SOP 90: Analyzing Trends

Marc Los Huertos

July 3, 2017

tufte-handout
verbatim amsmath natbib
graphicx
Time Series Analysis and Display [Marc Los Huertos]Marc Los Huertos
booktabs
multicol

Abstract

Time Series Analyses are very important in environmental monitoring.

1 Introduction

Trend analysis... versus time series...

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.

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis".

Time series data have a natural temporal ordering. This makes time series analysis distinct from other common data analysis problems, in which there is no natural ordering of the observations (e.g. explaining people's wages by reference to their respective education levels, where the individuals' data could be entered in any order). Time series analysis is also distinct from spatial data analysis where the observations typically relate to geographical locations (e.g.

accounting for house prices by the location as well as the intrinsic characteristics of the houses). A stochastic model for a time series will generally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time series models will often make use of the natural one-way ordering of time so that values for a given period will be expressed as deriving in some way from past values, rather than from future values (see time reversibility.)

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.

2 Simple Regression

$y = mx + b$, where x is time.

y x

$$y = mx + b + \epsilon \tag{1}$$

```
## Warning in file(file, "rt"):  cannot open file 'H:\My Documents\My
Webs\Mauna_Loa.csv':  No such file or directory
## Error in file(file, "rt"):  cannot open the connection
```

Okay, you know have created a data frame. To confirm this, type `str(maunaloa)` and you should see some strange text that describes the data frame. This function allow you to peer into the data frame structure. You you see it is a data frame and it has several variables and each one has certain characteristics and R even shows you some of the observations. This is a good thing to get into the habitat of check, for you want to ensure the data have been imported in a way that you expect.

Remember, a data frame is a list of vectors. To access the data inside the data frame, you can use the following command

```
#maunaloa$average
```

to dump the average CO_2 concentrations readings onto your screen as a vector. You should see some 627 observations, depending on how recent the data have been uploaded. So, the dollar symbol is used to drill into the data frame vectors. And when you look at the `str()` function again, you will see these dollar signs again.

3 Exploring Data: The First Step of Data Analysis

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 monthly average CO_2 concentrations. That means we need to get the values, named "average" from the data frame. 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(maunaloa)
```

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

```
#maunaloa$average
```

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

```
#length(maunaloa$average)
```

Okay, let's calculate the mean. In this case, it requires caution. Notice there are NAs in the data. NA is the R symbol for missing data and R requires the user to be fairly intentional about how to deal with missing data. [1]

Typing `mean(maunaloa$average)` 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.'

[1] 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.

Figure 1: Histogram of carbon dioxide concentrations at Mauna Loa, Hawaii.

```
#mean(maunaloa$average, na.rm=T)
```

Okay as of Jan 31, 2017, the average is 352.3159[2]. It will change next month when May 2010 is added to the data set. Now let's calculate the median and standard deviation.

```
#median(maunaloa$average, na.rm=T)
#sd(maunaloa$average, na.rm=T)
```

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(maunaloa)
```

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.

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(maunaloa$average)
```

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

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.

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

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

equation for a line, but with the addition of error. You are probably familiar with the equation for a line where,

$$y = m * x + b \tag{2}$$

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

$$y \sim \alpha + \beta * x + \epsilon \tag{3}$$

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.

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.

Okay, let's see if we can do this for our Mauna Loa data. Let's test if there is a significant change of carbon dioxide concentrations with time. Since both the predictor and response variables are continuous data, this analysis will be a linear regression, but the form and function of the linear model are exactly the same. The linear model would look something like this

$$CO_2 \sim \alpha + \beta * time + \epsilon \tag{4}$$

Translating this in R will take some additional tricks besides just getting the code figured out. First, we need to identify the predictor variable in the

Figure 2: Carbon dioxide concentrations sampled for each month at Mauna Loa, Hawaii.

Figure 3: Carbon dioxide concentrations at Mauna Loa, Hawaii.

data frame. There are three variables associated with time: year, month, and decimal.date. 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 today, we will ignore this violation and just create a linear model in bliss. So, let's use year as the predictor variable and assume there might be some error because each month have slightly different concentrations. For the response variable, we will use the monthly averages, "'average"'. 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,

```
#plot(average ~ year, data=maunaloa)
```

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(average ~ year, data=maunaloa)
```

From this model, we learn that the change in CO_2 is ppm $year^{-1}$.When I first made this handout the rate of increase was 1.441. Why do you think this rate has changed? Figure 4 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(average ~ year, data=maunaloa))
```

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

Figure 4: Default diagnostic plots for a linear model in R.

```
#par(mfrow=c(2,2))
#plot(lm(average ~ year, data=maunaloa))
```

4.1 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. Linear models should have diagnostic plots that do not have any obvious structure or pattern. In this case, Figure 4.1 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.

4.2 Problems with a Simple Regression Model

A properly specified model is shown in

5 More Sophisticated Approaches

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.

6 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.