests show that adding predictors lead to a rejection of the null hypothesis and the models in each case are said to be significant in comparison to the first one which shows that increasing the number of variables has lead to a better model.

3.2.3 Lahore's Results

```
Residuals:
    Min      1Q  Median      3Q     Max
-17.036  -3.036  -0.019   4.964  16.964
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.03630    0.05103   471.1   <2e-16 ***
Residual standard error: 6.32 on 15338 degrees of freedom
```

This model's coefficient is significant and i.e. β_0 is significantly different from zero as indicated by the p-value and at time, t=0 there is a temperature of 24.04°.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.396994   0.122662 190.744  < 2e-16 ***
Years        0.028481   0.004876   5.841 5.29e-09 ***
Residual standard error: 6.551 on 14241 degrees of freedom
Multiple R-squared: 0.00239,Adjusted R-squared: 0.00232
F-statistic: 34.12 on 1 and 14241 DF,  p-value: 5.29e-09
```

This model's coefficient is significant and i.e. β_0 is significantly different from zero as indicated by the p-value of the t-test and at time, t=0 in this model there is a temperature of 23.40°. In addition, the slope of the years component is contributing on average a 0.03 degrees per year to model. The model on the whole is considered to be significant as indicated by the p-value in the last line. In this case only a mere 0.24% of the model's variation can be attributed to the time (i.e. year variable).

```
Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) 23.42390    0.07447  314.535   <2e-16 ***
Years        0.02727    0.00296    9.211   <2e-16 ***
SINE        -0.89793    0.04713  -19.053   <2e-16 ***
COSINE      -7.30699    0.04712 -155.065   <2e-16 ***
Residual standard error: 3.976 on 14239 degrees of freedom
Multiple R-squared: 0.6324,Adjusted R-squared: 0.6324
F-statistic:  8167 on 3 and 14239 DF,  p-value: < 2.2e-1
```

This model's coefficient is significant and i.e. β_0 is significantly different from zero as indicated by the p-value of the t-test and at time, t=0 in this model there is a temperature of 23.42°. In addition, the slope of the years component is contributing on average a 0.03 degree per year to model. The model on the whole is considered to be significant in comparison to a model where there are no predictors as indicated by the p-value in the last line. Residual standard error again are used in the aforementioned model to test for homoscedasticity. The sine and cosine terms have coefficients of -.89 and -7.31 respectively. This contribution's significance can be '*judged* ' by a major increase in R-squared statistics in comparison to say the last model. In this case of 63.24% of the model's variation can be attributed to all the predictors involved in this harmonic regression model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-17.448  -4.448  -0.431   5.552  15.552
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.4483     0.1452   168.4   <2e-16 ***
Residual standard error: 7.112 on 2398 degrees of freedom
```

This model's coefficient is significant and i.e. β_0 is significantly different from zero as indicated by the p-value and at time, t=0 there is a temperature of 24.45°.

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                18.777897   0.247570  75.849  < 2e-16 ***
as.factor(Indicator)[T.1] 10.248404   0.201234  50.928  < 2e-16 ***
Years                       0.024273   0.008935   2.717  0.00664 **
Residual standard error: 4.928 on 2396 degrees of freedom
Multiple R-squared: 0.5203,Adjusted R-squared: 0.5199
F-statistic:  1300 on 2 and 2396 DF,  p-value: < 2.2e-16
```

Again all the coefficients are significant as indicated by the model, however, this time round the interpretation is that at that the years component does contribute around 0.02 degrees per year however, the indicator function i.e. contributes 10 degrees in the summers. The model as a whole is significant as shown by the p-value. 50.23% of the variation in temperature is explained by the above given model's predictors.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     20.78680    0.31313  66.384  < 2e-16 ***
as.factor(Indicator)[T.1]        6.23964    0.44177  14.124  < 2e-16 ***
Years                           -0.06432    0.01237  -5.201 2.15e-07 ***
as.factor(Indicator)[T.1]:Years  0.17746    0.01750  10.139  < 2e-16 ***
Residual standard error: 4.827 on 2395 degrees of freedom
Multiple R-squared: 0.5401,Adjusted R-squared: 0.5395
F-statistic: 937.4 on 3 and 2395 DF,  p-value: < 2.2e-16
```

Basic interpretation: All the coefficients are significant as indicated by the model, however, this time round the interpretation is that at that the years component does contribute around -0.06 degrees per year however, the indicator function i.e. contributes 10 degrees in the summers. The model as a whole is significant as shown by the p-value. 54.01% of the variation in temperature is explained by the above given model's predictors.

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      27.07495    0.65377  41.413  < 2e-16 ***
as.factor(Indicator)[T.1]        -2.32551    0.92654  -2.510   0.0121 *
Years                            -0.06660    0.01005  -6.625 4.26e-11 ***
SINE                             -2.04467    0.40994  -4.988 6.55e-07 ***
COSINE                           -9.93205    0.64497 -15.399  < 2e-16 ***
as.factor(Indicator)[T.1]:Years   0.18104    0.01423  12.726  < 2e-16 ***
as.factor(Indicator)[T.1]:SINE    3.67568    0.45962   7.997 1.96e-15 ***
as.factor(Indicator)[T.1]:COSINE  7.22891    0.97812   7.391 2.01e-13 ***
Residual standard error: 3.922 on 2391 degrees of freedom
Multiple R-squared: 0.6968,Adjusted R-squared: 0.6959
F-statistic: 784.8 on 7 and 2391 DF,  p-value: < 2.2e-16
```

The model itself is significant so we reject the null hypothesis. The coefficients of the model are all significant as indicated by their p-values, the R squared value is now 69.68% so approximately 70 percent of the model's variation is explained by its predictor variables.

3.2.4 F - test for linear model comparison

```
Model 1: Temperature ~ Years
Model 2: Temperature ~ Years + SINE + COSINE
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1  14241 611105
2  14239 225153  2    385952 12204 < 2.2e-16 ***
```

```
Model 1: Temperature ~ as.factor(Indicator) * Years
Model 2: Temperature ~ as.factor(Indicator) * (Years + SINE + COSINE)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   2395 55793
2   2391 36786  4     19007 308.85 < 2.2e-16 ***

Model 1: Temperature ~ as.factor(Indicator) + Years + SINE + COSINE
Model 2: Temperature ~ as.factor(Indicator) * (Years + SINE + COSINE)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   2394 42837
2   2391 36786  3    6050.8 131.09 < 2.2e-16 ***

Model 1: Temperature ~ 1
Model 2: Temperature ~ as.factor(Indicator) * (Years + SINE + COSINE)
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1   2398 121307
2   2391  36786  7     84521 784.8 < 2.2e-16 ***

Model 1: Temperature ~ as.factor(Indicator) + Years
Model 2: Temperature ~ as.factor(Indicator) * Years
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   2396 58188
2   2395 55793  1    2394.9 102.80 < 2.2e-16 ***

Model 1: Temperature ~ 1
Model 2: Temperature ~ as.factor(Indicator) * Years
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2398 121307
2   2395  55793  3     65514 937.43 < 2.2e-16 ***

Model 1: Temperature ~ as.factor(Indicator) + Years
Model 2: Temperature ~ as.factor(Indicator) + Years + SINE + COSINE
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   2396 58188
2   2394 42837  2     15351 428.96 < 2.2e-16 ***

Model 1: Temperature ~ 1
Model 2: Temperature ~ as.factor(Indicator) + Years + SINE + COSINE
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2398 121307
2   2394  42837  4     78470 1096.4 < 2.2e-16 ***
```

All the above model tests show that adding predictors lead to a rejection of the null hypothesis and the models in each case are said to be significant in comparison to the first one in each of the above cases.

3.3 BreuschPagan and Akaike information criterion tests

Model	DF	B-P test	Karachi	Lahore
1	Not applicable/0	Not applicable	Not applicable	Not applicable
2	1	Heteroscedastic	6.74E-17	5.82E-171
3	3	Heteroscedastic	1.98E-45	2.05E-138
4	Not applicable/0	Not applicable	Not applicable	Not applicable
5	2	Heteroscedastic	7.91E-121	1.22E-37
6	4	Heteroscedastic	1.62E-87	3.92E-118
7	3	Heteroscedastic	1.28E-116	6.83E-44
8	7	Heteroscedastic	4.68E-100	1.78E-78

Model	DF	AIC **Karachi**	% Δ in AIC **within models**	AIC **Lahore**	% Δ in AIC **within models**
1	0	81,142.50		100,093.10	
2	1	81,028.94	-0.140%	93,965.36	-6.122%
3	3	81,093.70	0.080%	79,747.88	-15.131%
4	0	14,070.53		16,224.00	
5	2	11,047.40	-21.486%	14,465.68	-10.838%
6	4	12,476.26	12.934%	13,734.82	-5.052%
7	3	12,445.98	-0.243%	14,366.75	4.601%
8	7	10,469.12	-15.884%	13,375.50	-6.900%

Comments

The fact that model 8 in both the locations are the best model according to the AIC criterion can be attributed to the fact that it has the lowest likelihood function for both the locations and also had the highest number of variables. Both are heteroscedastic and naturally have the highest R^2 for their locations.

4. Conclusion and Discussion

In general all the model had significant components except one. All the models as shown by B-P tests in section indicate that the models were were not homoscedastic. This in itself does not make the analysis redundant. However, I would like to stress on the fact that further improvements such as adding a greater number of predictor variables as well as using better imputation techniques instead of pure mean imputation, with generally greater stochasticity could perhaps lead to a better analysis. There are many factors which have lead me to conclude that although significant the changes in temperature are miniscule they have no real direct impact on the human body however, I do agree with Wagner (1996) that weather events have become more extreme partly due to the main measured variable i.e. temperature as well as other factors such as inadequacy of rescue services, the lack of general infrastructure in the location of interest. In my viewpoint, the purpose of this project was to primarily explore R's computational capabilities and in the process also apply these to a weather data set. My own achievement at the moment remains limited to a better understanding of data analysis, the problems involved, as well as the practical application of such studies. I also believe that such projects can help academic institutions further develop potential industry related partnerships.

Also, there are a few areas where such reports are utilised both within and outside Pakistan specially in feasibility studies such as the one undertaken by the Employment and Pakistan SNS (2006). Moreover, such studies developed and carried out with a greater number of parameters can aid governments, their policy and decision makers in their decision making progress. At the grass-roots level such a project may aid in the planning of flood barriers and the like which can help save several lives as well as the prevent the direct and indirect economic losses associated with such events. For example, in 2010, flash floods Pakistan experienced amongst one of the most devastating floods in history, which displaced several rural area families and stretched the nation's fundamental armed resources. UN Secretary-General Ban Ki-moon described the flash floods as,

> *"... a slow-motion tsunami. Its destructive powers will accumulate and grow with time"*

Figure 4.1: 2010 Flash Floods, courtesy IMC UK

4.1 Future work

I believe this project has a lot of potential to be further developed into a better model. In the near future, I intend to grow and extend this model with a higher number of predictors such as station pressure, humidity levels, extreme weather events with a deeper theoretical knowledge perspective gained through future courses like Bayesian Inference, Time Series Analysis, Generalised Linear Models along with possible corporate collaboration with the insurance industry and the Department of Geology, University of Canterbury.

4.2 Research Experience

Although, this was my first research experience but I strongly believe it gave me an excellent insight into the research carried out at Canterbury as well as its practical implications for a research based career. Also this particular project allowed me to apply the skills learnt in theoretical courses such as Monte Carlo Methods, Statistical Inference and Introductory Statistics. Finally, it also also gave me the chance to experience the problems faced by real life data analysts related to missing data, violation of important assumptions underlying the linear models. I may now safely admit that I can use R and LaTeXfor data analysis reports.

Acknowledgements

I would like to specifically thank and acknowledge Dr. Elena Moltchanova for her guidance and encouragement for letting me enrol in this summer project and explaining every aspect of the project. Also, I would like to thank the Department of Mathematics and Statistics for providing essential facilities as well as the necessary funding opportunity without which this would not have been possible.

A. R Command List

Commands I utilised in R

library - to load external packages such as session (saves, retrieves R session history), date and chron (used for date objects) , ggplot2 (graphics)

fix - manipulation of a data frame

attach - makes data frame objects easily available

detach - opposite of attach

rm - removes an item from the R Session

ls - lists R Session's objects

search - loads current list of R packages in use

names - provides name of an R object

as.numeric - turns an item into a numeric class

substr - used for extracting information from a vector

unclass, factor, length, with, by, list, data.frame, lapply, matrix, unlist, aggregate,rbind - normally used in combination with other commands, for data manipulation **format, subset, seq, nrows, rownames**

class - checks the object's class,

round, as.character, as.integer - used for object manipulation

na.omit - used in handling NaN values

mean, summary, range,anova, sd - statistical functions

as.Date, years, mdy.date, julian - handling date objects

value - extracts values

lm - to construct linear models

B. R Code

```
###########Laying out Lahore’s complete timeline - to calculate the #############
######################### total number of NaN values ############################
start=as.Date("1957-07-01")
end=as.Date("2011-12-30")
timeline <- seq(start1,end1,by=’1 day’)
length(wlhe$Month)-length(timeline)

########### building of data frames for actual analysis ########################
wlhe_main <- data.frame(Date=as.Date(wlhe$Proper_Date),value=as.integer(wlhe$Temperature))
wlhe_main_1 <- data.frame(Date=as.Date(timeline), Temperature= with(wlhe, value[match(timeline,
Date)]))

#################### replacing missing values with mean ########################
wlhe_main_1$Temperature[is.na(wlhe4$Temperature)] <- mean(wlhe_main_1$Temperature, na.rm = TRUE)

#################### further data set amendments ###############################
############# to convert julian dates to no. of years and calculate#############
######### sine and cosine components for Harmonic regression ###################
rownames(wlhe_main_1) <- seq(length=nrow(wlhe_main_1))
wlhe_main_1$Day<- month.day.year(unclass(wlhe_main_1$Date))$day
wlhe_main_1$Year<- month.day.year(unclass(wlhe_main_1$Date))$year
wlhe_main_1$Month <- month.day.year(unclass(wlhe_main_1$Date))$month
wlhe_main_1$Years <- julian(wlhe_main_1$Month, wlhe_main_1$Day, wlhe_main_1$Year)/365.224
wlhe_main_1$SINE<- with(wlhe_main_1, sin((2*pi*wlhe_main_1$Years)))
wlhe_main_1$COSINE<- with(wlhe_main_1, cos((2*pi*wlhe_main_1$Years)))

####################### subsetting the main hot and cold - lahore ###########################
hot_lhe <- subset(wlhe_main_1, Month==c(5:8))
rownames(hot_lhe) <- seq(length=nrow(hot_lhe))
cold_lhe <- subset(wlhe_main_1, Month==c(10,11,12,1))
rownames(cold_lhe) <- seq(length=nrow(cold_lhe))
cold_lhe$Indicator<- with(cold_lhe, 0)
hot_lhe$Indicator<- with(hot_lhe, 1)

######### extreme comparison data set with only the summers and winters ########
ecomp_lhe<- rbind(hot_lhe,cold_lhe)
```

```
########################### mean value curves ##############################################
###################### for summer and winters alike ############################
mean_cold_lhe <-data.frame(mean_dates= aggregate(x=as.numeric(as.Date(cold_lhe$Date)),
by = list(factor(cold_lhe$Year)), FUN = "mean"),mean_years = aggregate(x =cold_lhe$Temperature,
by = list(factor(cold_lhe$Year)), FUN = "mean"))
mean_cold_lhe$Date<- with(mean_cold_lhe,as.Date(chron(round(mean_cold_lhe$mean_dates.x)),
format="%m/%d/%y"))


mean_hot_lhe <-data.frame(mean_dates= aggregate(x=as.numeric(as.Date(hot_lhe$Date)),
by = list(factor(hot_lhe$Year)), FUN = "mean"),mean_years = aggregate(x =hot_lhe$Temperature,
by = list(factor(hot_lhe$Year)), FUN = "mean"))
mean_hot_lhe$Date<- with(mean_hot_lhe,as.Date(chron(round(mean_hot_lhe$mean_dates.x)),
format="%m/%d/%y"))

########################################## plots##################################
plot_comp_lhe <- ggplot(wlhe_main_1, aes(as.Date(Date), Temperature)) +
geom_smooth(data=cold_lhe, method="lm", formula=y~poly(x, 2),colour="dark blue", se=F, size=2) +
geom_smooth(data=hot_lhe, method="lm", formula=y~poly(x, 2),colour="dark red", se=F, size=2)  +
geom_point(data=mean_cold_lhe, colour="dark red") +
geom_point(data=mean_hot_lhe) + opts(title = "Lahore Temperature Outlook");plot9_lhe

plot_density_lhe<- qplot(wlhe_main_1$Temperature, geom="density",
main="Lahore Temperature Outlook") + xlab("Temperature");plot10_lhe #
###############################################################################
###the density plot ''geom'' makes the assumption that the data is unbounded , #
#########continuous and smooth which it may or may not be in practice#########


######################  comparative plot #######################################
histogram_lhe<- ggplot(wlhe_main_1, aes(Temperature)) +
geom_histogram(aes(y=..density..), data = cold_lhe, fill="lightcyan3", colour = "lightcyan3",
alpha = 0.8, binwidth=1.2) +
geom_histogram(aes(y=..density..), data = hot_lhe, fill="orangered",colour = "orangered",
alpha = 0.6, binwidth=1.2) +
opts(panel.background=theme_rect(fill="NA", colour="black"))

######## comparative histogram overlaid by a density plot #####################
histogram_lhe + stat_density(aes(y=..density..), adjust=1.4, data=wlhe_main_1, fill="NA",
colour="red", size=1.05) + xlab("Temperature (C)") + opts(axis.text.x=theme_text(size=10))+
opts(axis.title.x=theme_text(size=10, face="bold"))
```

```
################################ Normal Q-Q plots ##############################
par(mfrow=c(2,2))
qqnorm(studres(m1_l), xlab="Linear Model 1") + abline(0,1)
qqnorm(studres(m2_l), xlab="Linear Model 2") + abline(0,1)
qqnorm(studres(m3_l), xlab="Linear Model 3") + abline(0,1)
qqnorm(studres(m4_l), xlab="Linear Model 4") + abline(0,1)
par(mfrow=c(2,2))
qqnorm(studres(m5_l), xlab="Linear Model 5") + abline(0,1)
qqnorm(studres(m6_l), xlab="Linear Model 6") + abline(0,1)
qqnorm(studres(m7_l), xlab="Linear Model 7") + abline(0,1)
qqnorm(studres(m8_l), xlab="Linear Model 8") + abline(0,1)

############################whole data models###################################
summary(m3_l<- lm(Temperature~1, data=wlhe_main_1))
summary(m2_l<- lm(Temperature~ Years + SINE + COSINE, data=wlhe_main_1))
summary(m1_l<- lm(Temperature~Years, data=wlhe_main_1))


####################   winter/summer extremes models  ##########################
summary(m4_l<- lm(Temperature~as.factor(Indicator)+Years + SINE + COSINE, data=ecomp_lhe))
summary(m5_l<- lm(Temperature ~as.factor(Indicator)+Years,data = ecomp_lhe))
summary(m6_l<- lm(Temperature~1, data=ecomp_lhe))


############################## multiplication #################################
summary(m7_l<- lm(Temperature~as.factor(Indicator)*(Years+ SINE + COSINE), data=ecomp_lhe))
summary(m8_l<- lm(Temperature ~as.factor(Indicator)* Years,data = ecomp_lhe))



############################# F-tests to compare the plots#####################
anova(m1_l,m2_l)
anova(m8_l,m7_l)
anova(m4_l,m7_l)
anova(m6_l,m7_l)
anova(m5_l,m8_l)
anova(m6_l,m8_l)
anova(m5_l,m4_l)
anova(m6_l,m4_l)
```

Bibliography

(2006). Fish farming inland costal, production, processing and marketing.

(2009). Solstice. Britannica.

Cox, N. J. (2006). Speaking stata: In praise of trigonometric predictors. *Stata Journal*, 6(4):564–567.

Hill, T. and Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining.*

Krohn, S. (2000). Danish wind turbines: an industrial success story. *Danish Wind Industry Association*, 21.

Lacey, M. Inference in linear regression.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (1998). *Applied Regression Analysis.* Springer New York.

Scheffer, J. (2002). Dealing with missing data. *Research letters in the information and mathematical sciences*, 3(1):153–160.

Wagner, D. (1996). Scenarios of extreme temperature events. *Climatic Change*, 33:385–407. 10.1007/BF00142585.