SOP 90: Analyzing Climate Trends with R

Marc Los Huertos

July 10, 2017

Contents

1 Introduction **2**
- 1.1 Rational . . . 2
- 1.2 Trend Analysis versus Time Series Analysis . . . 3
- 1.3 Generalized Steps . . . 3
- 1.4 Goals for This Document . . . 3

2 Frequentists and Bayesian Statistics **4**
- 2.1 Bayesian Statistics . . . 4
- 2.2 Frequentist Statistics . . . 5
  - 2.2.1 Categorical versus Continuous Data . . . 5
  - 2.2.2 Four Frequentists Approaches . . . 6

3 Understanding the Data: The First Step **6**
- 3.1 Data Source and Metadata . . . 6
- 3.2 Evaluating the Structure of Data . . . 6
- 3.3 Evaluating for Completeness . . . 8
- 3.4 Evaluating the Central Tendencies . . . 8
- 3.5 Evaluating Spread . . . 10
- 3.6 'Null Hypotheses' as the Foundation of Frequentist Statistics . . . 11
- 3.7 Type I and Type II Errors . . . 11
- 3.8 Mathematical Mechanisms to Test Hypothesis . . . 11

4 Linear Models in R **12**

5 Simple Regression **13**
- 5.1 What is Linear Regression? . . . 13
- 5.2 Regression and Climate Change . . . 14
- 5.3 Creating Monthly Averages of Daily Maximum Temperatures . . . 17
- 5.4 Creating Monthly Means . . . 17
- 5.5 Testing all the Months . . . 21
  - 5.5.1 TMIN . . . 21
- 5.6 TMIN . . . 23
- 5.7 TMIN . . . 26

5.7.1 Departure from Mean 27
5.7.2 Experimental Portion — Precipitation 27
5.8 Problems with a Simple Regression Model 28
5.8.1 Model Diagnostics 28

6 The 'Null' Hypothesis versus Information Criteria 30
6.1 Model Comparison 30
6.2 AIC to make statements about strength of evidence 30

7 Relaxing Model Assumptions 30
7.1 Using Sources of Error in the Model 30
7.2 Generalized Least Square (GLS) and Autocorrelation 30
7.3 Adding Seasonality 31

8 More Sophisticated Approaches 31

9 Advanced Methods 31

10 Time Series Analysis 31

11 References 32

Abstract

Trend and Time Series Analyses are very important in environmental monitoring. For our purposes, developing methods to analyze climate data can using a range of tools, from relatively simple methods to advanced statistical modelling.

This document is designed to introduce a few tools to analyze regularly (i.e. daily or monthly) collected data. First, we will use a standard regression model, mixed-effects models, and finally more advanced time-series modelling approaches.

1 Introduction

1.1 Rational

In an age of industrialization and waste hazardous waste production, monitoring schemes have become ubiquitous to provide early warning, tracking environmental quality

To detect change or provide warning for public safety, we also need tools to assess if there are trends or if there are changes that might pose a hazard.

For our purposes, the trend analysis or time series analysis is used to evaluate the contested nature of climate change – in particular, to determine if there weather changes at a regional level.

1.2 Trend Analysis versus Time Series Analysis

Trend analysis statistics are a set of tools used to detect patterns of change exceed the relative variation of the system. To do this, statistical models try to partition the sources of variation – sources of an effect versus sources of random varation, for example.

Methods might includce linear regression, the Mann-Kendall Trend Test, seasonal Mann-Kendall Test, correlated seasonal Mann-Kendall Test, partial Mann-Kendall Trend test, (Seasonal) Sen's slope, partial correlation trend test and change-point test after Pettitt.

Alternatively, researchers might use time series analysis to evaluate the signal of the series by decomposing it for internal cycles, e.g. seasons and time of year, and determine if the signal is stationary or non-stationary, i.e. non-changing with time versus changing with time. A time series is a sequence of data points, measured typically at successive points in time spaced at uniform time intervals. Examples of time series are the daily closing value of the Dow Jones index and the annual flow volume of the Nile River at Aswan. Time series are very frequently plotted via line charts. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, and communications engineering.

1.3 Generalized Steps

Presentation of trend data

Presentations of time-trend data should usually include the following:

Graphical plots displaying the observed data over time Comment on any statistical methods used to transform the data Report average percent change An interpretation of the trends seen Interpretation of trend data

1.4 Goals for This Document

This document provides EA students with limited statstical training to use various statistical tools to analyze climate data. To accompiish this, we have tried to explain the theoretical background and demonstrate the steps to analyze climate data. We also try to explain how the "null" hypothesis is used in frequentist statistics, the meaning of type I and type II errors, and how we draw conclusions using statistical tools. Finally, we demonstrate the steps to use some of the common tools to analyze climate data using temperature and precipitation data collected from Bangkok, Thailand.

NOTE: I am not a statsitian. But I have written this because I have never found an adequate guide for undergraduates that is both accessible and complete enough for our project. However, no document is perfect.

First, this document is not complete. Second, there are areas of confusion for me and these probably come out in the document. Even when I understand the concepts and tools, my explanation may lead to confusion. In any case, I

suggest you use this as ONE resource and not the only one. In addition, if you are confused by sections, please let me know because I work hard to improve it with each iteration.

2 Frequentists and Bayesian Statistics

Statistics was born well before computer software had been developed, the development of statistical tools have both a historical context that is well beyond our project, but should be appreciated so we can use these tools without too much dificulty.

First, we need to know there is a distinction between the Frenquentist and Bayesian statitician. Although the tools and approach are quite distinct, few students appreciate how different these approaches are until they are graduate school – and being forced to 'pick' one versus the other in how they approach data analysis problems.

I will reverse the typical order in most texts and describe Bayesian statistics first because we can use this to explicitly talk about probability and then we'll shift to Freqentist statistics – which most of the tools in this text rely.

2.1 Bayesian Statistics

Bayesian statistics, named for Thomas Bayes (17011761), is a theory in the field of statistics in which the evidence about the true state of the world is expressed in terms of degrees of belief known as Bayesian probabilities.

Bayesian statistics is a system for describing epistemological uncertainty using the mathematical language of probability. In the 'Bayesian paradigm,' degrees of belief in states of nature are specified; these are non-negative, and the total belief in all states of nature is fixed to be one. Bayesian statistical methods start with existing 'prior' beliefs, and update these using data to give 'posterior' beliefs, which may be used as the basis for inferential decisions.

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference is an important technique in statistics, and especially in mathematical statistics. Bayesian updating is particularly important in the dynamic analysis of a sequence of data. Bayesian inference has found application in a wide range of activities, including science, engineering, philosophy, medicine, sport, and law.

To demonstrate Baysian approach, let's evaluate the distribution of temperatures from Bangkok, Thailand.

https://www.r-bloggers.com/a-simple-intro-to-bayesian-change-point-analysis/

http://www.flutterbys.com.au/stats/tut/tut7.2b.html

https://www.r-bloggers.com/bayesian-linear-regression-analysis-without-tears-r/

2.2 Frequentist Statistics

Frequentist inference has been associated with the frequentist interpretation of probability, specifically that any given experiment can be considered as one of an infinite sequence of possible repetitions of the same experiment, each capable of producing statistically independent results.[1] In this view, the frequentist inference approach to drawing conclusions from data is effectively to require that the correct conclusion should be drawn with a given (high) probability, among this notional set of repetitions. However, exactly the same procedures can be developed under a subtly different formulation. This is one where a pre-experiment point of view is taken. It can be argued that the design of an experiment should include, before undertaking the experiment, decisions about exactly what steps will be taken to reach a conclusion from the data yet to be obtained. These steps can be specified by the scientist so that there is a high probability of reaching a correct decision where, in this case, the probability relates to a yet to occur set of random events and hence does not rely on the frequency interpretation of probability. This formulation has been discussed by Neyman,[2] among others.

There are two major differences in the frequentist and Bayesian approaches to inference that are not included in the above consideration of the interpretation of probability:

In a frequentist approach to inference, unknown parameters are often, but not always, treated as having fixed but unknown values that are not capable of being treated as random variates in any sense, and hence there is no way that probabilities can be associated with them. In contrast, a Bayesian approach to inference does allow probabilities to be associated with unknown parameters, where these probabilities can sometimes have a frequency probability interpretation as well as a Bayesian one. The Bayesian approach allows these probabilities to have an interpretation as representing the scientist's belief that given values of the parameter are true [see Bayesian probability - Personal probabilities and objective methods for constructing priors]. While "probabilities" are involved in both approaches to inference, the probabilities are associated with different types of things. The result of a Bayesian approach can be a probability distribution for what is known about the parameters given the results of the experiment or study. The result of a frequentist approach is either a "true or false" conclusion from a significance test or a conclusion in the form that a given sample-derived confidence interval covers the true value: either of these conclusions has a given probability of being correct, where this probability has either a frequency probability interpretation or a pre-experiment interpretation.

2.2.1 Categorical versus Continuous Data

When we measure environmental data, they can be either continuous or categorical. Categorical data are sometimes called discrete data, e.g. count data might be considered discrete and categorical — if they are relatively number of possible values (perhaps, less than 3-4). Predictor variables can also be either

Table 1: 2 x 2 Matrix of Inference Methods –Going to insert graphics to give little pics of analysis...

		Response	
		Categorical	Continuous
Predictor	Categorical	Tests of Association 	ANOVA
	Continuous	Logistic Regression	Linear Regression

categorical or continuous — where we think of values that can be integers with a great range and values between integers (i.e. values with decimals).

Based on our understanding of the data, the choices of analysis become limited but also easier to choose from (Table

2.2.2 Four Frequentists Approaches

Table 1 describes the statistical approaches available to freqentists in relation to the characteristics of the data.

In our example, the temperature is obviously continuous. Date can be treated as continuous when their are lots of them, but as describe below, it might also be considered categorical – in part because it's 'ordered' and it's divisible by day – and not finer resolution, which is decidedly not continuous. However, because there are so many in these records, we can ignore that to continue our analysis.

3 Understanding the Data: The First Step

The results of all ecological studies, including time-series designs should be interpreted with caution:1

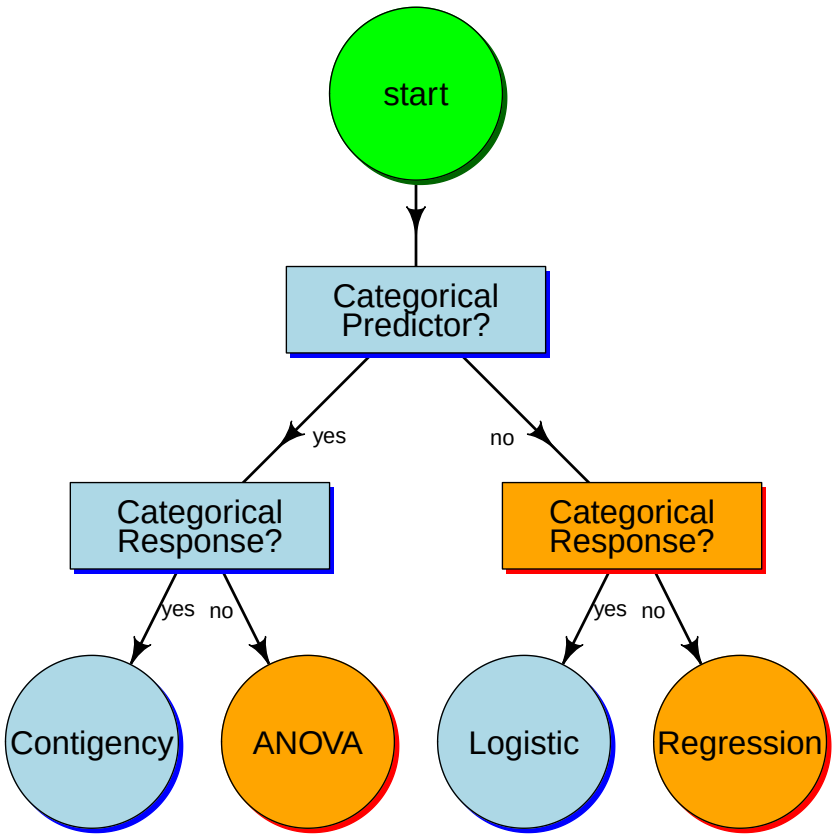
Data on exposure and outcome may be collected in different ways for different populations Migration of populations between groups during the study period may dilute any difference between the groups Such studies usually rely on routine data sources, which may have been collected for other purposes Ecological studies do not allow us to answer questions about individual risks

3.1 Data Source and Metadata

3.2 Evaluating the Structure of Data

Time series data have a natural temporal ordering. This makes time series analysis distinct from other common data analysis problems

Figure 1: Decision tree for frequents inference depending on the predictor and response data types.



This is analogous to selection a column or row of numbers in Excel to find the mean and you can usually find it by just looking at your spreadsheet to find the data of interest. In R you have to think a bit about what you want. Using the `str` command is good start, but we could also just look at the top of the observations to see which variables are of interest. To this we use the function `head()`, which is short for header, which shows the variable names and the first six observations.

```
head(Thailand)
```

```
##             STATION STATION_NAME     DATE  PRCP TAVG TMAX
## 1 GHCND:TH000048456 DON MUANG TH 19430101 -9999 27.6 33.9
## 2 GHCND:TH000048456 DON MUANG TH 19430102 -9999 26.8 31.7
## 3 GHCND:TH000048456 DON MUANG TH 19430103 -9999 27.2 32.8
## 4 GHCND:TH000048456 DON MUANG TH 19430104 -9999 27.3 33.3
## 5 GHCND:TH000048456 DON MUANG TH 19430105 -9999 27.8 32.2
## 6 GHCND:TH000048456 DON MUANG TH 19430106 -9999 27.1 32.8
##   TMIN    NewDate
## 1   NA 1943-01-01
## 2 21.7 1943-01-02
## 3 21.1 1943-01-03
## 4 21.1 1943-01-04
## 5 21.1 1943-01-05
## 6 21.1 1943-01-06
```

3.3 Evaluating for Completeness

NA is the R symbol for missing data and R requires the user to be fairly intentional about how to deal with missing data. Missing data usually mean the dataset is biased. In contrast to many software packages, R forces you to acknowledge the implications of missing data, which can be annoying, like a parent reminding you to clean your room or brush your teeth or take a shower once in the while. But the trade is worth it: you have dealt explicitly with missing data.

3.4 Evaluating the Central Tendencies

One of the first things you should do with your data is determine some of the central tendencies. For example, the mean, median, and standard deviation. Also some graphing of the data is also important. For example, what does the distribution of the data look like?

Let's start with the easy stuff. We want to get the mean of the maximum temperatures. That means we need to get the values, named TMAX from the data frame.

Okay, so we want "average." But typing average by itself doesn't show us anything except an error. Let's try `str` again. Notice the dollar symbols. These

symbols are use to signify a list of values inside the data frame. To access this list, we type

```
Thailand$TMAX
```

So, now we can get the number of observations, i.e. the length of the vector, by typing

```
length(Thailand$TMAX)

## [1] 27026
```

Okay, let's calculate the mean. In this case, it requires caution. Notice there are NAs in the data.

Typing `mean(Thailand$TMAX)` gives an a ambiguous result, `NA`. Try it. R is basically saying that the mean can not be calculated because of missing values, thus the mean is also missing. So, can we not calculate the mean when data are missing? No, we just have to tell R what to do with missing data. In this case, we tell R to remove them, with the argument `na.rm="TRUE"`, where True can be abbreviated to T. na.rm="TRUE" roughly translates to 'please remove all the NAs.'

Okay as of July 10, 2017, the average is 32.9461248[1]. It will change next month when May 2010 is added to the data set. Now let's determine the median and standard deviation.

```
median(Thailand$TMAX, na.rm=T)

## [1] 33

sd(Thailand$TMAX, na.rm=T)

## [1] 2.336892
```

If you would like a summary of each of the variables, the function is pretty easy to remember–but the output is not exceptionally pleasing.

```
summary(Thailand)

##                STATION             STATION_NAME        DATE
##  GHCND:TH000048456:27026   DON MUANG TH:27026   Min.   :19430101
##                                                 1st Qu.:19610848
##                                                 Median :19800265
##                                                 Mean   :19797426
##                                                 3rd Qu.:19980830
##                                                 Max.   :20170630
```

[1] How many significant figures should you report? Have I reported this correctly?

```
##
##       PRCP              TAVG              TMAX            TMIN
##  Min.   :-9999.0   Min.   :-9999.0   Min.   :19.30   Min.   : 2.40
##  1st Qu.:    0.0   1st Qu.:   27.1   1st Qu.:31.60   1st Qu.:23.00
##  Median :    0.0   Median :   28.4   Median :33.00   Median :24.50
##  Mean   :-1532.3   Mean   : -255.2   Mean   :32.95   Mean   :24.02
##  3rd Qu.:    1.0   3rd Qu.:   29.6   3rd Qu.:34.40   3rd Qu.:25.60
##  Max.   :  484.1   Max.   :   34.4   Max.   :40.80   Max.   :30.10
##                                      NA's   :5066    NA's   :7760
##     NewDate
##  Min.   :1943-01-01
##  1st Qu.:1961-08-31
##  Median :1980-02-29
##  Mean   :1980-03-04
##  3rd Qu.:1998-08-29
##  Max.   :2017-06-30
##
```

Nevertheless, the output gives you a really good idea regarding the central tendencies of the entire data set. Granted typing code might seem like a major step backwards in the computer world, but after a few weeks you will appreciate not having the search through arcane menus to find which button to push–even worse, in these push-button software systems, it often hard to figure out what they are doing. In the case of R, you have a really good idea of what it did, but were much more engaged in the process.

3.5 Evaluating Spread

When the mean and median diverge, it means that the distribution is skewed in some way. Let's see what the distribution looks like by creating a histogram.

```
hist(Thailand$TMAX)
```

The one you have made probably does not look that pretty, but with some more advanced coding, this is what it might look like (Figure 2).

Congratulations, you have made it through the next step in R! You now know how to do an exploratory analysis and even generate a basic histogram to view the distribution of a data set. Next, we use a standard statistical technique to determine the slope of the line and weather the line is statistically significant–but first we need to understand something about hypothesis testing.

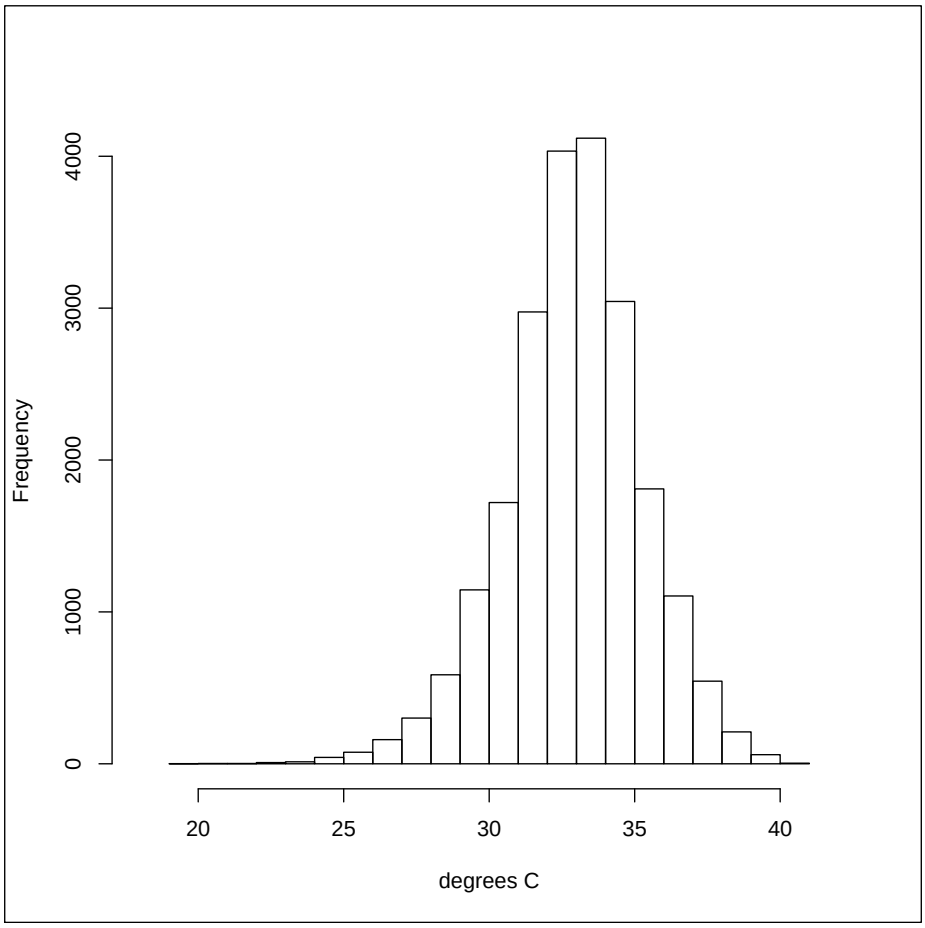


Figure 2: Histogram of Maximum Temperatures, Bangkok, Thailand.

3.6 'Null Hypotheses' as the Foundation of Frequentist Statistics

3.7 Type I and Type II Errors

3.8 Mathematical Mechanisms to Test Hypothesis

Using the linear model, we can analyze several types of data, when the response variable is continuous. If the have a predictor variable that is categorical, then we often analyze the data using the method known as analysis of variance or ANOVA. If the predictor variable is continuous, then we often analyze data using a regression analysis.

Table 2: 2 x 2 Matrix of Inference Methods –Going to insert graphics to give little pics of analysis...

		Reality	
		Truth	Not True
Statistical Result	Belief	Correct Decision 	Type I
	Non-belief	Type II	Correct Decision

4 Linear Models in R

The use of the linear model is the cornerstone of statistics. So ubiquitous it is rarely explained coherently. The linear model can be summarized at the equation for a line, but with the addition of error. You are probably familiar with the equation for a line where,

$$y = m * x + b \tag{1}$$

This equation defines a line, where m is the slope, b is the y-intercept, and the x and y are coordinates. The linear model is based on this form and is usually written as

$$y \sim \alpha + \beta * x + \epsilon \tag{2}$$

The order is usually changed, where the intercept is first, followed by the slope and x variable and the addition of error or noise. The error is usually symbolized as ϵ. In general, in a statistical model, Greek letters are used and instead of an equals sign, we use a tilde, meaning that that left side of the equation is a function of the right side. Luckily, this is the approximate form that R expects, so if you understand this, you will have a pretty good idea of how to code a linear model in R.

The function to build a linear model is `lm()`. This function is extremely powerful and can be easily implemented, but this is a good time to see what the help menus look like in R.

```
help(lm)
```

I am not showing it here, but you should see a long complex looking help page window pop up. All help files in R are structured the same way, so in spite of the uninterpretable text, written by and for computer programmers, the structure will become familiar. Beginning with the description, the help screen describes the function, how to use it, and give some examples. Admittedly, I rarely understand much of the text, but I find the examples to be very useful!

In fact, I suggest you paste the example into R and see what happens, I find this one of the best ways to learn R. Use an example that I know works, then change it to make it do what I want it to do.

5 Simple Regression

5.1 What is Linear Regression?

Linear regression is the most basic and commonly used predictive analysis. Regression estimates are used to describe data and to explain the relationship between one dependent variable and one or more independent variables. At the center of the regression analysis is the task of fitting a single line through a scatter plot. The simplest form with one dependent and one independent variable is defined by the formula:

$$y = a + b * x. \tag{3}$$

Sometimes the dependent variable is also called the response. The independent variables are also predictor variables. However, Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points. It consists of 3 stages:

1. analyzing the correlation and directionality of the data,
2. estimating the model, i.e., fitting the line, and
3. evaluating the validity and usefulness of the model.

There are three major uses for Regression Analysis: 1) causal analysis, 2) forecasting an effect, 3) trend forecasting. Other than correlation analysis, which focuses on the strength of the relationship between two or more variables, regression analysis assumes a dependence or causal relationship between one or more independent and one dependent variable.

Firstly, it might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, age and income.

Secondly, it can be used to forecast effects or impacts of changes. That is, regression analysis helps us to understand how much the dependent variable will change when we change one or more independent variables. Typical questions are, How much additional Y do I get for one additional unit of X?.

Thirdly, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. Typical questions are, “What will the price for gold be 6 month from now?” “What is the total effort for a tasks?”

5.2 Regression and Climate Change

One of the ourcomes of the linear regression is to estimate the best fit line

$$y = mx + b + \epsilon, \tag{4}$$

where ϵ is an estimate of the error. In addition, two other estimates are provided, one for the slope, m, and the y-intercept, b.

But these estimates are also hypotheses, where the null hypothesis is:

slope is zero Rejecting the null hypothesis would be support the alternative hypothesis, or the estimate of the slope.

y-intercept is zero Rejecting the null hypothesis would support the alternative hypothesis, the estimate of the y-intercept.

Okay, let's see if we can do this for our Bangkok data. Let's test if there is a significant change of daily maximum temperatures (TMAX) with time. Thus, in general terms, Maximum temperature is a function of time, or $TMAX = f(Time)$.

$$TMAX \sim \alpha + \beta * time + \epsilon \tag{5}$$

Translating this in R will take some additional tricks besides just getting the code figured out. First, we need to identify the predictor variable, 'NewDate', in the data frame which we created in SOP85.

Because these data are in a time series, they are serially correlated, meaning that the June sample will be more like the July sample than the August sample. In addition, the June 2010 sample will be similar to the June 2009 sample. These correlation violate the assumption of independence, but for now, we will ignore this violation and just create a linear model in bliss.

For the response variable, we will use the daily maximum temeratures, TMAX. Remember there are some missing data, it will be interesting to note how R deals with that.

First, let's create a plot of data using `plot()`, whose format is `plot(x, y)` or `plot(y  x)`. We will use the later for now,

Finally, there is one important difference between the linear model that we used in the `aov()` function. This time we use the `lm()` function that arrange the results more in-line with a regression model. This syntax is still pretty straight forward,

```
lm(TMAX ~ NewDate, data=Thailand)

##
## Call:
## lm(formula = TMAX ~ NewDate, data = Thailand)
##
## Coefficients:
```

Figure 3: Maximum daily temperatures for Bangkok, Thailand.

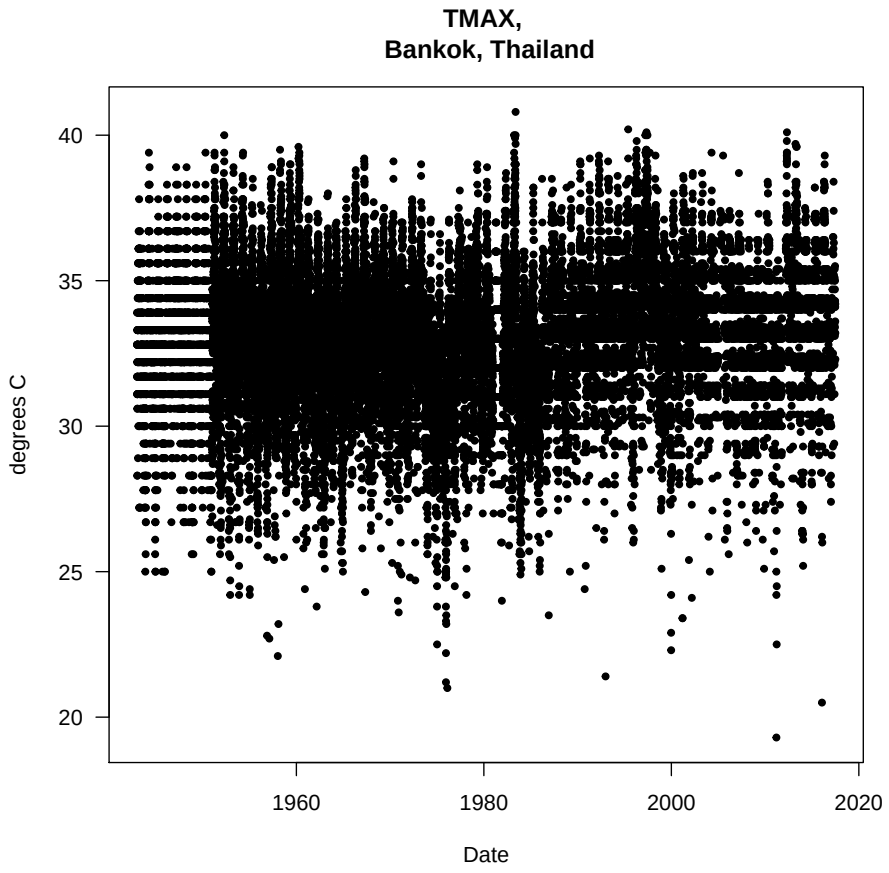
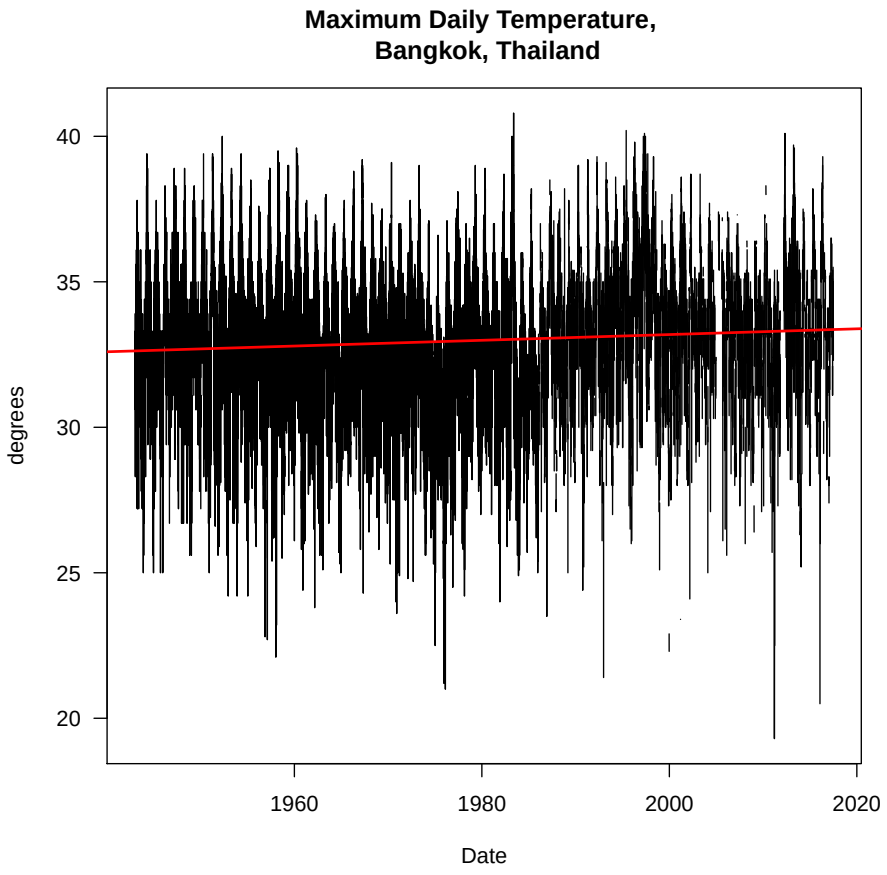


Figure 4: TMAX in Bankok, Thailand.




```
## (Intercept)      NewDate
##   3.289e+01    2.702e-05
```

From this model, we learn that the change in $TMAX$ is 0 degrees $year^{-1}$. Figure 5.2 shows the increasing concentrations, but also the seasonal variation. Statisticians have more advanced methods to analyze these data then what we have done, but for our purposes the implications are the same. Greenhouse gas emissions are increasing and the estimated rates suggest an increasing rate.

Now let's ask if this value is significant, by putting the linear model into a ANOVA-like table. There are a number of functions that do this and we have seen the `anova()` function above. For linear regression, however, the `summary()` function gives a more complete output.

```
summary(lm(TMAX ~ NewDate, data=Thailand))

##
## Call:
## lm(formula = TMAX ~ NewDate, data = Thailand)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9974  -1.3193   0.0782   1.4717   7.7771
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.289e+01  1.631e-02 2016.35   <2e-16 ***
## NewDate     2.702e-05  2.140e-06   12.62   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.329 on 21958 degrees of freedom
##   (5066 observations deleted due to missingness)
## Multiple R-squared:  0.007202,Adjusted R-squared:  0.007157
## F-statistic: 159.3 on 1 and 21958 DF,  p-value: < 2.2e-16
```

Here we find the that the slope and intercept are highly significant, we have some information on the residuals, and R^2 estimates, etc.

5.3 Creating Monthly Averages of Daily Maximum Temperatures

One of the first things to note is how messy the data look and there are lots of sources of variation. For example, we expect months to respond differently to the climate change. To assess this, we will now analyze the data for monthly means of the maximum temperatures.

5.4 Creating Monthly Means

To create monthly means, we need to disagragate the NewDate variable into a month and year variables.

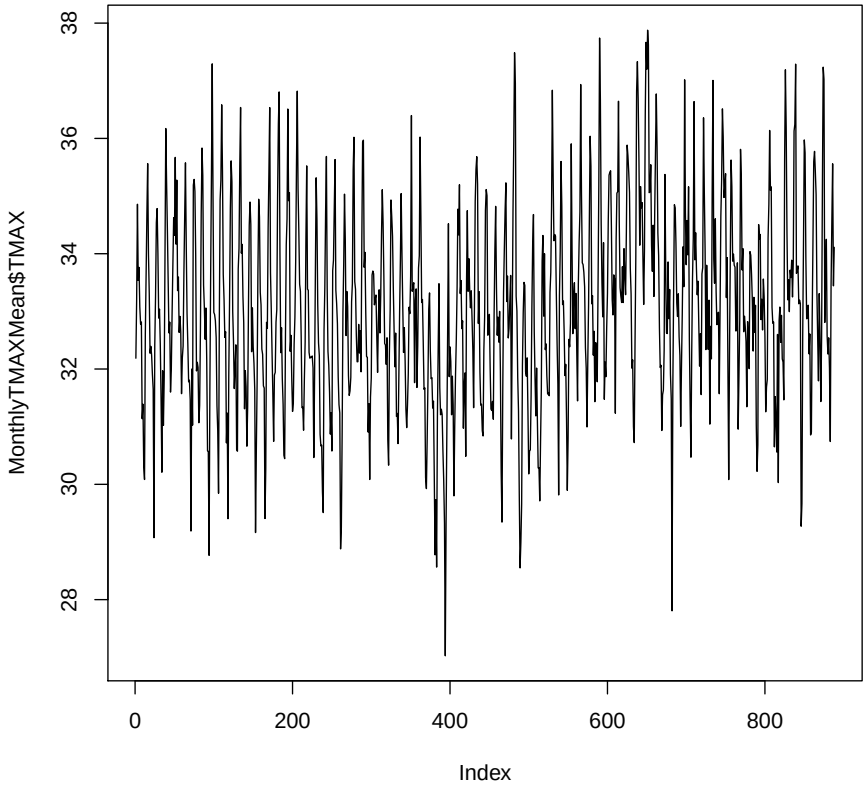
First we can use the `as.Date()` function to extract a portion of the date, where %m is for month and %Y is for a four digit year. Then, we create new variables in our dataframe, one for month and one for year.

```
Thailand$Month = format(as.Date(Thailand$NewDate), format = "%m")
Thailand$Year = format(Thailand$NewDate, format="%Y")
```

After creating the month and year as separate variables, we can use them to caculate the mean using the `aggregate()` function. In the code below, we can also calculate the standard deviation too, although I haven't used this measure in this document, several students have asked for this for their analysis.

```
MonthlyTMAXMean = aggregate(TMAX ~ Month + Year, Thailand, mean)

MonthlyTMAXMean$YEAR = as.numeric(MonthlyTMAXMean$Year)
MonthlyTMAXMean$MONTH = as.numeric(MonthlyTMAXMean$Month)
str(MonthlyTMAXMean)

## 'data.frame': 888 obs. of  5 variables:
##  $ Month: chr  "01" "02" "03" "04" ...
##  $ Year : chr  "1943" "1943" "1943" "1943" ...
##  $ TMAX : num  32.2 33.2 34.9 33.5 33.8 ...
##  $ YEAR : num  1943 1943 1943 1943 1943 ...
##  $ MONTH: num  1 2 3 4 5 6 7 8 9 10 ...

MonthlyTMAXSD = aggregate(TMAX ~ Month + Year, Thailand, sd)

MonthlyTMAXSD$YEAR = as.numeric(MonthlyTMAXSD$Year)
MonthlyTMAXSD$MONTH = as.numeric(MonthlyTMAXSD$Month)
MonthlyTMAXSD$NewDate = MonthlyTMAXSD$YEAR + (MonthlyTMAXSD$MONTH - 1)/12

head(MonthlyTMAXSD)

##   Month Year     TMAX YEAR MONTH  NewDate
## 1    01 1943 1.523317 1943     1 1943.000
## 2    02 1943 1.668648 1943     2 1943.083
## 3    03 1943 1.969395 1943     3 1943.167
## 4    04 1943 2.521970 1943     4 1943.250
## 5    05 1943 2.100818 1943     5 1943.333
## 6    06 1943 1.041763 1943     6 1943.417
```

```
plot(MonthlyTMAXMean$TMAX, ty='l')
```



Below is Standard Deviation

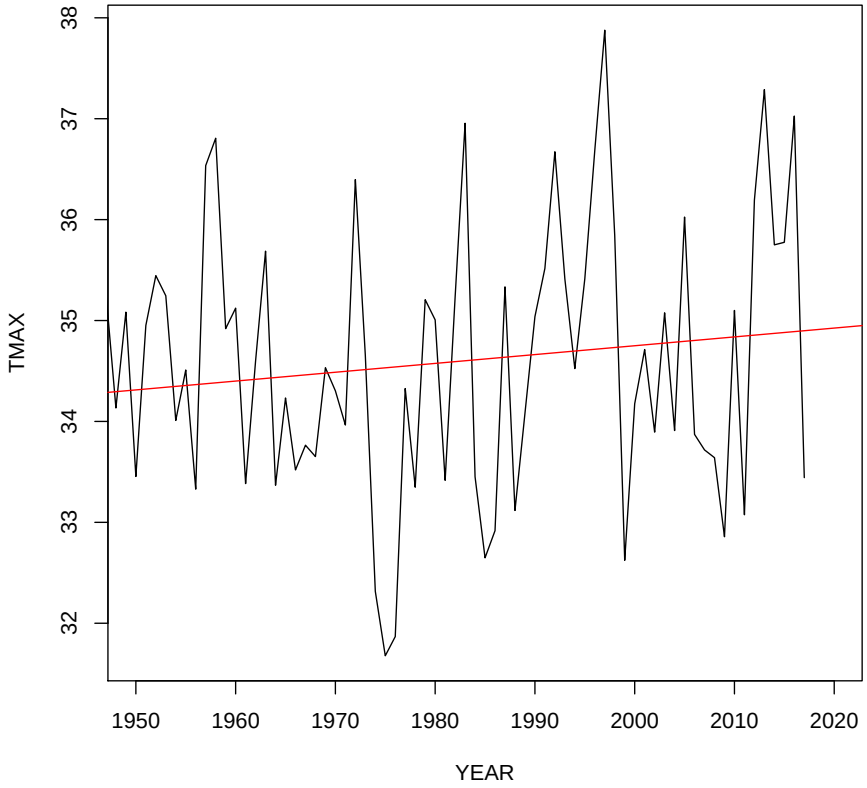
```
#plot(MonthlySD£TMAX, ty='l')

#plot(TMAX~ NewDate, data=MonthlySD, ty='l')
#SD.lm <- lm(TMAX~NewDate, data=MonthlySD)
#summary(SD.lm)

#abline(coef(SD.lm), col="red")
```

Selecting for 1 Month – May

Perhaps, we can get a better handle on this stuff if we analyze for just one month at a time – certainly easier to visualize!

```
#plot(MonthlyTMAXMean£TMAX[MonthlyTMAXMean£Month=="05"], ty='l')
plot(TMAX~YEAR, data=MonthlyTMAXMean[MonthlyTMAXMean$Month=="05",], ty='l', xlim=c(1950, 202
May.lm <- lm(TMAX~YEAR, data=MonthlyTMAXMean[MonthlyTMAXMean$Month=="05",])
```

```r
summary(May.lm)

##
## Call:
## lm(formula = TMAX ~ YEAR, data = MonthlyTMAXMean[MonthlyTMAXMean$Month ==
##     "05", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.85376 -0.93210 -0.04633  0.81231  3.15347
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.238621  13.753642   1.253    0.214
## YEAR         0.008756   0.006946   1.261    0.211
##
## Residual standard error: 1.302 on 73 degrees of freedom
## Multiple R-squared:  0.0213,Adjusted R-squared:  0.007897
## F-statistic: 1.589 on 1 and 73 DF,  p-value: 0.2115

abline(coef(May.lm), col="red")
```



Now, the change is 0.0088 degress C/year or 0.876 degress C/100 years with a probability of 0.2115. Although we can't reject the null hypothesis, we find the method to be fairly straightforward!

5.5 Testing all the Months

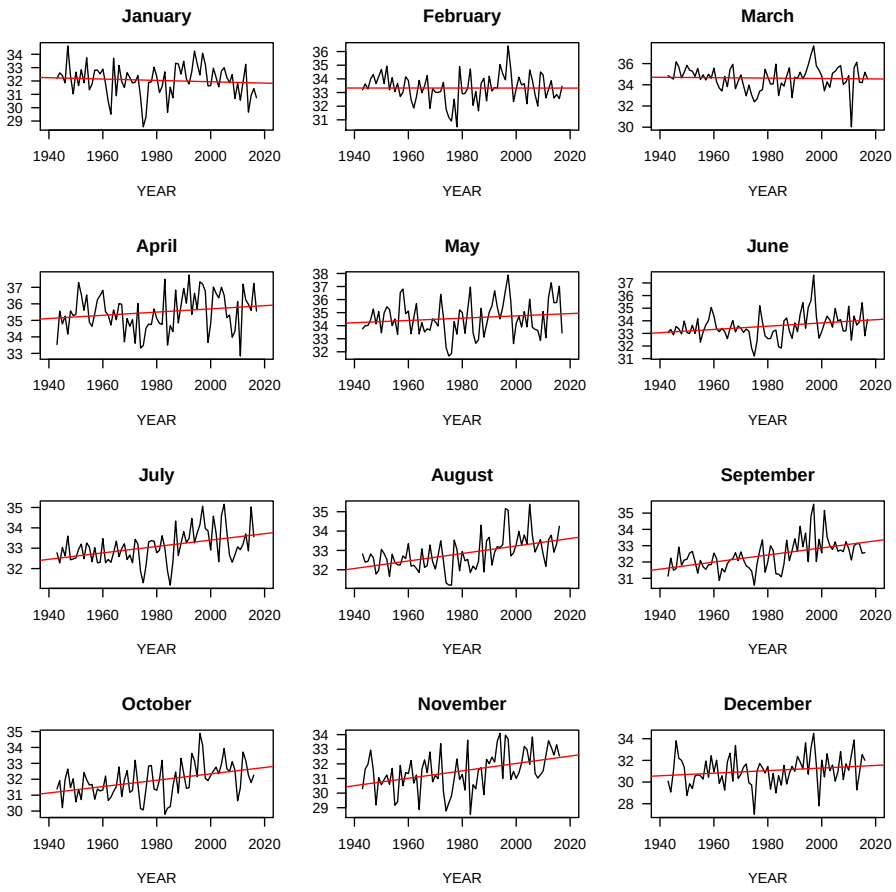
I think you should evaluate every month and see what happens. You might also consider looking at the TMIN as well. Could be important![2]

Below, I have create code to evaluate all of the months at once, but you may prefer to go through each month manually and change the number from 5 to other months of the year.

5.5.1 TMIN

1. We create a monthly TMIN mean for each month.

[2] What about multiple hypotheses in one dataset!


January
February
March
April
May
June
July
August
September
October
November
December
YEAR
1940
1960
1980
2000
2020

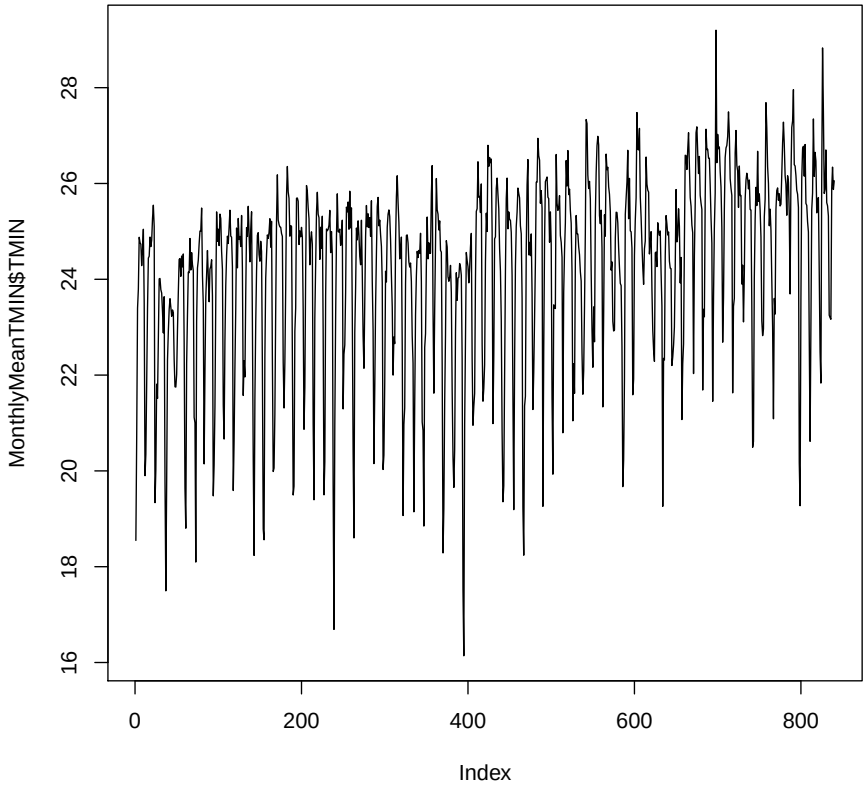
```r
MonthlyMeanTMIN = aggregate(TMIN ~ Month + Year, Thailand, mean)

MonthlyMeanTMIN$YEAR = as.numeric(MonthlyMeanTMIN$Year)
head(MonthlyMeanTMIN)
```

```
##   Month Year     TMIN YEAR
## 1    01 1943 18.54828 1943
## 2    02 1943 20.73077 1943
## 3    03 1943 23.39655 1943
## 4    04 1943 23.79259 1943
## 5    05 1943 24.87692 1943
## 6    06 1943 24.76429 1943
```

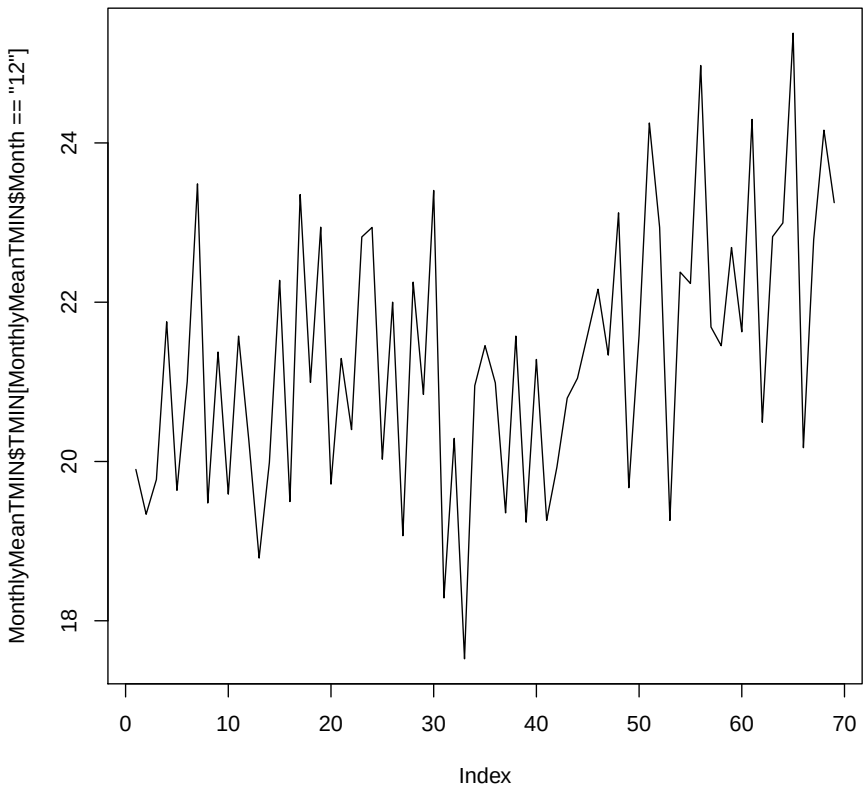
5.6 TMIN

2. Now we plot the mins, and again, find tons of scatter.

```r
plot(MonthlyMeanTMIN$TMIN, ty='l')
```

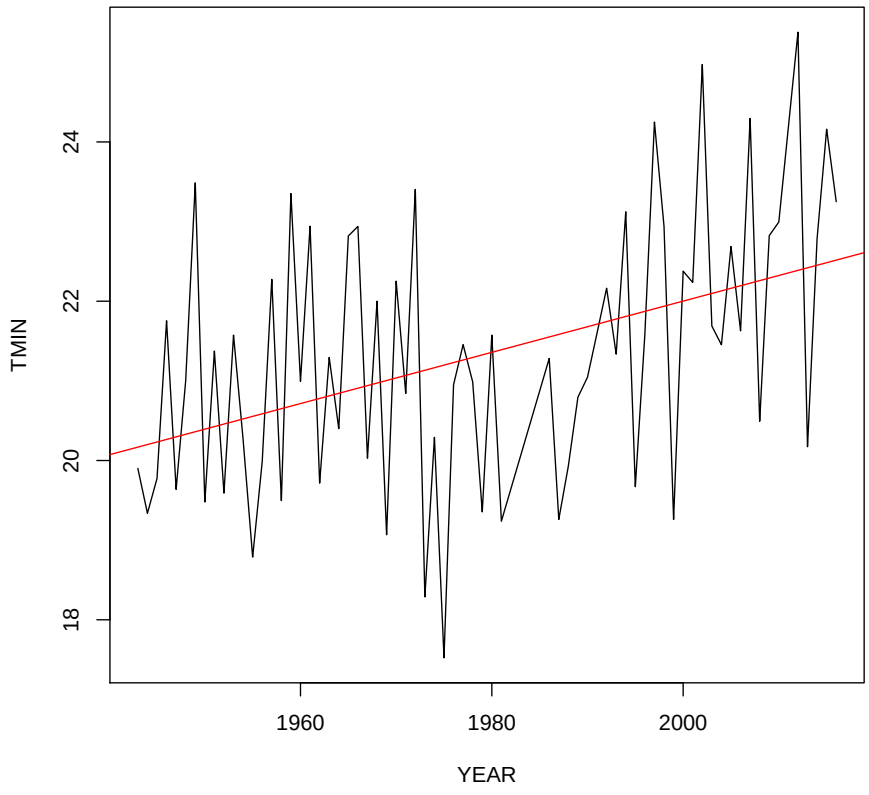



```
plot(MonthlyMeanTMIN$TMIN[MonthlyMeanTMIN$Month=="12"], ty='l')
```




```
plot(TMIN~YEAR, data=MonthlyMeanTMIN[MonthlyMeanTMIN$Month=="12",], ty='l')
Dec.lm <- lm(TMIN~YEAR, data=MonthlyMeanTMIN[MonthlyMeanTMIN$Month=="12",])
summary(Dec.lm)

##
## Call:
## lm(formula = TMIN ~ YEAR, data = MonthlyMeanTMIN[MonthlyMeanTMIN$Month ==
##     "12", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6746 -0.8672 -0.2573  1.0020  3.1247
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.320213  16.700175  -2.534 0.013617 *
```

```
## YEAR          0.032161   0.008439   3.811 0.000303 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.525 on 67 degrees of freedom
## Multiple R-squared:  0.1782,Adjusted R-squared:  0.1659
## F-statistic: 14.52 on 1 and 67 DF,  p-value: 0.0003034

abline(coef(Dec.lm), col="red")
```



5.7 TMIN

3. In this case, we get a slope, 0.0321607 degress/year and a probability of 3.034×10^{-4} and an r-sqaured of 0.178. Cool! As we might expect, the a small amount of the variance is explained by the "Month." Many things predict

temerpature, that year is one, is quite problematic.

4. What we have not determined is the cause. So, be careful when you describe the results, cause and effect cannot be analyzed using this method.

5.7.1 Departure from Mean

```
#PRCP_mean = mean(LosAngeles£PRCP, na.rm=T)

#plot(PRCP~NewDate, data=LosAngeles)
#abline(h=PRCP_mean, col="blue")
```

5.7.2 Experimental Portion — Precipitation

Precipiation might depend more on the departure from the mean (often referred as as normal, whatever that means!). I think it's worth pursuing, but haven't finished the analysis yet.

First, we need a "mean" – The IPCC uses 1961-1990 as a norm, I don't know what is the standard for California, so we should look that up.

Second, we need to remove the missing values and evalaute which years have complete years. If you are missing rainy months, then the whole year should be thrown out – but what about partial years in the drought season?

Third, we will need to decide what level of aggredation – monthly, yearly, etc.

Fourth, in CA the water year starts in Oct 1. Should we follow the same convention?

```
#LosAngeles£PRCP[LosAngeles£PRCP==-9999] <- NA
#YearlySum = aggregate(PRCP ~ Year, LosAngeles, sum)
#YearlySum£YEAR = as.numeric(YearlySum£Year)
#YearlyMean = mean(YearlySum£PRCP)
```

A yearly mean, based on the annual sum for the entire records. Not sure this is appropriate.

Figure has points of the yearly sum of rainfall and the blue line mean. The greenline is the trend and red line is a five year running average, I think! I am still trying to understand what the code is doing.

```
#plot(PRCP~YEAR, data=YearlySum, las=1, ty="p")
#abline(h=YearlyMean, col="blue")
#YearlySum.lm = lm(PRCP~YEAR, data=YearlySum)
#abline(coef(YearlySum.lm), col="green")

#n <- 5
#k <- rep(1/n, n)
```

```
#k

#y_lag <- stats::filter(YearlySum£PRCP, k, sides=1)
#lines(YearlySum£YEAR, y_lag, col="red")
```

```
#summary(YearlySum.lm)
```

5.8 Problems with a Simple Regression Model

Regression models, like all statistics, rely on certain assumptions. Violations of these assumptions reduces the validity of the model. If the violations are serious, then the model could be misleading or even incorrect.

Here is a list of assumptions to produce a valid regression model:

Homogeneity of Variance

something else

Assumptions about e , the error term: t i. E(et) = 0, zero mean ii. E(et) = s , constant variance 2 2 iii. E(et , Xt) = 0, no correlation with X iv. E(e , e) , no autocorrelation. t t-1 v. e Normally distributed (for hypothesis testing). t 3. Assumption four is especially important and most likely not to be met when using time series data.

Autocorrelation.

1. It is not uncommon for errors to track themselves; that is, for the error a time t to depend in part on its value at t - m, where m is a prior time period.

5.8.1 Model Diagnostics

With every statistical test done, researchers validate their model in some way or anther. Often this entails the use of diagnostics, a standardize battery of procedures to check to see if the data are following the assumptions.

In R four plots are created by default. To see them all at the same time, we need to change the graphical parameters so the graphics window expects four panels, in this case a 2 rows and two columns.

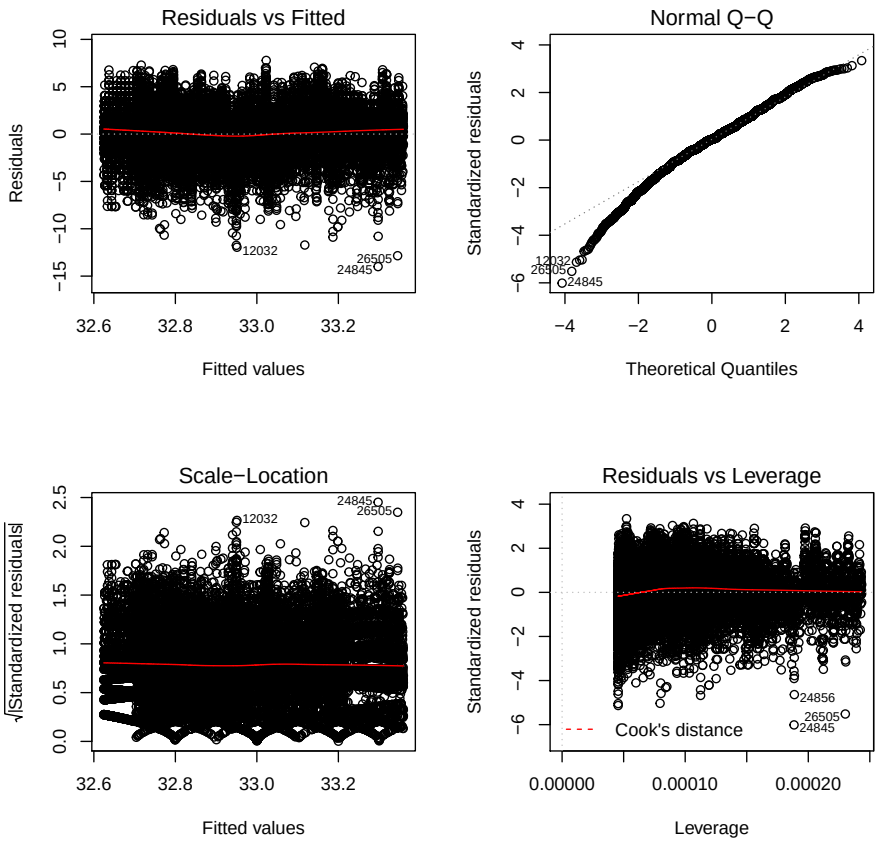
```
par(mfrow=c(2,2))
```

Try not to get bogged down in the code at this point. But it is a useful thing to remember.

To determine the validity of linear model assumptions (e.g. normality or heterogeneity of variance), you have probably used statistical tests; in contrast statisticians almost exclusively look at diagnostic plots. Why? When assumptions are violated the tests to determine violations do not perform well. So, let's see how to look at these assumptions graphically with these diagnostic plots.

Figure 5: Default diagnostic plots for a linear model in R.

```
par(mfrow=c(2,2))
plot(lm(TMAX ~ NewDate, data=Thailand))
```



Linear models should have diagnostic plots that do not have any obvious structure or pattern. In this case, Figure 5.8.1 should show a great deal remaining structure in the residuals. Although for today, we are not going to try to interpret these figures, but you should notice there is a ton of unaccounted structure, i.e. variance, in the model. This is due, in part, to a violation of independence; these data are serially correlated and the model does not account for that and is inappropriate because of this. It also appears that a straight-line model does not fit well and a curvilinear should be investigated.

A properly specified model is shown in

6 The 'Null' Hypothesis versus Information Criteria

6.1 Model Comparison

6.2 AIC to make statements about strength of evidence

7 Relaxing Model Assumptions

7.1 Using Sources of Error in the Model

Instead of letting autocorrelation be 'hidden' problem in the data, we can incorporate the correlation structure into the model and use it to our advantage – create a better, i.e. unbiased estimate of the model parameters.

7.2 Generalized Least Square (GLS) and Autocorrelation

```
library(nlme)

#TMAX.gls = gls(TMAX ~ NewDate, data = Thailand, na.action=na.omit)
#summary(TMAX.gls)
#TMAX.gls2 = gls(TMAX ~ NewDate, data = Thailand, correlation = corAR1(form=~1), na.action=
#summary(TMAX.gls2)

#anova(TMAX.gls, TMAX.gls2)
```

7.3 Adding Seasonality

8 More Sophisticated Approaches

9 Advanced Methods

You may want to examine the GAM package in R, as it can be adapted to do some (or all) of what you are looking for. The original paper (Hastie & Tibshirani, 1986) is available via OpenAccess if you're up for reading it.

Essentially, you model a single dependent variable as being an additive combination of 'smooth' predictors. One of the typical uses is to have time series and lags thereof as your predictors, smooth these inputs, then apply GAM.

This method has been used extensively to estimate daily mortality as a function of smoothed environmental time series, especially pollutants. It's not OpenAccess, but (Dominici et al., 2000) is a superb reference, and (Statistical Methods for Environmental Epidemiology with R) is an excellent book on how to use R to do this type of analysis.

10 Time Series Analysis

Time series analysis

Time series analysis refers to a particular collection of specialised regression methods that use integrated moving averages and other smoothing techniques to illustrate trends in the data. It involves a complex process that incorporates information from past observations and past errors in those observations into the estimation of predicted values.

Moving averages provide a useful way of presenting time series data, highlighting any long-term trends whilst smoothing out any short-term fluctuations. They are also commonly used to analyse trends in financial analysis. The calculation of moving averages is described in more detail here.

Methods for time series analyses may be divided into two classes: frequency-domain methods and time-domain methods. The former include spectral analysis and recently wavelet analysis; the latter include auto-correlation and cross-correlation analysis. In time domain correlation analyses can be made in a filter-like manner using scaled correlation, thereby mitigating the need to operate in frequency domain.

Whether or not you wish to forecast or not has nothing whatsoever to do with correct time series analysis. Time series methods can develop a robust model which can be used simply to characterize the relationship between a dependent series and a set of user-suggested inputs (a.k.a. user-specified predictor series) and empirically identified omitted variables be they deterministic or stochastic.Users at their option can then extend the "signal" into the future i.e. forecast with uncertainties based upon the uncertainty in the coefficients and the uncertainty in the future values of the predictor . Now these two kinds of empirically identified "omitted series" can be classified as 1) deterministic

and 2) stochastic. The first type are simply Pulses, Level Shifts , Seasonal Pulses and Local Time Trends whereas the second type is represented by the ARIMA portion of your final model. When one omits one or more stochastic series from the list of possible predictors, the omission is characterized by the ARIMA component in your final model. Time series modelers refer to ARIMA models as a "Poor Man's Regression Model" because the past of the series is being used as a proxy for omitted stochastic input series.

11 References