**Introductory Time Series with R**

**Chapter 1 – Time Series Data**

**1.1 Purpose**

Generally speaking, time series analysis can be used to quantify the random variation inherent in the data, as well as to highlight the important/significant features of the data, thus allowing a clearer understanding of the past and the capability to knowledgeably predict the future.

To this last point, once a time series model is created it essentially can be used to simulate the future. A few examples include assessing different strategies for inventory control by simulating demand for a product, or simulating daily rainfall to investigate long-term environmental effects of proposed policies.

**1.2 Time Series**

As you probably suspect, variables that are measured sequentially over time form the basis of time series analysis. The period of time between observations of a particular variable is called the *sampling interval* and will vary depending on the problem. For signal processing data, the sampling interval could be a second. For national economic data, the sampling interval could be a year.

The book takes a statistical approach to time series, in that it assumes that the historical observations of the given variable are realizations of sequences of random variables. A few other important features of time series include:

* Trends and/or seasonal variations which must be taken into account to arrive at a forecast

Think about economic output of a popular beach town by month over the course of many years. The overall *trend* of output tends to go up with time as population grows, and there are inherent *seasonal variations* since there are more visitors during the summer months,

* Observations close together in time tend to be *serially dependent* on one another

This just means that observations close together in time tend to be correlated.

* Fitted models can be used as a basis for statistical tests

Say you have a time series model that forecasts a certain level of demand for a product next month, then that month comes and the actual demand was somewhat higher that forecasted. Is this increase statistical evidence that there is some underlying change in sales that must now be accounted for? Or is it to be expected because of the random variation inherent in that variable?

* 1. **Plots, trends, and seasonal variation**

Although the concept of a trend – or a systematic change in a time series that is not periodic – is straight-forward to grasp, it is difficult to discern whether a time series is demonstrating a trend that is part of a longer cycle or simply random and therefore subject to unpredictable change. These random (aka stochastic) trends are especially common in economic and financial time series. It is important to note that regression models are not appropriate for stochastic trends.

Forecasting relies on extrapolation, which in turn relies on the assumption that present trends continue. This is tricky because this assumption cannot be verified in any empirical way, but if we can identify likely causes of a trend, we can justify extrapolating it out to the near future. This argument is buffeted by the fact that a trend is likely to change slowly (absent some shock to the system) and therefore linear extrapolation is warranted.

There are two important things to keep in mind when extrapolating from a time series:

* Only use linear extrapolation (i.e., don’t use higher order polynomials even if they work well to describe the historic series)
* It is better to use linear extrapolation from the more recent values of the given variable, as opposed to basing the extrapolation on all available data

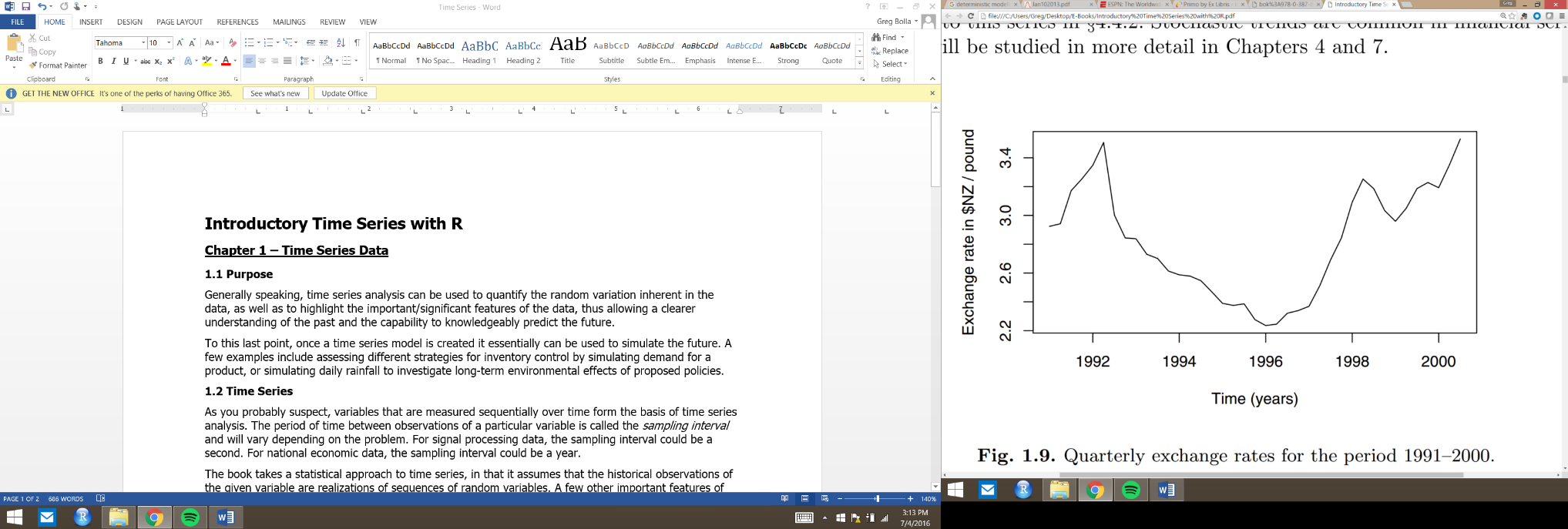
*1.4.3 Multiple time series: Electricity, beer and chocolate data*

This goes into the standard “correlation does not imply causation” of examining multiple time series variables, but includes a couple important new concepts. One is the intersect function in R to obtain the intersection of two series that overlap in time.

Another is how to remove trends and/or seasonal effects before comparing multiple time series. A common approach is to work with the residuals of a regression model that uses deterministic terms to represent the trend and seasonal effects. This is discussed more in Chapter 5.

*1.4.4 Quarterly exchange rate: GBP to NZ dollar*

This highlights an important difference between stochastic trends which emphasized randomness versus more deterministic trends like GDP (which trends upward as more time passes/the population grows).

Below is the graph of exchange rate between the New Zealand dollar and British pound. While deterministic trends are apparent through a similar graph, such marked patterns are less likely to be seen in stochastic trends.

In this case, the trend seems to change direction at unpredictable times rather than displaying a consistent pattern. However, if you had two separate time series – one from 1991 to the beginning of 1996 and another from 1996 to 2001 – you may think the overall trend was downward or upward depending on which time series you were analyzing. This underscores the importance of extrapolating stochastic trends without properly understanding the underlying causes. To reduce the risk of making inappropriate forecasts, statistical tests (Chapter 7) can be used to tests for a stochastic trend. Additionally, in these cases, a *random walk* model (Chapter 4) can sometimes provide a good fit to stochastic trend data to incorporate the inherent randomness.

* 1. **Decomposition of Series**

*1.5.3 Estimating trends and seasonal effects*

One of the simpler methods to estimate a trend at time is to calculate a *moving average* centered at . A moving average is an average of a specified number of time series values around each value in the time series (with the exception of the first and last few observations, since they obviously would not have the required values before or after to calculate the moving average). The length of the moving average is chosen to average out seasonal effects; so for monthly temperature data, as an example, you would want the length of the moving average to be over the course of the full year so you’re not giving too much weight to temperatures of a particular season.

In the temperature example, the centered moving average can be estimated at by the following equation:

The reason and are multiplied by ½ is to not give too much weight the timer series observation 6 months in the past/future (which is the same month of the year).

We can then estimate the monthly (in this case, it generalizes to any other frequency) additive effect, , by subtracting :

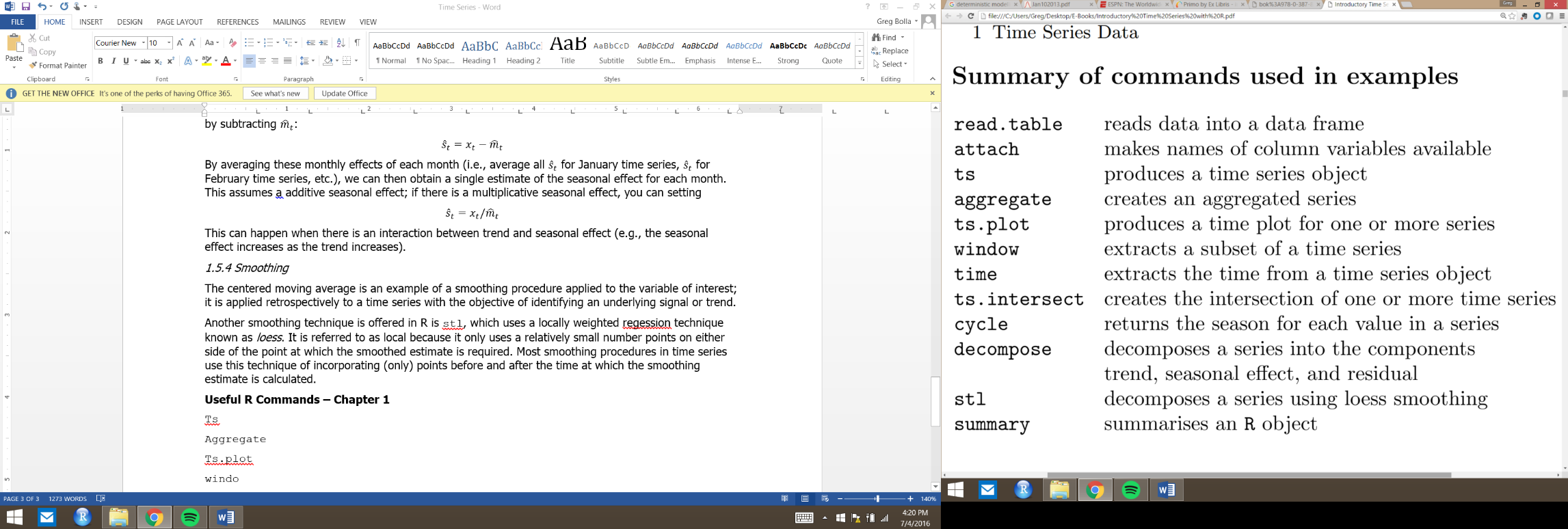
By averaging these monthly effects of each month (i.e., average all for January time series, for February time series, etc.), we can then obtain a single estimate of the seasonal effect for each month. This assumes a additive seasonal effect; if there is a multiplicative seasonal effect, you can setting

This can happen when there is an interaction between trend and seasonal effect (e.g., the seasonal effect increases as the trend increases).

*1.5.4 Smoothing*

The centered moving average is an example of a smoothing procedure applied to the variable of interest; it is applied retrospectively to a time series with the objective of identifying an underlying signal or trend.

Another smoothing technique is offered in R is stl, which uses a locally weighted regession technique known as *loess*. It is referred to as local because it only uses a relatively small number points on either side of the point at which the smoothed estimate is required. Most smoothing procedures in time series use this technique of incorporating (only) points before and after the time at which the smoothing estimate is calculated.

**Useful R Commands – Chapter 1**

**Chapter 2 – Correlation**

**2.1 Purpose**

Once the trend and seasonal effects are identified, can remove these two aspects of the outcome variable, thus leaving only the random component which can be modelled against. However, due to the fact that the outcome variable tends to be correlated with itself at different time periods, it may not be the best choice to model with the assumption that it is an independent random variable.

By identifying the correlations between the outcome variable at different time periods, we can improve our forecasts and generate realistic time series for simulations.

Plots of serial correlation, known as the *correlogram*, can be used to visualize this correlation of the outcome variable has with itself (called *autocorrelation*).

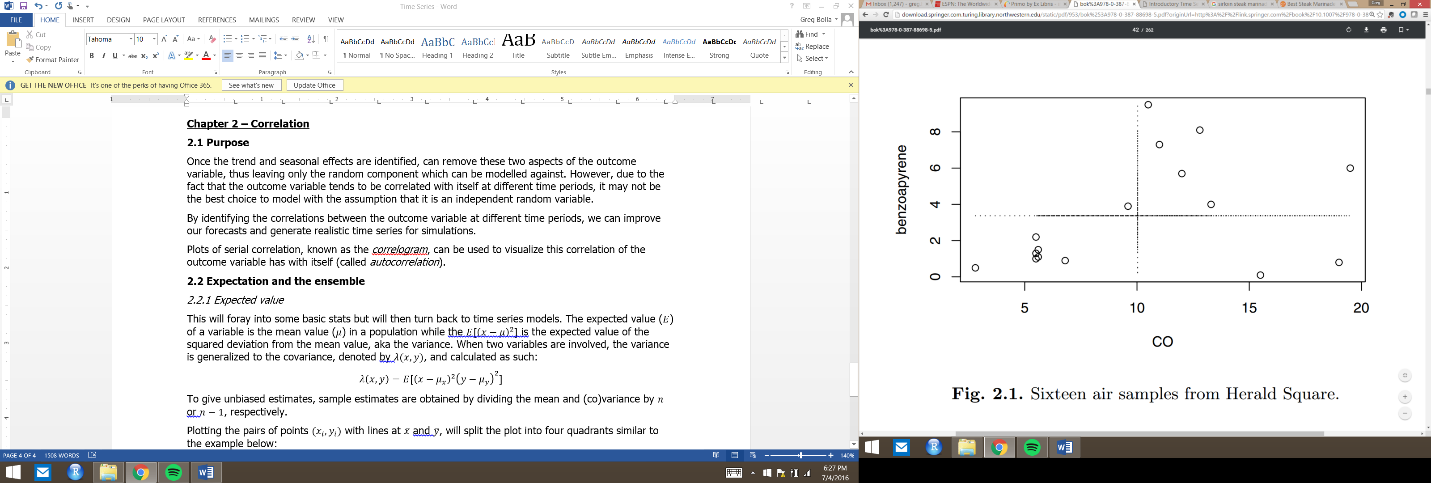
**2.2 Expectation and the ensemble**

*2.2.1 Expected value*

This will foray into some basic stats but will then turn back to time series models. The expected value () of a variable is the mean value () in a population while the is the expected value of the squared deviation from the mean value, aka the variance. When two variables are involved, the variance is generalized to the covariance, denoted by , and calculated as such:

To give unbiased estimates, sample estimates are obtained by dividing the mean and (co)variance by or , respectively.

Plotting the pairs of points with lines at and , will split the plot into four quadrants similar to the example below:



Which quadrant points lie can tell you whether that point positively or negatively contributed to the covariance. Points in the lower left have both and negative values, so the contribution to covariance is positive. Top right has positive contribution to covariance (since both and are positive), while bottom right and top left have negative contribution to covariance.

This means that as tends to increase as increases, most points will be in the lower left/upper right quadrants, and as tends to decrease as increases (or vice versa), points will be concentrated in the upper left/lower right.

Once covariance has been calculated, we can find the correlation, , of and :

*2.2.2 The ensemble and stationarity*

The mean function of a time series model is the expected value of the variable at the given point in time:

At a given point in time, , there are countless possible values of the variable of interest. The predicted value is simply the mean of all of these possible values, collectively known as the *ensemble* which constitutes the entire population. If we have a time series model predicting values into the future, we can simulate more than one time series. However, with historical data, we only have a single time series to work with. So what we do is estimate the mean sample point using the corresponding observed value.

If the mean function is constant, meaning that the joint probability distribution does not change when shifted in time, we say that the time series model is *stationary* in the mean. Similarly, the mean and variance will remain constant of the time series. Keep in mind that we are talking about the mean/variance of the random component of the variable of interest, and therefore the trend/seasonal variation are taken out of the outcome variable.

*2.2.3 Ergodic series*

A time series model that is stationary in the mean is *ergodic* in the mean if the time average for a single time series tends to the ensemble mean as the length of the time series increases. This essentially means that given enough time and observations, a time series model will hit on all potential points included in the ensemble.

Although you tend to think of time series analysis in regards to things like environmental or economic time series which are single realizations of an underlying model, there are cases in which we can have multiple time series arising from the same model. Think about the acceleration of a particular type of car model, 0-60 mph. Two separate prototypes of the same model may not have the exact same acceleration due to slight differences in manufacture.

*2.2.4 Variance Function*

The variance of the population (i.e., ensemble) can be estimated usinge the sample variance and the equation you’ve seen before:

As you might expect, this could actually underestimate the variance when you have short time series, due to the fact that sequential observations tend to be correlated with one another (and thus clustered closer together than what you might expect from an independent random variable). However, this does not present a huge issue since the bias decreases rapidly as increases.

*2.2.5 Autocorrelation*

As mentioned previously, the correlation a variable has with itself at different times is known as *autocorrelation* or *serial correlation*. Additionally, the model can be described as *second order stationary* if the correlation between variables depends only on the number of time steps (known as the *lag*) separating them.

If a time series model is second-order stationary, we can define an *autocovariance function (acvf)*, , as a function of the lag :

The function does not depend on because the ensemble remains the same across all time periods.

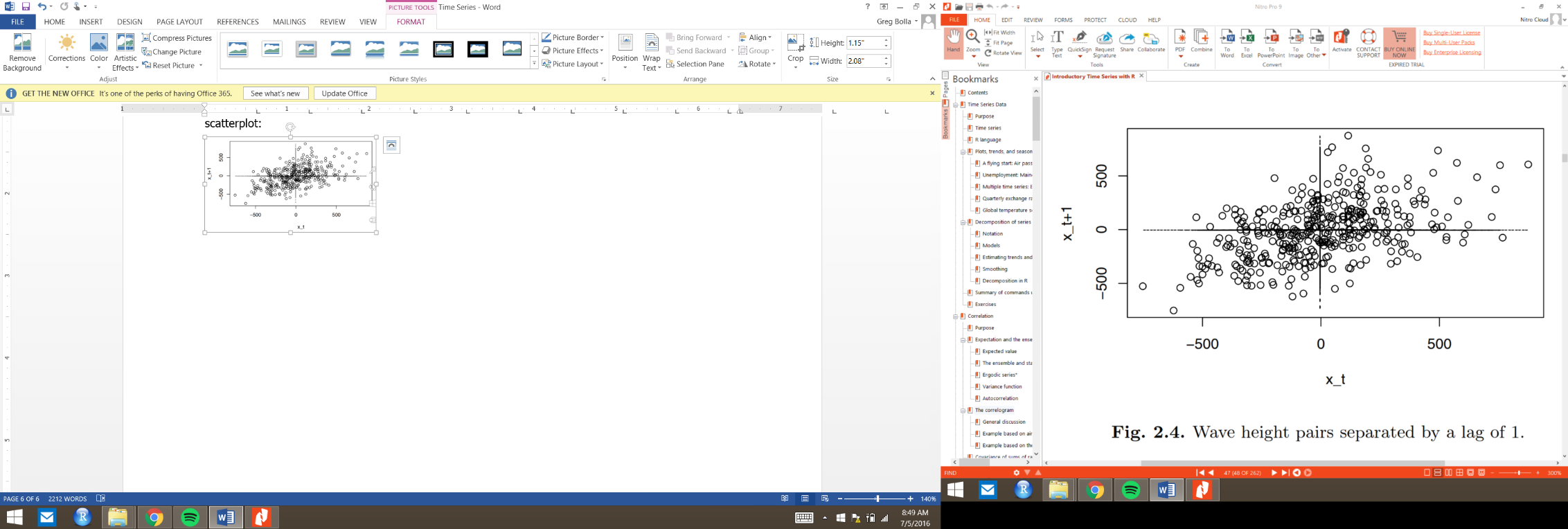
Once you have found the autocovariance, you can then find the *autocorrelation (acf)*, , at a given lag :

Both acvf and acf can be calculated from the sample. First the sample acvf, :

And then the sample acf, , using the sample acvf:

If the lag is 0, the autocorrelation is always 1, which makes intuitive sense.

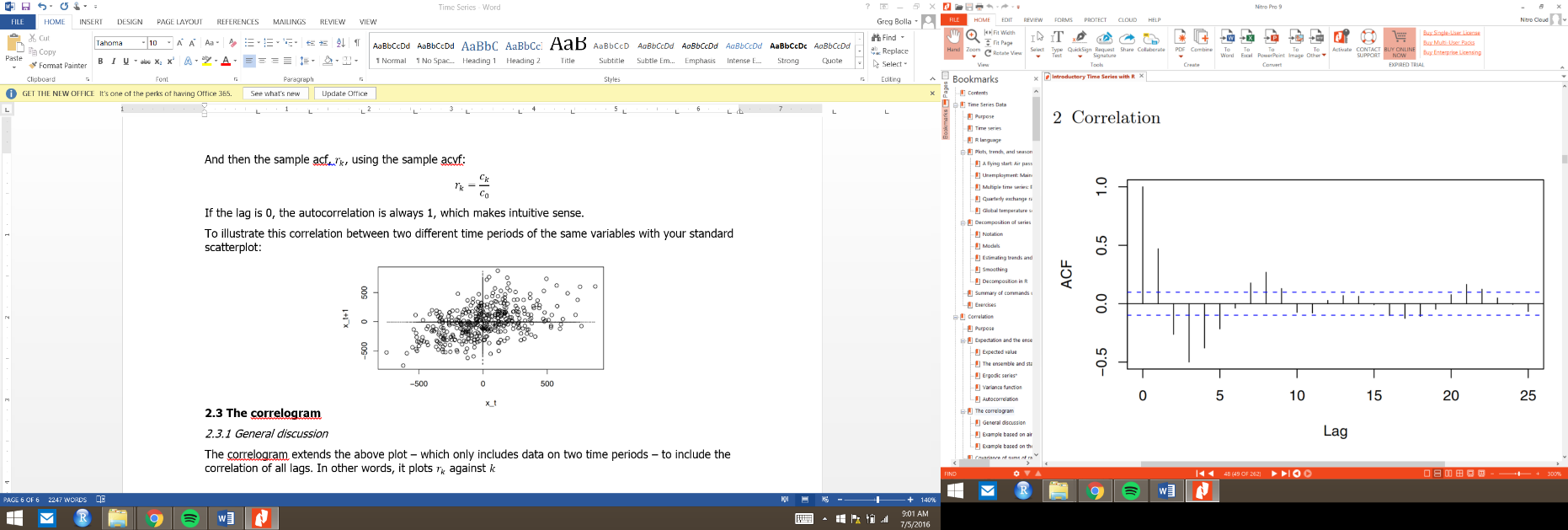
To illustrate this correlation between two different time periods of the same variables with your standard scatterplot:



**2.3 The correlogram**

*2.3.1 General discussion*

The correlogram extends the above plot – which only includes data on two time periods – to include the correlation of all lags. In other words, it plots against , as in the example below:



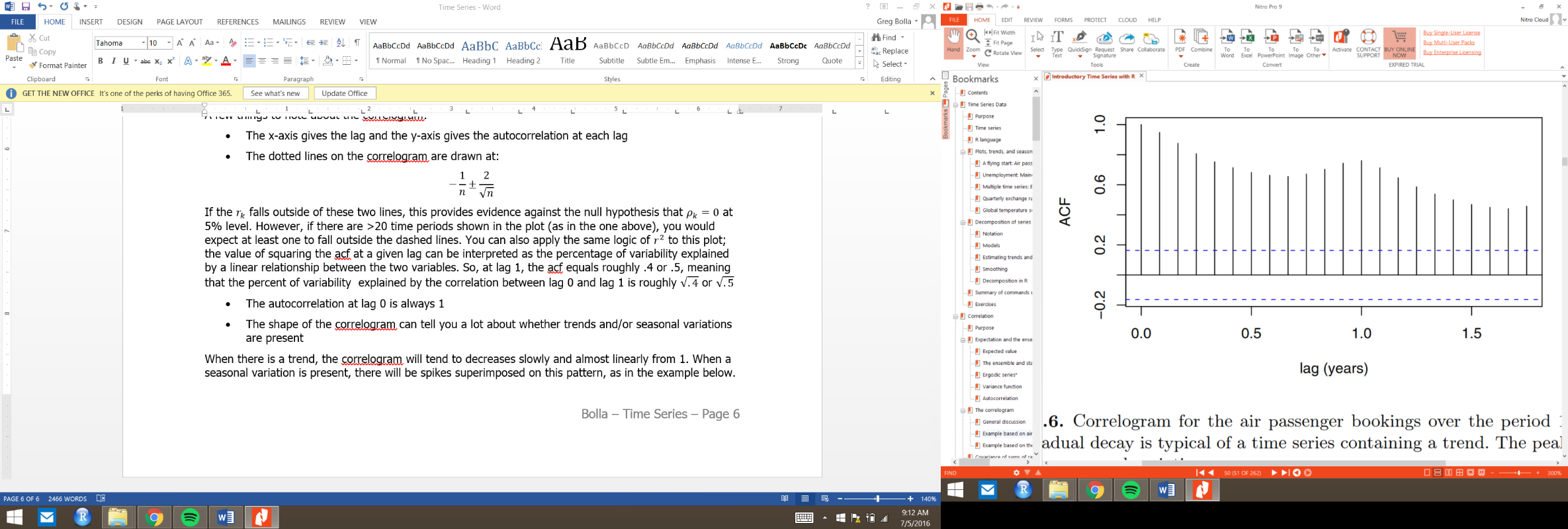
A few things to note about the correlogram:

* The x-axis gives the lag and the y-axis gives the autocorrelation at each lag
* The dotted lines on the correlogram are drawn at:

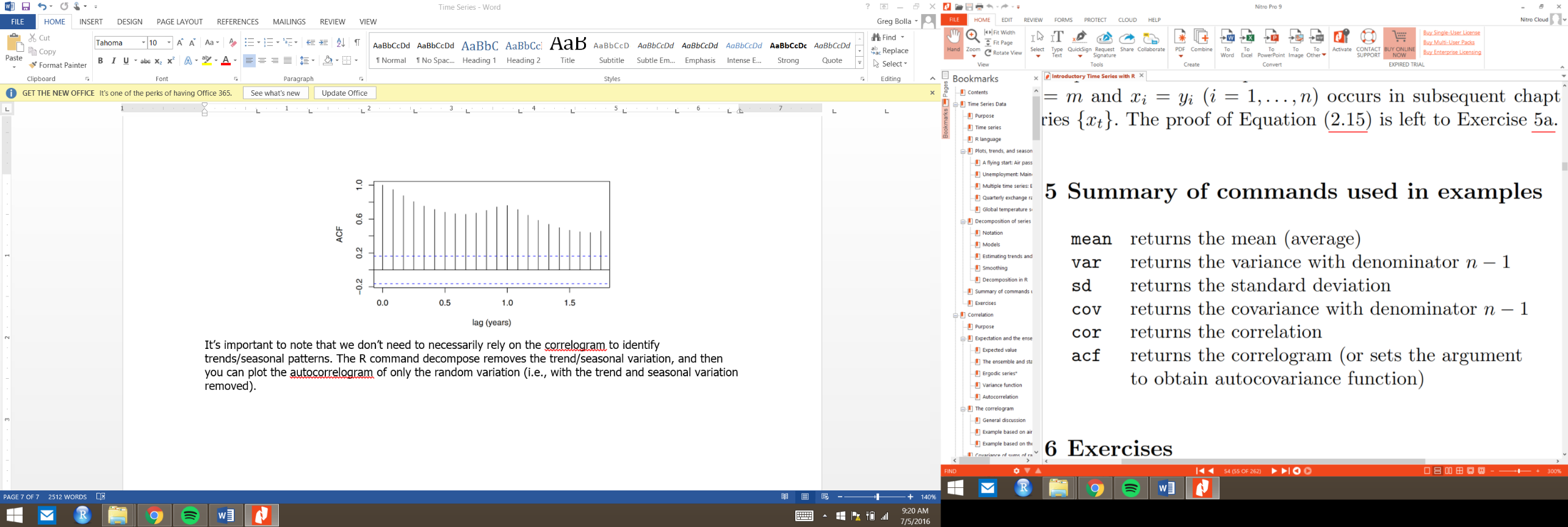
If the falls outside of these two lines, this provides evidence against the null hypothesis that at 5% level. However, if there are >20 time periods shown in the plot (as in the one above), you would expect at least one to fall outside the dashed lines. You can also apply the same logic of to this plot; the value of squaring the acf at a given lag can be interpreted as the percentage of variability explained by a linear relationship between the two variables. So, at lag 1, the acf equals roughly .4 or .5, meaning that the percent of variability explained by the correlation between lag 0 and lag 1 is roughly or

* The autocorrelation at lag 0 is always 1
* The shape of the correlogram can tell you a lot about whether trends and/or seasonal variations are present

When there is a trend, the correlogram will tend to decreases slowly and almost linearly from 1. When a seasonal variation is present, there will be spikes superimposed on this pattern, as in the example below.



It’s important to note that we don’t need to necessarily rely on the correlogram to identify trends/seasonal patterns. The R command decompose removes the trend/seasonal variation, and then you can plot the autocorrelogram of only the random variation (i.e., with the trend and seasonal variation removed).

**Useful R commands – Chapter 2**

**Chapter 3 – Forecasting Strategies**

**3.1 Purpose**

One of the more efficient and accurate ways of forecasting one variable is to find a related variable that leads it by one or more time intervals.

However, it is sometimes tricky/impossible to find a leading variable for the one you are investigating. The book goes into detail on two approaches that are popular when this is the case. The first is the Bass model, which I won’t go into much detail since its main application is in sales/marketing. The second is the so-called Holts-Winter method, which makes extrapolations based on present trends continuing, and implements adaptive estimates for these trends.

**3.2 Leading variables and associated variables**

*3.2.2 Building approvals publication*

The book uses an example from the Australian Bureau of Statistics (ABS) to explain the concepts around leading variables. Essentially, the ABS tracks building approvals and building activity; as you would imagine, if you’re trying to forecast building activity, building approvals would be a good candidate for a leading variable.

To quantify how much of relationship there is between a leading / lagging variable, we can use the *cross-correlation function*. To plot this relationship, we can use a *cross-correlogram*. Both concepts are similar to what was explained in the Chapter 2.

Cross-Correlation

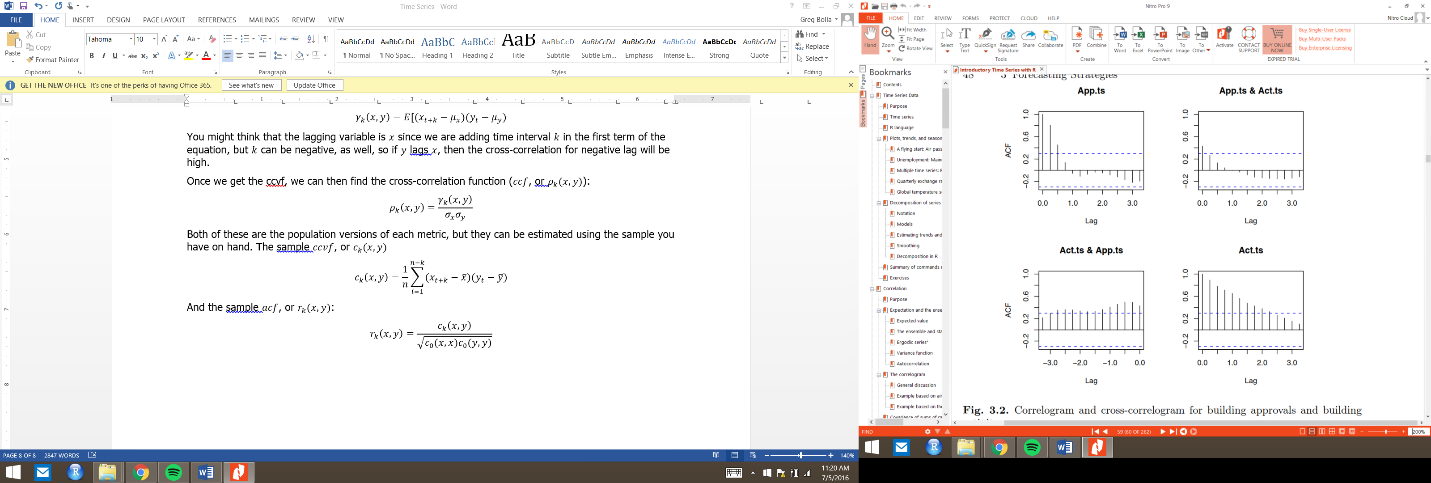
As we did with acvf, we need to define the cross covariance function ( or ), between two variables as a function of the lag.

You might think that the lagging variable is since we are adding time interval in the first term of the equation, but can be negative, as well, so if lags , then the cross-correlation for negative lag will be high.

Once we get the ccvf, we can then find the cross-correlation function (, or ):

Both of these are the population versions of each metric, but they can be estimated using the sample you have on hand. The sample , or

And the sample , or :

The function in R ccf produces the following cross-correlogram for building approvals and building activity:

Note how much of the negative lag between activity and approvals (lower left) are higher than what we would consider the null hypothesis of at a 5% significance level.

**3.3 Bass Model**

The Bass Model quantifies a theory of adoption and diffusion of a new product by society. While it would be useful in some problems, I don’t exactly see how it would apply to economic time series data so I’m going to skip this section.

**3.4 Exponential smoothing & the Holt-Winters method**

*3.4.1 Exponential smoothing*

Exponential smoothing, and the Holt-Winters method built using this concept, could be useful in economic series data. What it does it use historical data – with more emphasis on recent historical data – to extrapolate a trend out into the future. Note that this method assumes no systematic tend or seasonal effects (or alternatively, they have been removed).

A value of a variable at a given point in time (denoted by ) can be described as such:

where is the non-stationary mean of the process at time and are independent random deviations with a mean 0 and standard deviation

R uses the notation in place of so the book does the same throughout the rest of this section.

Given all this, a reasonable estimate of the mean at time is a weighted average of our observation at time and our estimate of the mean at time :

where

In this case, is known as the *exponentially weighted moving average* at time , and the value of is known as a smoothing parameter. If is close to 1, there is little smoothing and is approximated . When is close to 0, there is considerable smoothing and doesn’t take the most recent observation into account very much.

Athough there are different methods of finding , R does it automatically by minimizing the sum of squared one-step-ahead prediction errors. The one-step-ahead prediction error, , is calculated as such:

Ussing this above equation, we can re-write to incorporate a proportion of this prediction error:

\*Note the (very slight) difference between a, and alpha,

We can extend all of this to include more than just the most recent observation, employing the below equation:

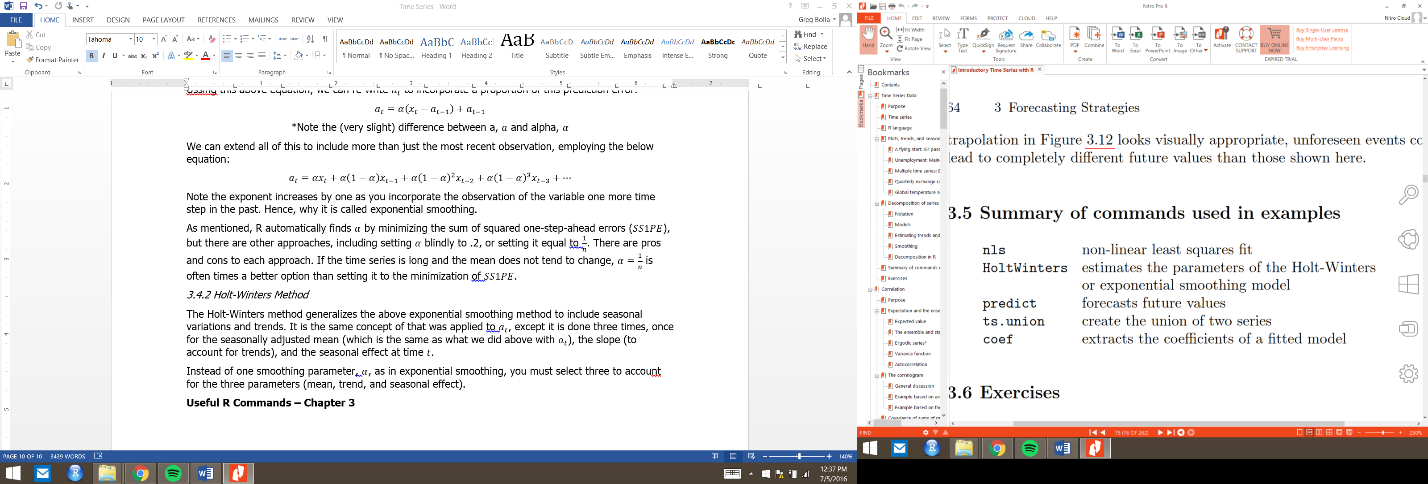
Note the exponent increases by one as you incorporate the observation of the variable one more time step in the past. Hence, why it is called exponential smoothing.

As mentioned, R automatically finds by minimizing the sum of squared one-step-ahead errors (), but there are other approaches, including setting blindly to .2, or setting it equal to . There are pros and cons to each approach. If the time series is long and the mean does not tend to change, is often times a better option than setting it to the minimization of .

*3.4.2 Holt-Winters Method*

The Holt-Winters method generalizes the above exponential smoothing method to include seasonal variations and trends. It is the same concept of that was applied to , except it is done three times, once for the seasonally adjusted mean (which is the same as what we did above with ), the slope (to account for trends), and the seasonal effect at time .

Instead of one smoothing parameter, , as in exponential smoothing, you must select three to account for the three parameters (mean, trend, and seasonal effect).

**Useful R Commands – Chapter 3**