Reconnect 2017 Module:
Time Series Applications in Energy-Related Issues
Instructor Version

Eugene Fiorini
Truman Koehler Professor of Mathematics
Muhlenberg College, PA

Kevin Shirley
Associate Professor and Director of Actuarial Science
Appalachian State University, NC

September 21, 2018

Abstract

This module introduces undergraduates to the statistical field of time series through a series of applications to energy-related examples. Topic 1 introduces the basic terminology of time series and the concepts of smoothing, weighted and moving averages. Topic 2 covers stationary series and introduces the reader to the autocorrelation function and the role it plays in building time series models. Topics 3 focuses on forecasting using regression and autoregressive (AR) models. Topic 4 uses regression to estimate solutions to a differential equation to model resource extraction. The class activities and discussions throughout the module are designed to be in-class activities. Additional student activities are designed as additional homework that can be assigned by the instructor.

Preliminary Material

Description

This module introduces upper level undergraduates to the basic concepts of time series through a series of applications to energy-related examples. The module is organized into four sections (Topics). Topic 1 introduces the basic terminology of time series as well as the concepts weighted and moving averages and their application to the technique of smoothing. Topic 2 introduces the reader to the autocorrelation function and the role it plays in building time series models. The autocorrelation function is also used to describe (weak) stationary and non-stationary series. Topics 3 and 4 focus on forecasting energy prices and resource production using AR and differential equation models. Each type of model is introduced through a series of activities designed to build understanding through discovery.

The module contains two types of activities: class activities/discussions and student activities. The class activities/discussions throughout the module are designed to be completed in class. Students are guided through a series of steps intended to help the student discover the concepts. The student activities are designed to be completed outside of the classroom setting and can be given as additional assignments by the instructor.

Module Summary

In this module, the student will explore applications related to energy prices and production. The data used is time dependent. The models may include deterministic and stochastic variables. Time series methods will be introduced and used to construct the time dependent models and regression will be used to obtain parameter values. Model variables will be tested for significance.

In 1956, a geoscientist named M. King Hubbert, predicted that peak oil production in the continental US would peak between 1965 and 1970. To make this prediction, Hubbert used the logistic model and historical US oil production data. Hubbert prediction was accurate for the time, although could not have anticipated significant improvements in oil extraction technology, such as hydraulic fracturing. Students will apply a similar technique to coal production data ending at various time periods in order to make predictions about coal production and peak production. Students can then observe the market and hydraulic fracturing effects on such predictions.

Fuel prices play an important role in our everyday lives. For motorists, most observe the price of gasoline while driving into work. Homeowners are constantly reminded of the retail prices of electricity, heating oil, and natural gas when the bill arrives. The prices affect our budgeting, savings, decision to commute or not, and the appliances we purchase. Businesses and industry also make decisions based on resource prices and availability. For some, this information is so vital that they devote resources to modeling and forecasting these items. For example, forecasting energy prices and usage plays an important role in economic analysis performed by today's energy providers. Local energy providers may want to forecast high usage days by their customers ahead of time to ensure and plan energy delivery. Students will gain some insight into this activity by exploring models used to forecast retail fuel prices.

Target Audience

The target audience is upper division students with some basic knowledge of inferential statistical procedures and methodologies.

Prerequisites

Prerequisites include a course in probability and statistics that include topics such as properties of probability, hypothesis testing, maximum likelihood, ANOVA, and regression.

Mathematical Fields

Statistics, Probability, Calculus, Discrete Mathematics.

Application Areas

The results of this module could be applied to understand production of energy resources and electricity and its distribution over time. It could also be applied to make predictions for future production and demand of energy resources and electricity.

Goals and Objectives

The main goal of this module is to improve students' mathematical and statistical modeling skills through critical analysis of time series data associated with energy production and usage.

The objectives are to improve students' ability to

- identify trends and seasonal adjustments in correlated data;
- explain and construct a time series model that accurately represents the data;
- use the model to predict short-term trends;
- identify deviations and how additional data points impact the time series model.

Technology/Software Needs

The open software R and RStudio, the open source and enterprise-ready professional software for R. It is possible to use Excel in place of R in some parts of the module. Instructions for using R, RStudio, and Excel are included in appendices.

The Module

Introduction

Statistical procedures rely on certain assumptions being met to produce valid results. For example, common assumptions when applying a two-sample t-test include random sampling, normality of data distribution, adequacy of sample size, and equality of variance in standard deviation. The importance of the random sampling assumption relates to efforts to obtain a sample that is as representative of the population as possible. Random sampling helps minimize (eliminate) systematic bias.

More accurately, to justify statistical procedures through mathematical theorems (and thus, have some confidence in the interpretation of results), these theorems rely on random sampling. A random sample can be thought of as a sequence of independent, identically distributed (iid) random variables.[1] That is, random sampling and independent, identically distributed (iid) are basically the same. It is more common to use the phrase random sample in statistics; whereas it is more common to use iid in probability. Upper case Roman letters are generally used to represent random variables in statistics. Therefore, we represent a sequence of iid random variables by

$$X_1, X_2, X_3, \ldots \tag{1}$$

The purpose of this module is to discuss how to analyze a sequence of random variables when they are time dependent and not independent, identically distributed. That is, given a sequence of random variables $X_1, X_2, X_3, \ldots$, if for each $t \in \mathbb{N}$, X_t is dependent on some or all of the random variables $X_1, X_2, X_3, \ldots, X_{t-1}$, how do we statistically analyze the data if we cannot assume iid?

There are many instances in which the sequence of random variables is not independently, identically distributed. Common examples include such data as economic forecasts, census analysis, and utility studies, among many others. Other examples include forecasting natural phenomena such as weather and earthquakes, as well as signal processing and pattern recognition.

Class Discussion: Time Series Analysis

The examples given above (economic forecasts, census analysis, utility studies, natural phenomena forecasts) all have something in common in how the data is collected. Before reading ahead, spend some class time discussing how data is collected for these and other examples. What characteristics do the data from these examples have in common? For instance, if we are interested in measuring economic factors to determine how the economy is performing, what would you suggest to collect meaningful data? Can you think of other examples of data with similar characteristics?

Notes to Instructor. *This opening discussion is intended to introduce students to the concept of time series analysis. There are two characteristics usually associated with time series data. The first is that in all these examples, as well as others, the data is time dependent. Thus, time-dependent data (data measured over identified time intervals) satisfies the statement "for each $n \in \mathbb{N}$, X_n is dependent on some or all of the random variables $X_1, X_2, X_3, \ldots, X_{n-1}$." Another characteristic of time-dependent data is the time intervals over which the data are collected tend to be of equivalent length. This helps simplify the mathematical analysis used to prove time series techniques. Some other examples include monitoring vital health factors in patients such as heart rate, blood pressure, vitamin levels, etc.; public support of politicians or public policy issues; additional economic factors such as new housing starts or consumer confidence; and transportation factors such as peak highway traffic periods, record of on-time flight takeoffs, or train arrivals.*

[1] http://www.math.uah.edu/stat/sample/

The study of time-dependent data represents an area of statistics called time series. Time series analysis techniques take into consideration the time-dependence of the data in order to create a model fitted to the historical data and to make a forecast or prediction of future data values. These forecasts can be point predictions or interval predictions for a future value. If a forecast is made at time $t-1$ for X_t then it is called a lead - 1 forecast.

Glossary of Terms

There are a few terms that will be useful when working through the activities in this module. Readers who are familiar with these terms can proceed directly to the next section that introduces the model.

A **time series** is defined as a sequence of an indexed set of random variables, $\{X_i : i \in \mathbb{N}\}$ where the index set is based on successive equally-spaced time intervals. For example, a patient's blood pressure may need to be measured twice a day (every 12 hours) over the course of two months while taking a specific medication. **Time series analysis** consists of those techniques and methodologies applied to a time series to extract meaningful information from the data. **Time series forecasting** is the application of a time series model to data in order to predict future data values based on previous observations.

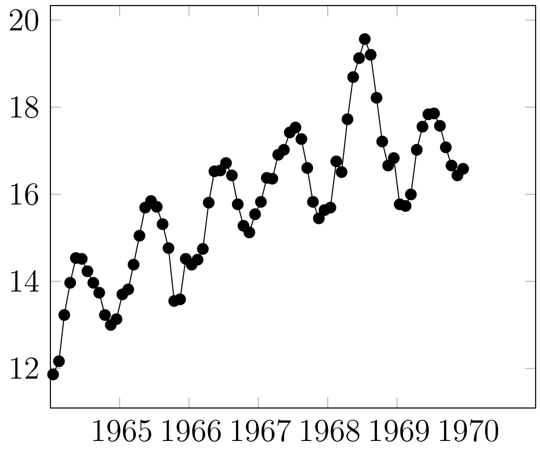
Because time series are based on temporal indexing, a common method of representing time series is graphically. This suggests there is a relationship between time series analysis and regression analysis. **Regression analysis** is a statistical process measuring the relationship between and dependent variable and one or more independent variables. However, there are some critical differences between time series analysis and regression analysis.

Class Discussion. *Before reading ahead, spend some class time discussing what some of the critical differences are between time series analysis and regression analysis. The relationship between regression analysis and time series analysis is further explored later in the module.*

Notes to Instructor. *One critical difference is, once again, the time index. Regression analysis does not take time dependence into account in its analysis. Although it is possible to fit a regression curve to a time series, this model assumes "independence" between the various data points. This does not mean a regression curve is useless when analyzing a time series. It could provide some insight into the data. For example, if the mean of the data is a deterministic function of time, $E[X(t)] = \mu(t)$, then regression may be useful in determining this curve. Common examples of such deterministic functions include the modeling of linear trends, periodic trends, and exponential trends in the data. Furthermore, depending on reasons for the analysis, a regression model might be sufficient for the problem at hand.*

A **mathematical or statistical model** is an abstract representation that uses mathematical or statistical language and concepts to describe the behavior of a system. Mathematical models are used in the natural sciences (physics, biology, earth science), engineering sciences (computer science, artificial intelligence), and the social sciences (economics, psychology, sociology, political science). A model can be used to help explain different components of a system and make predictions about its future. Mathematical models can take many forms: dynamical systems, statistical systems, differential equations, graphical systems, and game theoretic models, as well as many others. Models are also identified in terms of whether there is an element of randomness or not. **Deterministic** models can be described as models in which the output is completely determined by parametric values and initial conditions. That is, there is no element of random variation. **Stochastic** models, on the other hand, possess some inherent randomness. The same parametric values and initial conditions could lead to different outcomes.

Graphical representations of times series are a common means of presenting the data. A **time series plot** simply is the variable plotted against the appropriate time intervals. Below is a time series plot (2) of Lake Huron levels between January 1965 and December 1970.[2] Time series plots can provide a quick assessment of potential trends or seasonal patterns in some time series datasets.



(2)

Regression and time series models will produce a *predicted value* for the dependent variable $\hat{y}$. The *observed value* y in most cases will not exactly equal the predicted value. The **residual** e is the difference between the observed value and the predicted value.

$$e = y - \hat{y} \tag{3}$$

This module concentrates on a few of the more commonly used time series models. These models include the **moving average** (MA) **model**, the **autoregressive** (AR) **and model and autocorrelation function** (ACF), and the **autoregressive integrated moving average** ($ARIMA$) **model**. These models relate the present value to past values and past prediction errors.

[2] https://datamarket.com/data/set/22pw/monthly-lake-erie-levels-1921-1970

Seasonality is a characteristic of time series which regular variations recur in the data in fixed periodic intervals. Common periodic intervals include yearly, quarterly, monthly, and daily.

Topic 1: Elementary Time Series - Smoothing

We begin our discussion of elementary time series and moving average models with a class activity and discussion. The example uses the mathematical notion of average to introduce the concept of a moving average. Recall that the standard arithmetic average of t values $\{x_1, x_2, \ldots, x_t\}$ is given by

$$\bar{x} = \frac{1}{t}\sum_{i=1}^{t} x_i = \frac{x_1 + x_2 + \cdots + x_t}{t}$$

A **weighted average** of t values $\{x_1, x_2, \ldots, x_t\}$ is given by

$$\omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_t x_t$$

where

$$\omega_1 + \omega_2 + \cdots + \omega_t = 1$$

We will make use of these basic definitions of average in the following activity.

Class Activity: Comparing Moving Weighted Averages

Table 1 represents United States Coal Production in millions of short tons (2000 lbs) from 1997 through 2016[3].

Year	Coal (tons)	Year	Coal (tons)	Year	Coal (tons)	Year	Coal (tons)
1997	1,089.90	2002	1,094.30	2007	1,146.60	2012	1,016.50
1998	1,117.50	2003	1,071.80	2008	1,171.80	2013	984.80
1999	1,100.40	2004	1,112.10	2009	1,074.90	2014	1,000.05
2000	1,073.60	2005	1,131.50	2010	1,084.40	2015	896.90
2001	1,127.70	2006	1,162.70	2011	1,095.60	2016	728.20

Table 1: US Coal Production 1997-2016

1. Construct a time series plot of the data. Describe some trends or patterns in the data.

2. Table 2 contains the running averages of US coal production for consecutive years. That is,

$$\frac{x_t + x_{t-1}}{2} \text{ for } t = 1998, 1999, \ldots, 2016$$

[3] *Independent Statistics & Analysis*, U.S. Energy Information Administration, https://www.eia.gov/

Years	Average	Years	Average	Years	Average	Years	Average
1997-98	1103.7	2002-03	1083.05	2007-08	1159.2	2012-13	1000.65
1998-99	1108.95	2003-04	1091.95	2008-09	1123.35	2013-14	992.425
1999-2000	1087	2004-05	1121.8	2009-10	1079.65	2014-15	948.475
2000-01	1073.6	2005-06	1147.1	2010-11	1090	2015-16	812.55
2001-02	1094.3	2006-07	1154.65	2011-12	1056.05		

Table 2: US Coal Production Two-Year Running Average

On the same graph as the original data, construct a time series plot of the two-year running average data points found in table 2.

3. Construct a table consisting of the three year running averages (and average of length 3) for the data in table 1. On the same graph as the previous two plots, construct a time series plot of these “running averages of length 3” as well.

$$\frac{x_t + x_{t-1} + x_{t-2}}{3} \text{ for } t = 1999, 2000, \ldots, 2016$$

4. Compare the three time series plots. What observations can you make? In particular what do you notice between the time series plot of the original data and the 3-point average time series plot? If we continued to increase the number of points used in calculating this ”running” average, what do you would imagine the time plots would look like?

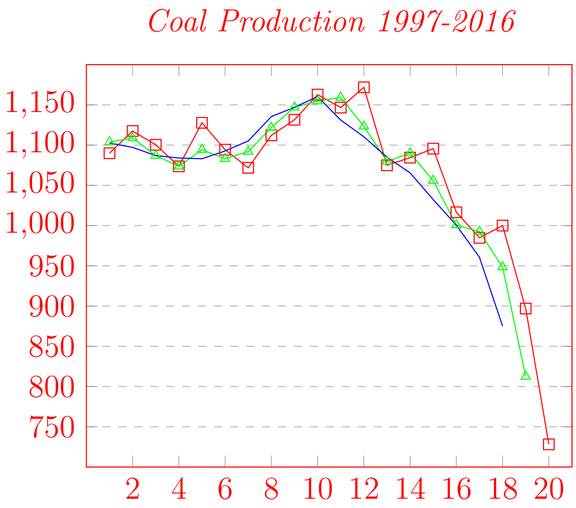
Notes to Instructor. *Some potential answers and comments to the class activity.*

1. *The students should recognize that coal production was relatively at the same levels from 1997 through 2008. Around 2008 it appears coal production began dropping off dramatically. Some discussion about what could have caused the drop starting around 2008 could include such items as the development of new technology (fracking and thus cheaper natural gas), improved efficiency with burning coal, or moving to alternative fuel sources. The time series plot is the red plot below with red squares.*

2. *The time series plot for the average of two years is included in the plot below (green triangles).*

3. *The time series plot for the average of length 3 is included in the plot below (blue).*

4. *The students should recognize that by taking increasingly larger averages, the plots begin to appear to have a ”smoothing” effect. Note that the average of length 3 plot appears to “smooth out” the time series plot of the original data. The concept of a moving or running average being used to smooth data in order to better observe trends in discussed after the activity.*



Smoothing Time Series with Weighted Averages

The previous activity demonstrates the concept of **smoothing** in time series. Smoothing is a type of *filter* that helps reveal trends or patterns in the data. A common method of smoothing is moving averages. A **moving average** is commonly described as a (possibly weighted) average calculated for each time point using observed values that surround each of those values[4]. For example, the class activity above asked you to calculate the average of length 3 for each time period using the previous two time periods x_{t-2} and x_{t-1} in the calculation:

$$\frac{x_t + x_{t-1} + x_{t-2}}{3} = \mu_t$$

We could have also calculated a moving average of length 3 using the previous observed value x_{t-1} and the next observed value x_{t+1},

$$\frac{x_{t-1} + x_t + x_{t+1}}{3} = \mu_t^*$$

In fact, we could have just as easily calculated a weighted average of length 3

$$\omega_{t-1}x_{t-1} + \omega_t x_t + \omega_{t+1}x_{t+1} = \mu_t^{**}$$

The previous example suggests weighted averages play a role in smoothing away seasonality in data. Seasonality can obscure critical trends and patterns within time series data. Smoothing away seasonality allows for easier identification of these trends. Applying a weighted average to time series data is one method of smoothing away the seasonality. However, how do we know which weighted average works best with a given time series dataset?

[4] https://onlinecourses.science.psu.edu/stat510

Class Activity: Smoothing with Weighted Averages

Tables 3 and 4 represent the total monthly electricity generated (in thousands of megawatt hours) in the United States from January 2011 (2011-01-15) through December 2016 (2016-12-15)[5]. This dataset is represented graphically in the time series plot (4). The time series plot shows the quantities in millions of megawatt hours.

1. Study the time series plot and data for a few moments. What is the seasonality? Is the seasonality monthly? Quarterly? Semi-annually? Yearly? That is, describe how electricity usage in the United States fluctuates seasonally. Justify your conclusion.

2. Before applying weighted averages to smooth away the seasonality, what sort of trends would you suggest exist in this data? That is, beyond the seasonal fluctuations, has electricity usage in the United States increased on average over the past 6 years? Has it decreased? Has electricity usage remained constant over the past 6 years?

3. We will compare several different moving averages to gain some insight into which type of moving average works best with this dataset. Construct a table of values for each weighted average using R or Excel. Then plot the resulting values on a time series plot.

$$\frac{1}{7}x_{t-3} + \frac{1}{7}x_{t-2} + \frac{1}{7}x_{t-1} + \frac{1}{7}x_{t} + \frac{1}{7}x_{t+1} + \frac{1}{7}x_{t+2} + \frac{1}{7}x_{t+3}$$

$$\frac{1}{12}x_{t-3} + \frac{1}{6}x_{t-2} + \frac{1}{6}x_{t-1} + \frac{1}{6}x_{t} + \frac{1}{6}x_{t+1} + \frac{1}{6}x_{t+2} + \frac{1}{12}x_{t+3}$$

$$\frac{1}{13}\sum_{i=t-6}^{t+6} x_i$$

$$\frac{1}{24}x_{t-6} + \frac{1}{12}\sum_{i=t-5}^{t+5} x_i + \frac{1}{24}x_{t+6}$$

4. Compare the four time series plots. Which of the four plots best smooths away the seasonality in the data? Justify your response.

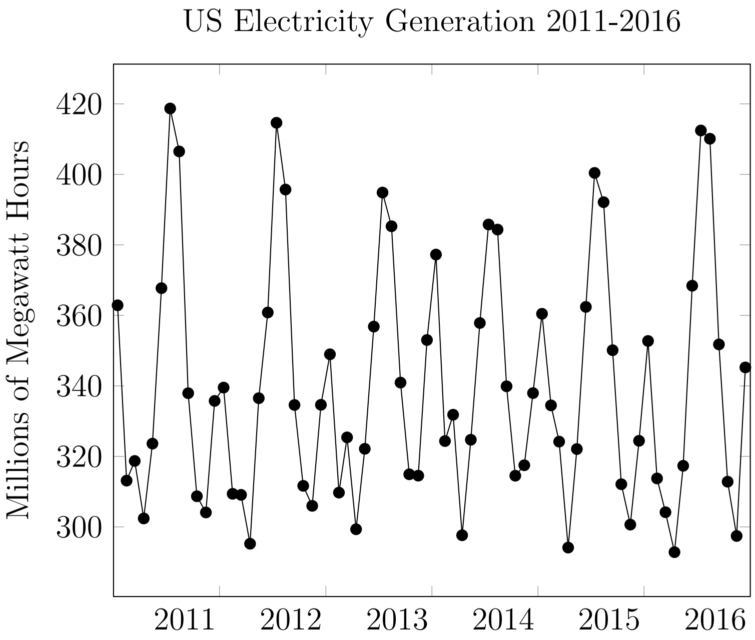
[5] *Independent Statistics & Analysis*, U.S. EIA, https://www.eia.gov/electricity/data/

Month	mwh	Month	mwh	Month	mwh
2011-01-15	362872	2012-01-15	339526	2013-01-15	348967
2011-02-15	313127	2012-02-15	309389	2013-02-15	309728
2011-03-15	318710	2012-03-15	309090	2013-03-15	325399
2011-04-15	302401	2012-04-15	295229	2013-04-15	299333
2011-05-15	323628	2012-05-15	336516	2013-05-15	322156
2011-06-15	367727	2012-06-15	360825	2013-06-15	356823
2011-07-15	418693	2012-07-15	414641	2013-07-15	394846
2011-08-15	406511	2012-08-15	395700	2013-08-15	385286
2011-09-15	337931	2012-09-15	334586	2013-09-15	340941
2011-10-15	308699	2012-10-15	311652	2013-10-15	314925
2011-11-15	304102	2012-11-15	305976	2013-11-15	314540
2011-12-15	335740	2012-12-15	334635	2013-12-15	353021

Table 3: US Electricity Generation 2011-2013

Month	mwh	Month	mwh	Month	mwh
2014-01-15	377255	2015-01-15	360455	2016-01-15	352745
2014-02-15	324348	2015-02-15	334476	2016-02-15	313749
2014-03-15	331823	2015-03-15	324192	2016-03-15	304168
2014-04-15	297631	2015-04-15	294133	2016-04-15	292836
2014-05-15	324724	2015-05-15	322087	2016-05-15	317337
2014-06-15	357844	2015-06-15	362409	2016-06-15	368418
2014-07-15	385780	2015-07-15	400419	2016-07-15	412450
2014-08-15	384341	2015-08-15	392116	2016-08-15	410113
2014-09-15	339887	2015-09-15	350122	2016-09-15	351769
2014-10-15	314522	2015-10-15	312112	2016-10-15	312828
2014-11-15	317495	2015-11-15	300653	2016-11-15	297427
2014-12-15	337957	2015-12-15	324427	2016-12-15	345238

Table 4: US Electricity Generation 2014-2016



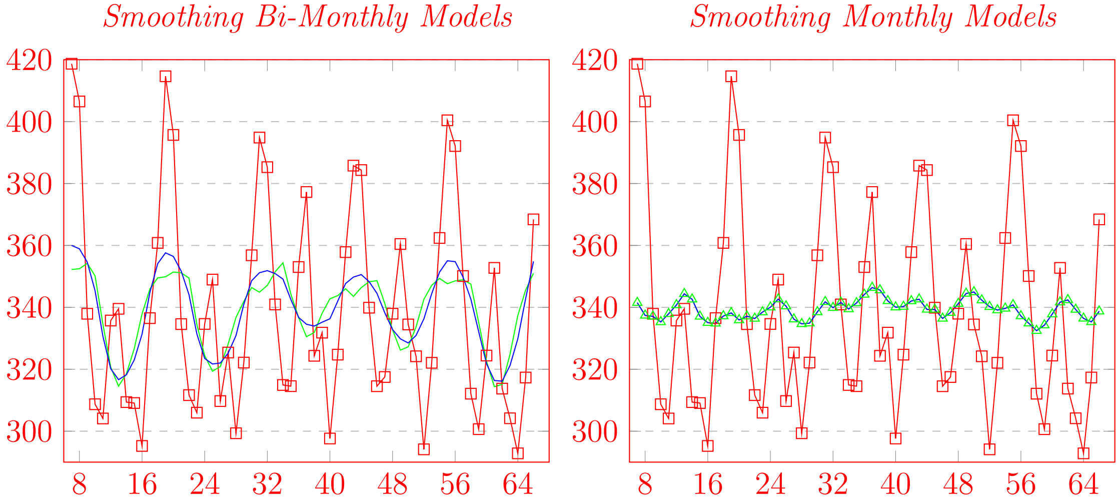
(4)

Notes to Instructor. *Some potential answers and comments to the class activity.*

1. *The students should recognize the data is monthly with a yearly seasonality. It could be argued that the seasonality is semi-annually with the "troughs" (of about equal depth - occurring in fall and spring) occurring every six months. However, the peaks in winter and summer are of different heights. The winter peaks suggest greater electricity usage occurs in the winter.*

2. *There is seasonal fluctuation. However, peaks and valleys are of roughly the same values each year over the past 6 years. That is, electricity usage appears to be relatively constant over the past six years, suggesting a steady (flat) trend. Smoothing away seasonality should result in a relatively flat trend line.*

3. *A plot of each of the weighted averages appears below.*



The time series plots that smooth away the seasonality best are the monthly averages.

$$\frac{1}{13}\sum_{i=t-6}^{t+6} x_i$$

$$\frac{1}{24}x_{t-6}+\frac{1}{12}\sum_{i=t-5}^{t+5} x_i+\frac{1}{24}x_{t+6}$$

In this case, the weighted average (that weighs the "outside" months less) is usually the standard model.

Note that the monthly models smooth away the seasonality better than the bi-monthly models. Because the data is recorded monthly with a yearly seasonality, the weights of the weighted average are based on a cycle of 12. The typical formula used to smooth away seasonality in quarterly data is based on a cycle of 4.

$$\frac{1}{8}x_{t-2}+\frac{1}{4}x_{t-1}+\frac{1}{4}x_t+\frac{1}{4}x_{t+1}+\frac{1}{8}x_{t+2}$$

Student Activity: Smoothing with Weighted Averages - Quarterly Natural Gas Prices

Table 5 represents the quarterly average residential natural gas prices (in US dollars per million cubic feet) in the United States from January 2011 (2011Q1) through December 2016 (2016Q4)[6].

Quarter	Price	Quarter	Price	Quarter	Price
2011Q1	11.16	2013Q1	9.74	2015Q1	9.66
2011Q2	13.43	2013Q2	12.56	2015Q2	12.43
2011Q3	17.48	2013Q3	16.93	2015Q3	17.02
2011Q4	11.47	2013Q4	10.34	2015Q4	10.38
2012Q1	10.44	2014Q1	10.2	2016Q1	8.79
2012Q2	12.98	2014Q2	13.56	2016Q2	11.4
2012Q3	16.2	2014Q3	17.48	2016Q3	17.31
2012Q4	10.79	2014Q4	10.88	2016Q4	10.29

Table 5: US Quarterly Average Residential Natural Gas Price 2011-2016

1. Construct a time series plot for the data in Table 5.
2. Describe the trends and seasonality that exists in the data.
3. Construct a table of values for each weighted average using R or Excel. Plot the resulting values on a time series plot. Compare the plots. Which of the plots best smooths away the seasonality in the data? Justify your response.

[6] *Independent Statistics & Analysis*, U.S. Energy Information Association, https://www.eia.gov/

$$\frac{1}{8}x_{t-2} + \frac{1}{4}x_{t-1} + \frac{1}{4}x_t + \frac{1}{4}x_{t+1} + \frac{1}{8}x_{t+2}$$

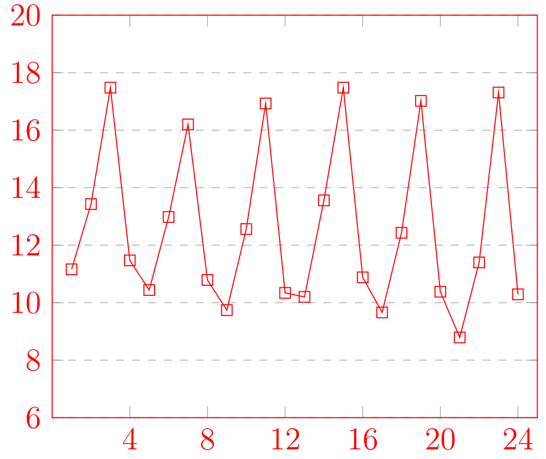
$$\frac{1}{5}\sum_{i=t-2}^{t+2} x_i$$

$$\frac{1}{4}\sum_{i=t}^{t+3} x_i$$

Notes to Instructor. *Some potential answers and comments to the student activity.*

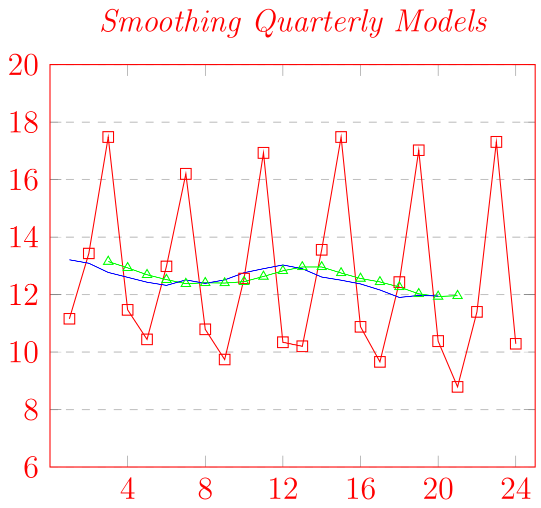
1. *A time series plot of the data is*

Time series plot of natural gas prices



2. *The trend for this time series is roughly constant with an approximate mean value of* 12.6. *The seasonality is clearly quarterly.*

3. *A plot of each of the weighted averages appears below.The green curve represents the weighted average. The blue curve represents the standard average. Both suggest reasonable smoothing of the data, perhaps the weighted average providing a slightly better representation.*



Student Activity: Smoothing with Weighted Averages - Hydroelectric

Table 6 represents the monthly electricity generated by Hydroelectric power (in thousand megawatt hours) in the United States from January 2013 (13-Jan) through December 2016 (16-Dec)[7].

Month	Price	Month	Price	Month	Price	Month	Price
13-Jan	24828.53	14-Jan	21633.79	15-Jan	24138.38	16-Jan	25426.41
13-Feb	20418.47	14-Feb	17396.13	15-Feb	22286.08	16-Feb	24149.73
13-Mar	20534.36	14-Mar	24257.13	15-Mar	24280.9	16-Mar	27024.9
13-Apr	25097.1	14-Apr	25439.91	15-Apr	22470.98	16-Apr	25475.33
13-May	28450.09	14-May	26543.89	15-May	20125.42	16-May	25362.38
13-Jun	27384.07	14-Jun	25743.88	15-Jun	20414.08	16-Jun	22902.14
13-Jul	27254.57	14-Jul	24357.4	15-Jul	21014.22	16-Jul	21246.81
13-Aug	21633.32	14-Aug	19807.25	15-Aug	19122.11	16-Aug	19359.06
13-Sep	16961.15	14-Sep	16074.33	15-Sep	16094.12	16-Sep	16280.94
13-Oct	17198.59	14-Oct	17159.21	15-Oct	16630.4	16-Oct	17248.91
13-Nov	17676.83	14-Nov	18624.92	15-Nov	19337.83	16-Nov	18814.83
13-Dec	21128.3	14-Dec	22328.79	15-Dec	23165.56	16-Dec	22537.89

Table 6: Monthly US Electricity Produced by Hydroelectric 2013-2016

1. Construct a time series plot for the data in Table 6.

2. Describe the trends and seasonality that exists in the data.

3. Construct a table of values for each weighted average using R or Excel. Plot the resulting values on a time series plot. Compare the plots. Which of the plots best smooths away the seasonality in the data? Justify your response.

[7] *Independent Statistics & Analysis*, U.S. Energy Information Association, https://www.eia.gov/

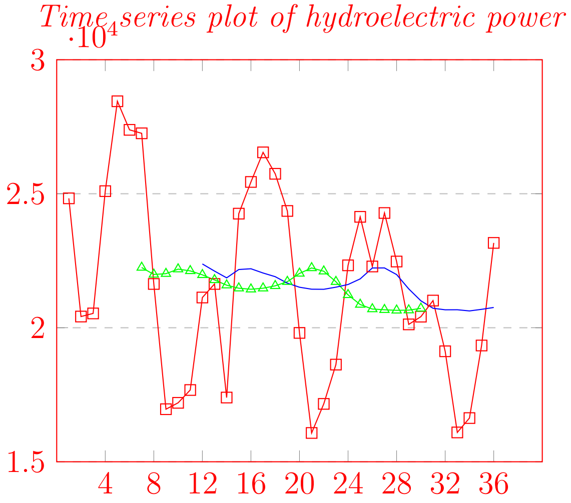
$$\frac{1}{24}x_{t-6} + \frac{1}{12}\sum_{i=t-5}^{t+5} x_i + \frac{1}{24}x_{t+6}$$

$$\frac{1}{13}\sum_{i=t-6}^{t+6} x_i$$

$$\frac{1}{12}\sum_{i=t}^{t+11} x_i$$

Notes to Instructor. *Some potential answers and comments to the student activity.*

1. *A time series plot of the data is*



2. *The time series has a slight downward trend over the two years. There appears to be an annual seasonality in the data.*

3. *A plot of each of the weighted averages appears in the above plot.The green curve represents the weighted average. The blue curve represents the standard average. Both suggest a better smoothing model of the data might exist. This could be due to the slightly downward trend in the data. This will be discussed in more detail in the other topics.*

Topic 2: Stationary Series and the Autocorrelation Function

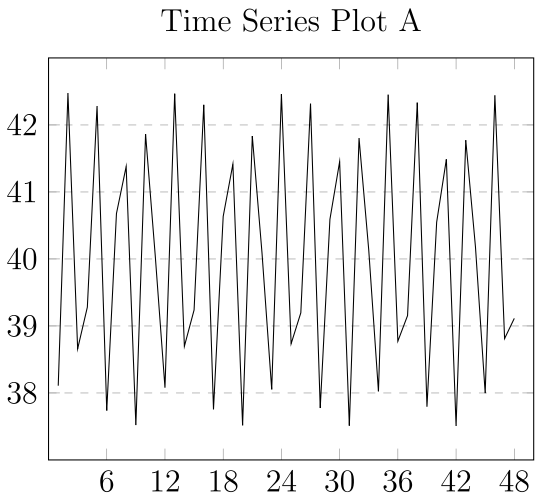
The previous section demonstrated the effect of weighted averages in smoothing away seasonality to better observe trends and patterns in time series data. The weighted averages can help give a description of the time series data, but more work needs to be done if we are to construct models that allow for inferences and predictions. We begin this process by introducing the autocorrelation function (ACF) and (weakly) stationary series.

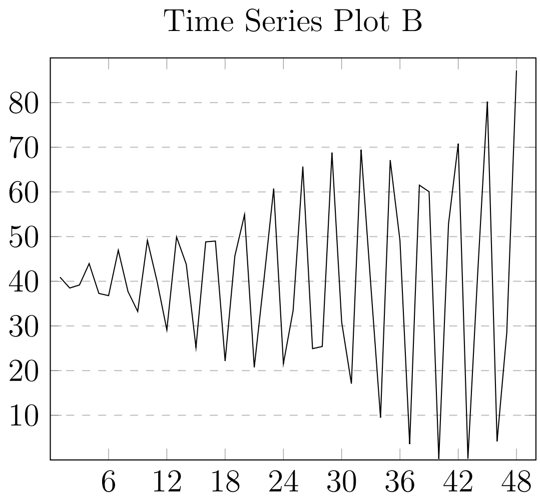
Class Activity: Stationary Series

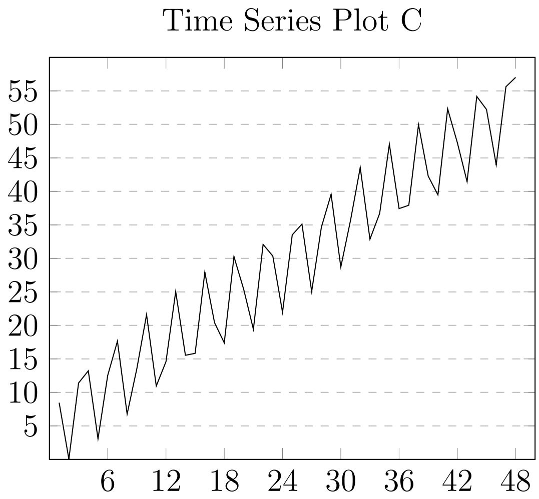
The diagram below shows the plots of three different time series. In this activity we will examine basic descriptive measures such as mean and variance. Recall that a time series can be described as a sequence of random variables $X_1, X_2, X_3, \ldots$ in which random variable X_t is dependent on some subset of the random variables $\{X_1, X_2, \ldots, X_{t-1}\}$. Note that for each time value t, random variable X_t has a distribution with mean μ_t and variance σ_t^2.

1. Examine the plots closely. What can you conjecture about the mean and standard deviation for X_t in each time series plot? For which plots is the mean relatively constant? For which plots does the mean seem to vary among the random variables $X_1, X_2, X_3, \ldots$? For which plots is the variance relatively constant? For which plots does the variance seem to vary among the random variables $X_1, X_2, X_3, \ldots$?

2. Recall that the covariance is the *joint* variability between two variables. Examine the three time series plots again. What can be conjectured about the covariance between random variables X_t and X_{t-h} for some $h \geq 1$? That is, are there some values of h for which the variability of random variables X_t and X_{t-h} are related (i.e. $Cov(X_t, X_{t-h}) \neq 0$)? Are there some values of h for which the variability of random variables X_t and X_{t-h} are not related (i.e. $Cov(X_t, X_{t-h}) = 0$)?







Notes to Instructor. *The following are some possible answers to the class activity.*

1. *Time series plot A shows a mean and variance that remain relatively constant throughout the time interval. There are clearly seasonal trends; however, a calculation of an overall mean would roughly produce a horizontal line. Likewise, variance remains relatively constant, once seasonality is taken into consideration. Time Series Plot B also shows a mean that remains relatively constant throughout the time interval. However, the time series plot shows more variability as time increases. This suggests the variance in this time series is not constant. In this case, variance increases over time. Time Series Plot C, on the other hand shows a relatively constant variance (the difference between minimum and maximum values remains constant); however, the mean tends to increase over time.*

2. *This question may require more scrutiny. Covariance can be determined by examining seasonality in the data. All three time series plots demonstrate seasonality of one form or another. For example, time series plot A shows a roughly yearly seasonality. This*

would suggest $Cov(X_t, X_{t-12}) \neq 0$ indicating a relationship between the variabilities of X_t and X_{t-12}. This may not be the only non-zero covariance.

3. *The point of this activity is to introduce the student to the concept of stationary series. Time series plot A suggests a stationary series, while Time series plots B and C do not. Class discussion could include references to the properties desirable in a stationary time series.*

The previous activity demonstrates some of the differences that occur with time series data. Time series Plot A suggests a scenario that, when seasonality is taken into account, is relatively "stable" throughout the time interval. That is, the mean and variance remain relatively constant. Time series plot B shows a relatively constant mean, but the data clearly increases in variance over the time interval. In the case of time series plot C, the variance is relatively stable, but the mean increases over the time interval. All three series show strong indications of seasonality. That is, there is correlation between X_t and X_{t-h} for some $h \geq 1$. Correlation between elements from the same series separated by a given interval h is called **autocorrelation**.

Class Activity: The Role of Trends and Seasonality

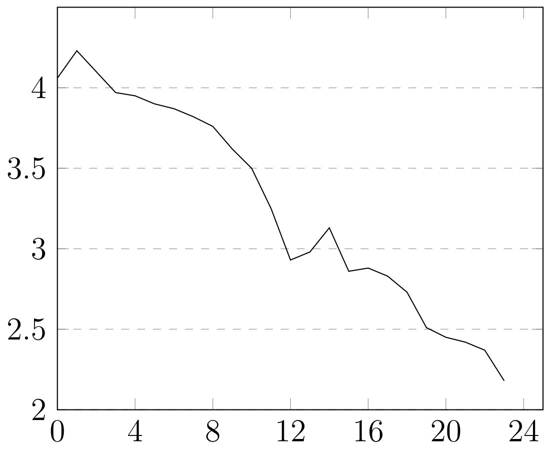
To better define and understand autocorrelation, we will look more closely at the role of trends and seasonality in time series data. The time series data in table (7) and the corresponding time series plot (5) represents the average monthly retail heating oil prices between January 2014 and December 2015 (in dollars per gallon)[8].

Month	Price	Month	Price	Month	Price	Month	Price
Jan 2014 ($t = 0$)	4.06	Jul-14	3.87	Jan-15	2.93	Jul-15	2.73
Feb-14	4.23	Aug-14	3.82	Feb-15	2.98	Aug-15	2.51
Mar-14	4.1	Sep-14	3.76	Mar-15	3.13	Sep-15	2.45
Apr-14	3.97	Oct-14	3.62	Apr-15	2.86	Oct-15	2.42
May-14	3.95	Nov-14	3.5	May-15	2.88	Nov-15	2.37
Jun-14	3.9	Dec-14	3.25	Jun-15	2.83	Dec 2015 ($t = 24$)	2.18

Table 7: Heating Oil Average Monthly Retail Prices

[8] "Short-term Energy Outlook Real and Nominal Prices," *Independent Statistics and Analysis*, U.S. Energy Information Administration, June 2017

Heating Oil Average Monthly Price



(5)

1. Study table (7) and time series plot (5). Describe the trend most likely associated with the time series data.

2. For the purposes of this activity, we will assume the time series data is quarterly. Use the weighted average model

$$W_t = \frac{1}{8}X_{t-2} + \frac{1}{4}(X_{t-1} + X_t + X_{t+1}) + \frac{1}{8}X_{t+2}$$

to calculate the values for $t = 2, 3, \ldots, 21$. Plot the resulting weighted averages $\{W_t : t = 2, 3, \ldots, 21\}$ on a graph.

3. For $t = 2, 3, \ldots, 21$, calculate and plot

$$Y_t = X_t - W_t$$

Examine the time series plot of $\{Y_t : t = 2, 3, \ldots, 21\}$. What was the effect of subtracting the weighted average W_t from the corresponding data point X_t?

4. To calculate the seasonal effect, we first re-index the data points. We are assuming seasonality is quarterly. With 24 data points, this gives 6 seasonal cycles. Let $k = 1, 2, 3, 4$ and $j = 0, 1, 2, 3, 4, 5$. Re-index $t = 0, 1, 2, \ldots, 23$ such that $t = t_{k,j}$ whenever $t = 6(k-1) + j \pmod 6$.

With each seasonal cycle, the mean for that cycle increases. That is, if M_j is the mean for cycle j, $j = 0, 1, 2, 3, 4, 5$, then $M_0 < M_1 < M_2 < M_3 < M_4 < M_5$. Calculate the mean estimate for each cycle:

$$\hat{M}_j = \frac{1}{4}(X_{1,j} + X_{2,j} + X_{3,j} + X_{4,j})$$

To estimate the seasonal effect, we will calculate averages across the data for a fixed place in each cycle. That is, if $k = 1$ (the start of each cycle), then we calculate $X_{1,j} - \hat{M}_j$ for $j = 1, 2, 3, 4, 5$ (every fourth data point) and average the values.

(a) Why do we subtract $\hat{M}_j$ from $X_{1,j}$?

(b) Why do we average every fourth data point for the seasonal component?

For $k = 1, 2, 3, 4$, estimate the four seasonal components

$$\hat{S}_k = \frac{1}{9}[(X_{k,1} - \hat{M}_1) + (X_{k,2} - \hat{M}_2) + \cdots + (X_{k,6} - \hat{M}_6)] = \frac{1}{6}\sum_{j=1}^{6}(X_{k,j} - \hat{M}_j)$$

If S_t is the seasonal component at time t, explain why $S_t = S_{t+4}$ for $t = 0, 1, 2, \ldots, 19$.

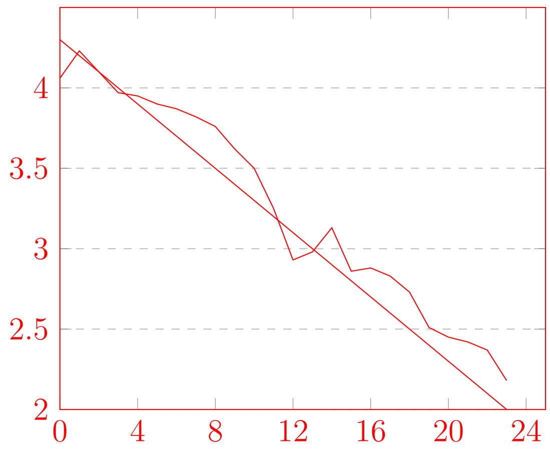
5. Part (3) eliminated the trend component to the time series data. The next step is to remove the seasonal component from the data. We do this by subtracting $\hat{S}_k$ from each corresponding data point. If Z_t represents what is left after removing the trend and the seasonal components, then

$$Z_t = Z_{k,j} = Y_t - \hat{S}_k = Y_{k,j} - \hat{S}_k = X_{k,j} - W_t - \hat{S}_k$$

Calculate $Z_t = Z_{k,j}$ for each $k = 1, 2, 3, 4$ and $j = 1, 2, 3, 4, 5$ ($t = 2, 3, \ldots, 21$). Plot the resulting time series Z_t. After the trend and seasonal components are removed from the data points, the plot of Z_t is not 0. Explain why there appears to be "residual" left over after subtracting the trend and seasonality effects.

Notes to Instructor. *The following are some possible answers to the class activity.*

1. *The trend appears to be linear and decreasing. An approximate equation for the trend line is* $y = -0.1x + 4.3$.

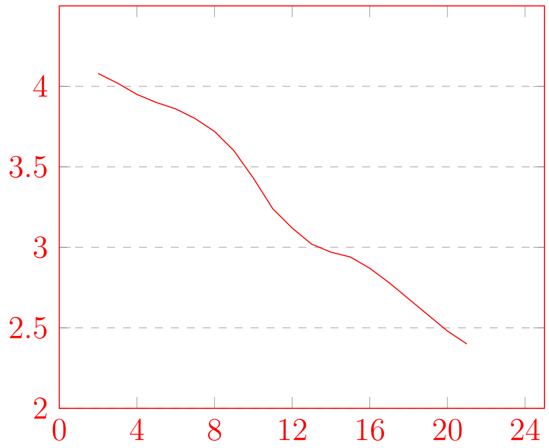


Heating Oil Prices with Trend

2. The table of weighted averages and corresponding plot appear below.

t	*ave.*	*t*	*ave.*	*t*	*ave.*	*t*	*ave.*
		6	*3.86*	12	*3.12*	18	*2.68*
		7	*3.8*	13	*3.02*	19	*2.58*
2	*4.08*	8	*3.72*	14	*2.97*	20	*2.48*
3	*4.02*	9	*3.6*	15	*2.94*	21	*2.4*
4	*3.95*	10	*3.43*	16	*2.87*		
5	*3.9*	11	*3.24*	17	*2.78*		

Table 8: Weighted Averages

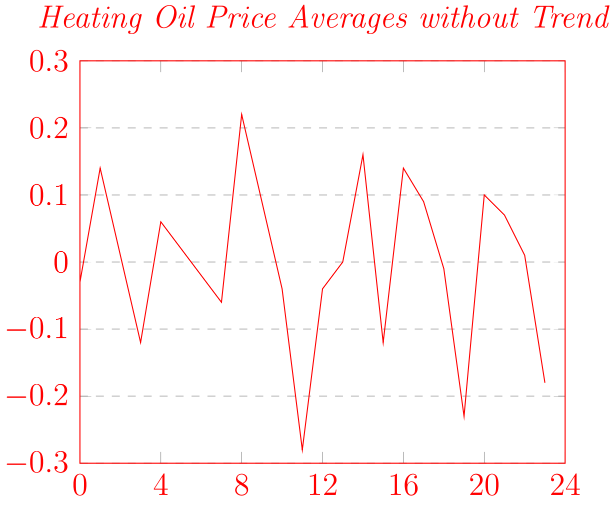


Heating Oil Prices Weighted Averages

3. *The effect of subtracting W_t from the data is the trend has been removed. Note the plot is no longer increasing. Note also that removing the trend effects brings out the seasonality.*

t	*diff.*	*t*	*diff.*	*t*	*diff.*	*t*	*diff.*
0	*-0.03*	6	*-0.02*	12	*-0.04*	18	*-0.01*
1	*0.14*	7	*-0.06*	13	*0*	19	*-0.23*
2	*0.01*	8	*0.22*	14	*0.16*	20	*0.10*
3	*-0.12*	9	*0.09*	15	*-0.12*	21	*0.07*
4	*0.06*	10	*-0.04*	16	*0.14*	22	*0.01*
5	*0.02*	11	*-0.28*	17	*0.09*	23	*-0.18*

Table 9: Trend Removed



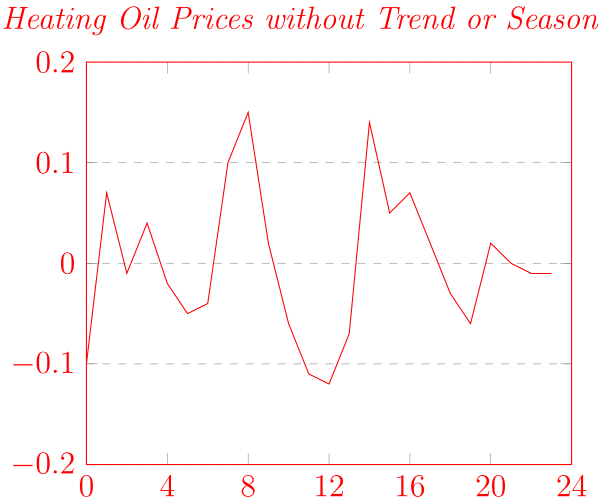
4. $M_0 = 4.09,\ M_1 = 3.89,\ M_2 = 3.53,\ M_3 = 2.97,\ M_4 = 2.74,\ M_5 = 2.35.$

   (a) *Subtracting M_j from each $X_{1,j}$ removes the trend created within each cycle so that what remains is only seasonality.*

   (b) *Every fourth point is used in each seasonal component average because it fixes the data points at the same place in the cycle. The expectation is that the effect of seasonality is the same at the equivalent place in the cycle.*

   $\hat{S}_1 = 0.08,\ \hat{S}_1 = 0.07,\ \hat{S}_1 = 0.02,\ \hat{S}_1 = -0.16.$ *We are assuming the seasonality is taken to be quarterly. This implies that $\hat{S}_t$ and $\hat{S_{t+4}}$ use the same values in calculating the average.*

5. *A plot of Z_t appears below.*



Trend and seasonality are the main components making up the data, but there is always some variation that remains. The question is whether there is more "pattern" to the data or if What remains is just "noise" in the data. By eliminating the trend and seasonality, the remains can be studied for effects of autocorrelation. This will be discussed in the next section. Generally, we write

$$X_t = W_t + S_t + \varepsilon_t$$

This class activity demonstrates how time series can be decomposed into component parts consisting of the trend W_t, the seasonality S_t, and random noise ε_t.

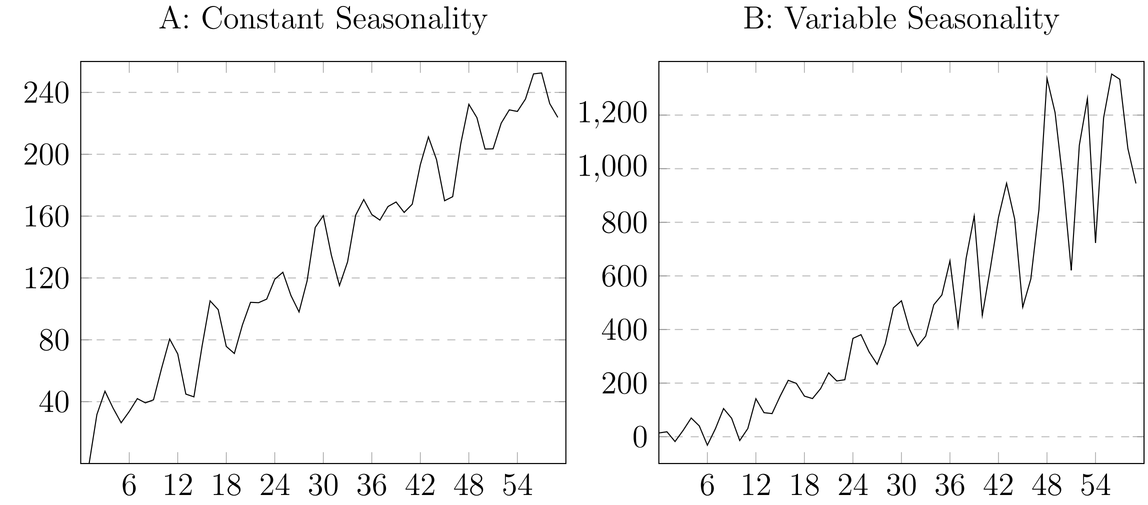
$$X_t = W_t + S_t + \varepsilon_t \tag{6}$$

We will summarize the previous discussion about seasonality and trend by presenting more formalized descriptions of these concepts.

- **Seasonality** (S_t) consists of patterns that repeat over a fixed time interval.
- The **trend** (W_t) consists of the underlying metrics of the data, increasing or decreasing over time.
- The **random noise** (ε_t) is the residual after the data information has been allocated to the trend and the seasonality.

Expression 6 represents an additive decomposition of a time series. This is not the only possible method of decomposition of a time series. Expression 6 is appropriate in situations where the seasonal variation remains constant throughout the time line (see diagram A in (8)). In the case where seasonality varies through the time line (diagram B in (8)), a more appropriate model is a multiplicative decomposition of the time series:

$$X_t = W_t S_t \varepsilon_t \tag{7}$$



(8)

Both the additive and multiplicative models decompose into a trend component (W_t), a seasonality component (S_t), and a residual (random noise) component (ε_t). Finding an appropriate representation for W_t, S_t, and ε_t is critical to constructing a reasonable model X_t for the time series. In the previous example we assumed seasonality was quarterly. Was this a reasonable assumption?

One way to test if the time series model is appropriate is to test if the residuals no longer contain any trace of a pattern. That is, after fitting data to a time series the residuals should be "white noise." They should have no autocorrelation. If there is autocorrelation in the residuals, this is an indication that the model is wrong.

We formally define the **autocorrelation function** between X_t and X_{t-h} as

$$ACF(X_t, X_{t-h}) = \frac{Cor(X_t, X_{t-h})}{StDev(X_t)StDev(X_{t-h})} \tag{9}$$

In the case where $StDev(X_t) = StDev(X_{t-h})$, we have

$$ACF(X_t, X_{t-h}) = \frac{Cor(X_t, X_{t-h})}{Var(X_t)} = \frac{\sum_{i=1}^{n-k} (X_i - \bar{X})(X_{i+k} - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{10}$$

Class Activity: The Role of Trends and Seasonality, Part II - Residuals

The autocorrelation function is commonly employed two different ways with time series. One use is to detect non-randomness in the residuals. That is, if the decomposition has captured all the pattern leaving only white noise in the residuals. The other use is to preemptively identify an appropriate time series model for the data.

When autocorrelation is used to detect non-randomness in the residuals, setting $h = 1$ (the first lag) is usually sufficient.When autocorrelation is used to determine an appropriate time series model, several autocorrelations are calculated (and plotted) for different lag values.

1. Using the heating oil average monthly retail prices data from the previous activity, calculate the lag-one autocorrelations ($h = 1$) for the residuals.

$$ACF(X_t, X_{t-1}) = \frac{Cor(X_t, X_{t-1})}{Var(X_t)} = \frac{\sum_{i=1}^{n-k}(X_i - \bar{X})(X_{i+k} - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

2. Based on your calculations, is the model from the previous class activity reasonable?

3. Calculate several more lags, $h = 1, 2, 3, 4, \ldots$, for the time series. Plot the lags on a graph. What do you think the autocorrelation plot suggests?

Notes to Instructor. *The following are possible answers to the class activity.*

1. *The R command to calculate the first lag is given by*

   $$my\_acf < -acf(dataset\$variable, lag.max = 1)$$

   The first lag is 0.306183

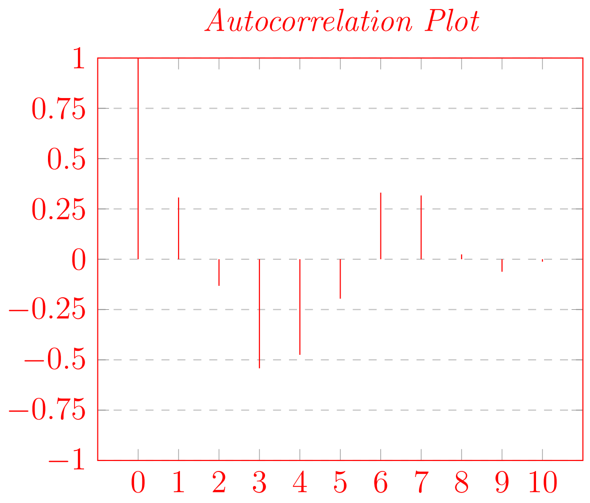
2. *The value of the first lag is small enough to suggest the model from the previous class activity is relatively reasonable. The lag value is not large enough to move to a more complicated model. Enough of the pattern is picked up by the given model to provide a reasonable interpretation.*

3. *The R command to calculate several lags is given by*

   $$my\_acf < -acf(dataset\$variable, lag.max = 10)$$

   Below is a table of the acf values along with an autocorrelation plot.

0	1	2	3	4	5	6	7	8	9	10
1.000	0.307	-0.132	-0.542	-0.475	-0.196	0.331	0.317	0.024	-0.062	-0.120



The autocorrelations taper with some positive and negative values. The second lag (-0.132) is slightly more than the square of the first lag ($0.307^2 = 0.094$, a reasonable rule of thumb that suggests the model is a reasonable fit. More about autocorrelations will be discussed in the next section.

This gives us the information we need to describe the criteria for (weakly) stationary time series. A **stationary time series** is one whose statistical measures, such as mean and variance, are constant over time. That is, a stationary time series satisfies

- $\mu = \mu_t = E[X_t]$ for all t (that is, μ_t is constant for all t)
- $Var(X_t) = \sigma^2$ for all t (that is, σ_t^2 is constant for all t)
- $Cov(X_t, X_{t-h}) = \sigma_h$ for all t and $h = 1, 2, 3, \ldots$ (that is, for each $h = 1, 2, 3, \ldots$, $Cov(X_t, X_{t-h})$ is constant and $ACF(X_t, X_{t-h}) = \frac{\sigma_h}{\sigma^2}$)
- The theoretical value of ACF of a particular lag h is constant across the series. That is, $ACF_t = ACF_{t-h}$ for some h.

Student Activity: Smoothing with Weighted Averages - Natural Gas Prices

This activity uses the time series data on U.S. quarterly U.S. prices of natural gas found in table 5 from January 2011 (2011Q1) through December 2016 (2016Q4).

1. Construct a time series plot for the data in Table 5.
2. Use the weighted average model

$$w_t = \frac{1}{8}x_{t-2} + \frac{1}{4}\sum_{i=t-1}^{t+1} x_i + \frac{1}{8}x_{t+2}$$

to calculate the values for $t = 6, 7, \ldots, n-6$. Plot the resulting weighted averages $\{W_t : t = 6, 7, \ldots, n-6\}$ on a graph.

3. For $t = 6, 7, \ldots, n-6$, calculate and plot

$$y_t = x_t - w_t$$

Examine the time series plot of $\{Y_t : t = 6, 7, \ldots, n-6\}$. Explain the effect of subtracting the weighted average W_t from the corresponding data point X_t. That is, does this account for the trend in the data?

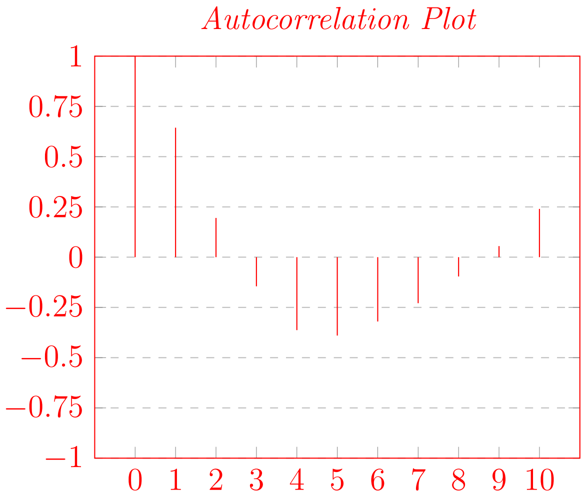
4. Calculate and plot the autocorrelation function. Interpret your results.

Notes to Instructor. *The following are possible answers to the class activity.*

1. *This is the same plot as the previous student activity.*
2. *This is the same result as the previous student activity.*
3. *The answers to this question are similar to the previous class activity. They can be easily calculated using R or Excel.*
4. *The autocorrelation values and plots appear below. The R command is*

$$my\_acf < -acf(dataset\$variable, lag.max = 10)$$

0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1.000	*0.644*	*0.195*	*-0.145*	*-0.363*	*-0.390*	*-0.320*	*-0.229*	*-0.096*	*0.055*	*0.240*



The first lag (0.644 > 0.4) shows some significance. This would suggest that a first order model is not the best choice. This will be discussed further in Topic 3. The autocorrelations due taper off, but show some pattern in the decay. Again, this suggests not all the pattern has been accounted for. More about autocorrelations will be discussed in the next section.

Student Activity: Trends, Seasonality, and ACF - Monthly Wind Electricity

Tables 10 and 11 represent the monthly electricity generated by wind (in thousands of megawatt hours) in the United States from January 2008 through March 2017[9].

	2008	2009	2010	2011	2012
Jan	4273.185	5950.825	6854.337	8550.495	13630.86
Feb	3851.749	5852.175	5431.858	10451.56	11051.71
Mar	4782.02	7099.063	8589.077	10544.65	14027.34
Apr	5225.282	7457.696	9764.456	12421.66	12709.03
May	5340.284	6261.961	8697.525	11772.16	12540.32
Jun	5140.376	5599.422	8049.021	10985.07	11972.16
Jul	4008.401	4954.941	6723.891	7488.629	8823.083
Aug	3264.406	5464.474	6685.855	7473.594	8469.419
Sep	3111.452	4650.708	7105.502	6869.029	8789.865
Oct	4756.401	6813.626	7943.808	10525.43	12635.9
Nov	4993.646	6875.183	9747.619	12438.55	11648.5
Dec	6615.899	6906.058	9059.297	10655.77	14523.52

Table 10: US Electricity Generated by Wind 2008-2012

	2013	2014	2015	2016	2017
Jan	14738.5	17911.21	15162.15	18531.42	20349.6
Feb	14075.59	14008.66	14921.55	20203.51	21691.63
Mar	15755.65	17735.88	15307.93	21979.27	25598.91
Apr	17476.27	18635.55	17867.15	20744.8	
May	16238.7	15601.37	17151.34	18795.47	
Jun	13748.11	15798.82	13421.27	16318.38	
Jul	11093.61	12187.39	13675.45	17594.61	
Aug	9633.884	10170.52	13080.03	13560.62	
Sep	11674.08	11519.77	13971.57	16430.38	
Oct	13635.02	14507.93	16380.04	20380.38	
Nov	15803.26	18866.93	19681.72	19342.23	
Dec	13967.06	14711.25	20098.37	22991.03	

Table 11: US Electricity Generated by Wind 2013-2017

[9] *Independent Statistics & Analysis*, U.S. Energy Information Association, https://www.eia.gov/

1. Construct a time series plot for the data in Tables 10 and 11.

2. Describe the trends and seasonality that exist in the data.

3. Use the weighted average model

$$W_t = \frac{1}{24}X_{t-6} + \frac{1}{12}\sum_{i=t-5}^{t+5} X_i + \frac{1}{24}X_{t+6}$$

   to calculate the values for $t = 6, 7, \ldots, n-6$. Plot the resulting weighted averages $\{W_t : t = 6, 7, \ldots, n-6\}$ on a graph.

4. For $t = 6, 7, \ldots, n-6$, calculate and plot

$$Y_t = X_t - W_t$$

   Examine the time series plot of $\{Y_t : t = 6, 7, \ldots, n-6\}$. Explain the effect of subtracting the weighted average W_t from the corresponding data point X_t. That is, does this account for the trend in the data?

5. Calculate and plot the autocorrelation function. Interpret your results.

Topic 3: Forecasting Fuel Prices - Regression and AR Models

In 1918, the price of gasoline was \$0.25 per gallon. By the time of the great depression, the average price had fallen to \$0.18 – \$0.19 per gallon. In our automobile/commuter based economy, the price of gasoline and diesel fuel continues to affect the everyday lives of most Americans. In this topic, we will explore historical gasoline prices and how to forecast future gasoline prices.

From a historical perspective, nominal gasoline prices have been quite volatile. This volatility may come as no surprise as the price of the major ingredient, crude oil, can be quite volatile. Although crude oil has not been scarce to the present, much of the world's supply comes from the middle east and other less politically stable countries. Prices have often been influenced by cartels and cartel like organizations. The supply of oil on the market has been controlled by the Texas Railroad Commissions in the 1930's to OPEC today. We will explore the relationship of the price of crude oil and price of gasoline.

The 1918 price of gasoline, \$0.25 per gallon, is \$3.92 per gallon inflation adjusted to 2015. Compare this to the average retail price in 2015 of \$2.36 per gallon. As seen in Figure 1, the long term trend is exponential. To understand the inflation adjusted prices, we can use regression to fit an exponential trend to annual average gasoline prices from 1918 to 2015[10]. The model takes the following form,

$$y(t) = a * e^{bt}, t = 0, 1, ..., 97 \tag{11}$$

[10] *Inflation Data*, Inflationdata.com, https://inflationdata.com/articles/inflation-adjusted-gasoline-prices/

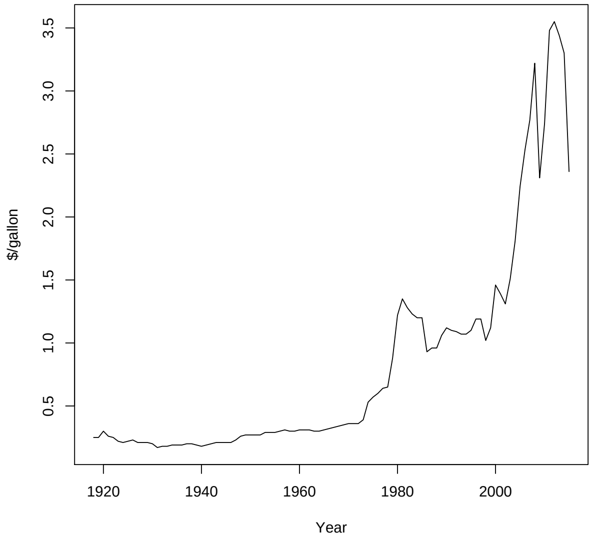


Figure 1: Average Annual Retail Gasoline Prices from 1918

where the $year = 1918 + t$, In order to use least squares linear regression to find the parameters a and b, we first transform the data by taking the natural logarithm of the average annual gasoline prices. By regressing the transformed data against the time variable we obtain parameter estimates for the model,

$$ln(y(t)) = ln(a) + bt \tag{12}$$

Regression yields the parameter estimates,

$$ln(a) = -2.12506 \text{ and } b = 0.0303. \tag{13}$$

Hence, our regression determined exponential model is

$$y(t) = 0.1194 * e^{0.0303t}. \tag{14}$$

Shown in Figure 2, our model indicates about a 3.03% annual continuously compounded growth rate in the nominal price of gasoline since 1918. The actual average annual inflation rate since 1918 is about 3.12% indicating that the average annual price of gasoline since 1918 in inflation adjusted terms has been approximately constant, growing a little less than the rate of inflation.

Class Activity: Using Exponential Trends

The previous discussion shows how average annual growth rates for prices with an exponential trend can be estimated using regression. We will use the method described in the discussion and the data given in Table 12 to compare the growth rates of oil and gasoline over the time period 1945 - 2015. Table 12 gives a sample of the annual average crude oil prices in \$/barrel, $Oil(t)$, every five years beginning in 1948. The table also gives the average annual retail gasoline prices in \$/gallon for the same years.

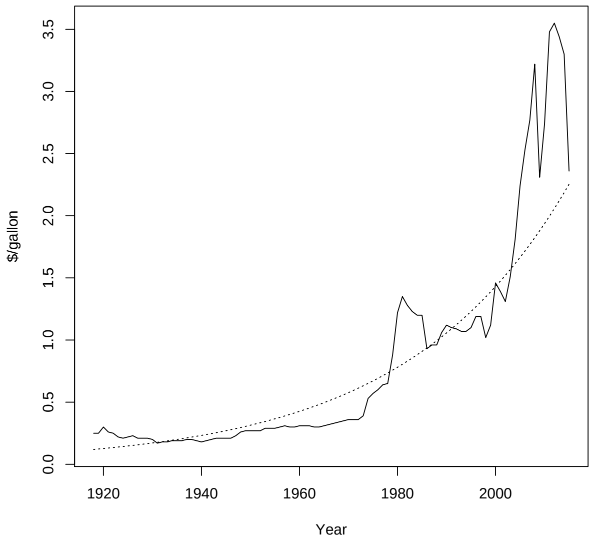


Figure 2: Retail Gasoline Prices and Exponential Model from 1918

Year	$Oil(t)$	$Gasoline(t)$	Year	$Oil(t)$	$Gasoline(t)$
1948	2.77	0.26	1983	29.08	1.23
1953	2.92	0.29	1988	14.87	0.96
1958	3.00	0.30	1993	16.75	1.07
1963	2.91	0.30	1998	11.91	1.02
1968	3.18	0.34	2003	27.69	1.51
1973	4.75	0.39	2008	91.48	3.22
1978	14.95	0.65	2013	91.17	3.44

Table 12: Sample Annual Oil and Gasoline Prices 1945-2015

1. Use linear regression to fit an exponential model to the sample data for $Oil(t)$, where $t = 0$ in 1945.

2. Use linear regression to fit an exponential model to the sample data for $Gasoline(t)$, where $t = 0$ in 1945.

3. The average annual rate of inflation in the US over the period 1945 - 2015 is approximately 3.7%. How does the continuously compounded growth rate of each compare with each other as well as the average inflation rate over the same time period?

Notes to Instructor. *The discussion to the class activity follows.*

1. *We transform the data for $Oil(t)$ in Table(12) and use regression to find the parameters for the model*

$$ln(Oil(t)) = ln(a) + bt + \epsilon$$

Regression yields the estimates,

$$ln(a) = 0.425645 \text{ and } b = 0.055197.$$

Hence, our analysis determines the exponential model is

$$y(t) = 1.530577 * e^{0.055197t}.$$

Our model indicates an approximate continuous rate of increase of about 5.5% annually for the averge price of crude oil in the US since 1945.

2. *We transform the data for Gasoline(t) in Table(12) and use regression to find the parameters for the model*

$$ln(Gasoline(t)) = ln(a) + bt + \epsilon$$

Regression yields the estimates,

$$ln(a) = -1.75169 \text{ and } b = 0.0405436.$$

Hence, our analysis determines the exponential model is

$$y(t) = 1.73481e^{0.0405436t}.$$

Our model indicates an approximate continuous rate of increase of just over 4% annually for the retail price of gasoline since 1945. Notice how the rate of increase of price of gasoline since 1945 has increased above the rate of increase since 1920.

3. *The price of oil is increasing at an annual rate about 1.8% higher than the rate of inflation and the average annual retail price of gasoline is increasing at a rate of just over the rate of inflation since 1945.*

Forecasing Gasoline Prices

In order to discuss forecasting, we will analyze the data over a shorter time horizon with average prices taken weekly. For our discussion, we consider the average weekly retail gasoline prices from the first week in 1995 to the end of 2000[11]. Using a moving average method to estimate trend and seasonal effects, Figure 3 shows the observed values, followed by the trend estimated by a centered moving average, seasonal effects, and the random component with trend and seasonal effects are removed. The seasonal effects on the price of gasoline reflect the changing seasonal demand. In the US, the summer months are going to be the months most heavily traveled, especially when travel is by automobile. Hence, we observe the largest positive seasonal effect on the price of gasoline, 0.0515, during the 26th week of the year (last week in June) . The largest negative seasonal effect is -0.0486 which occurs during the 7th week of the year, mid-February. Although as cubic model for the trend seems like a possibility for this six year time period, we shall see that there is a strong trend determined stochastically.

[11] Weekly US Regular Convential Retail Gasoline Prices," *Independent Statistics and Analysis*, U.S. Energy Information Administration, https://www.eia.gov/

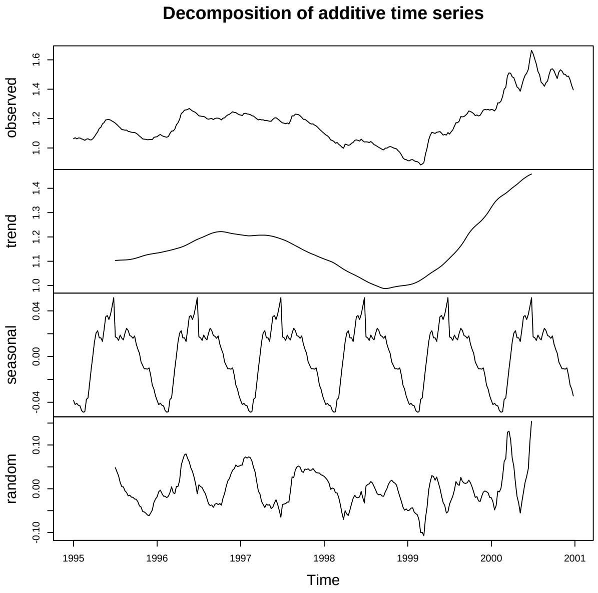


Figure 3: The decomposition of Gasoline Prices into a trend and seasonal effects

If our time series is stationary in the mean, variance, and covariance (2nd order stationary), then the ACF depends only on the time lag h. In this case, we can use the sample autocorrelation function (acf) to estimate ACF. First, we define the sample autocovariance function, c_k

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})$$

Then the sample autocorrelation function is

$$acf(k) = \frac{c_k}{c_0}$$

Values for the ACF and acf are between -1 and 1. A value of $acf(k)$ close 1 indicates a strong positive correlation between the values in the series that are separated by k-units of time. A correlogram is a graphical representation of $acf(k)$ for a finite number of time lags. To illustrate, let's consider the time series for weekly gasoline prices from 1995 through 2000, Figure 4. Notice, the correlogram begins with $acf(0) = 1$. This is included to provide scale for other values.The horizontal dashed line provides the reference for the 95% level of significance. The strong autocorrelation at each lag is indicative of the trend in the data. Recall the population ACF is defined on a second order stationary time series. A time series with a trend is not second order stationary since the mean will vary with time. Therefore, we will need to model the trend first in order to get accurate statistical information from the sample acf.

The trend for a time series is deterministic if we can specify a function of time that models this trend and the time series tends to revert back to the trend over time. Otherwise, a trend is stochastic and does not recover from the random shocks. In viewing the decomposition,

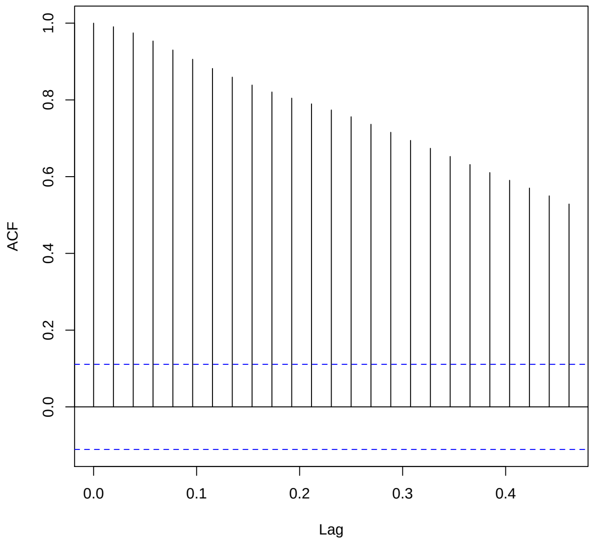


Figure 4: The sample autocorrelation function for weekly Gasoline Prices 1995 through 2000

one may wish to try modeling the trend with a polynomial of odd degree. A stochastic trend can by recognized by first differencing the data. To define the differences, we first define the backshift operator B,

$$By_t = y_{t-1}.$$

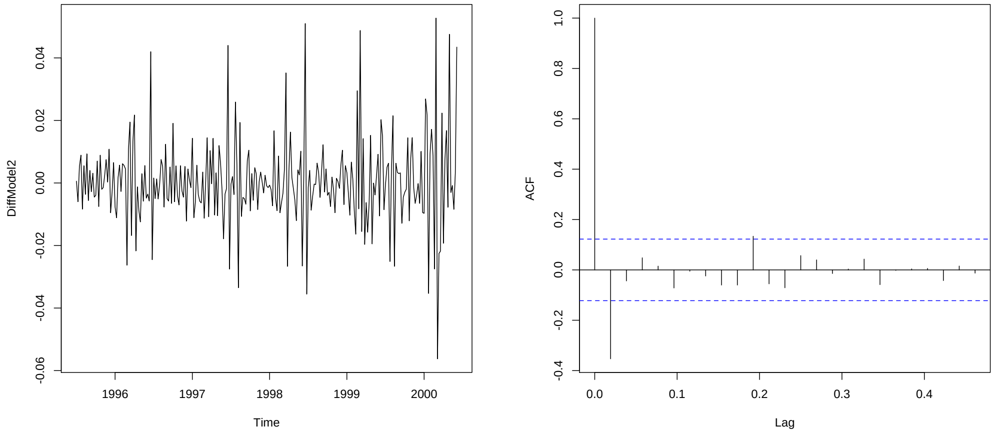
The first and second order differences are defined as follows:

$$\begin{aligned} \bigtriangledown y_t =& (1-B)y_t = y_t - y_{t-1} \\ \bigtriangledown^2 y_t =& (1-B)^2 y_t \end{aligned}$$

Differencing the seasonally adjusted gasoline prices twice (2nd order differencing), we obtain the time series and acf shown in Figure 5 . The autocorrelation function is well behaved and indicates a stationary process. The $acf(1)$ appears to be significant. If the $acf(1)$ is significant then the second order differences have a moving average structure remaining and a stochastic trend may be preferable. If $acf(1)$ is not significant then the the second order differences are best described as white noise, indicative of a quadratic trend. It is clear from Figure (3) that the trend is not quadratic. Taking higher order differences does not indicate a higher order polynomial is a preferred starting point for a model. This first step does not provide a model, but it gives us a starting point for specifying a model. We can begin by exploring AR(p) models and analyzing the residuals. If the trend had been deterministic, then we would begin by choosing the best curve for the trend. In our case, an AR(2) provides good place to begin analysis.

We begin our model construction by letting $x(t)$ be the seasonally adjusted data for gasoline prices from 1995 through 2000. Hence, $x(t)$ will the data minus the seasonal adjustment. Since we are specifying an AR(2) stochastic trend we have,

$$x(t) = \beta_1 x(t-1) + \beta_2 x(t-2) + C + \epsilon_t \tag{15}$$



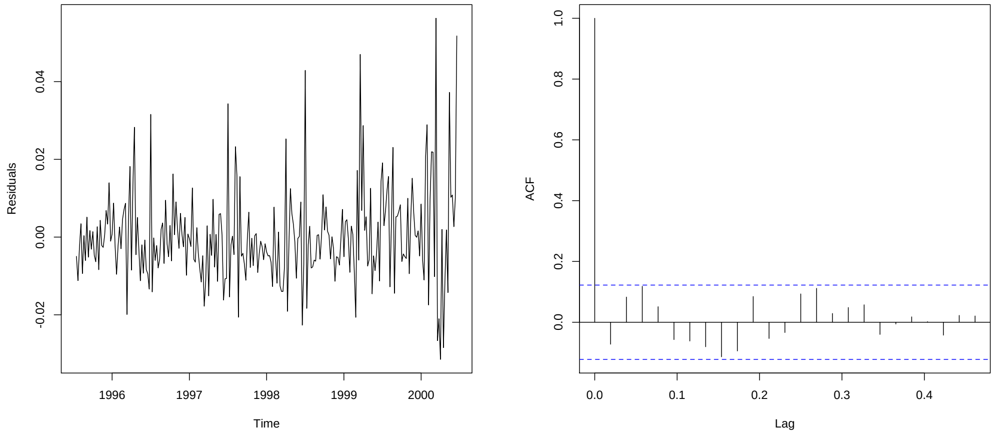
,

Figure 5: 2nd order difference plot and its acf

Using least squares regression we obtain the parameter estimates,

$$\hat{x}(t) = 1.50548x(t-1) - 0.50113x(t-2) - 0.00405 + \epsilon_t \tag{16}$$

Where the estimate for the constant is not significantly different from zero. The fit is quite good. The adjusted R^2 is 0.9907. The residual plot and correlogram are shown in Figure 6. It looks as though the trend has clearly been captured by the AR(2) process. The correlogram shows the residuals could have come from white noise. The plot of the residuals does show the possibility of increasing variance. This could be corrected by either using a multiplicative smoothing model or tranforming the data using the natural logarithm or we could seperately model the residuals. We will continue in our analysis using the AR(2) model in (16).



,

Figure 6: 2nd order difference plot and its acf

Let $y(t)$ be the model that includes the seasonal adjustments,

$$y(t) = x(t) + s(t) \tag{17}$$

where $\hat{s}(t)$ is the estimate of $s(t)$ and is the time series of seasonal adjustments produced by our smoothing process.

For forecasting and demonstration purposes, we examine weekly gasoline prices over a shorter time period. Consider weekly gasoline prices from January through April 2005.

Week	Gasoline(t)	Week	Gasoline(t)
1	1.745	9	1.904
2	1.771	10	1.979
3	1.802	11	2.039
4	1.839	12	2.095
5	1.896	13	2.137
6	1.89	14	2.196
7	1.873	15	2.251
8	1.878	16	2.198
		17	2.197

Table 13: Average Weekly Gasoline Prices Jan-Apr 2005

As seen in Figure (7), this data visually exhibits a strong linear trend with respect to time. We've previously seen a stochastic trend provides a model for the data over a different and longer time period. For comparison purposes, in this exposition we will build a predictive model using the linear trend, and let the reader build two different models in the exercises. For the linear trend, we use linear regression to determine estimates for parameters β_1 and β_2 the model,

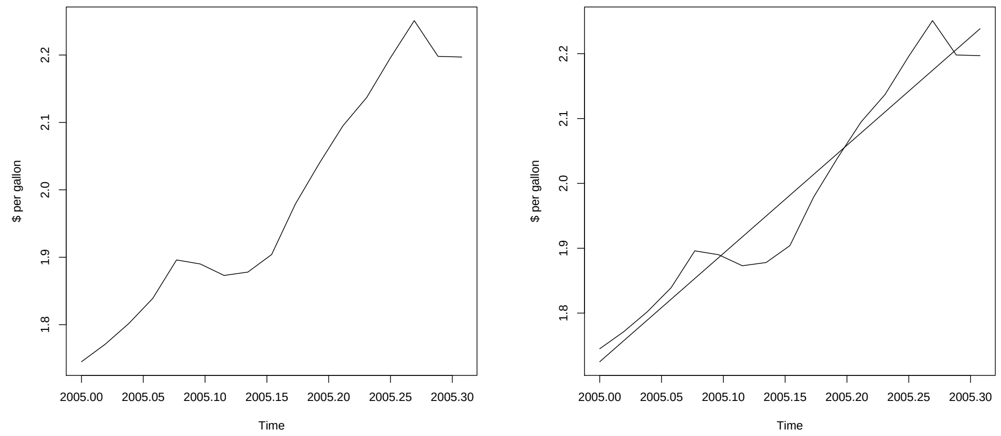
$$y(t) = \beta_1 t + \beta_2 + \epsilon_t \tag{18}$$

Regression yields the fitted equation,

$$\hat{y}(t) = 0.0321 \cdot t + 1.6929 + \epsilon_t \tag{19}$$

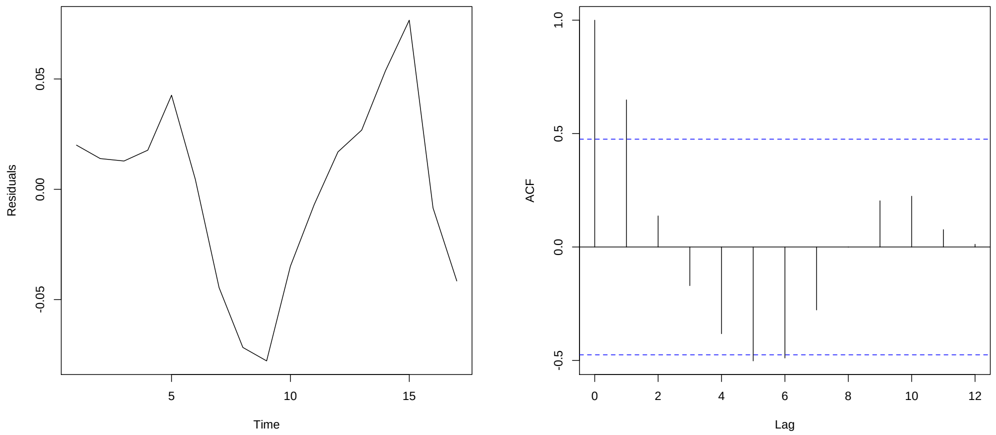
The statistics for this fit include the residual standard error: 0.0439 on 15 degrees of freedom and an adjusted R-squared of 0.9314. The plots of the model and data are shown in Figure (7).

Although a we have a strong linear fit, for predictive purposes, the model is still inadequate. We can see this in two ways. First, the residual standard error under certain conditions is an approximation of the one-period ahead forecast error. In this case, 0.0439 likely would not be an acceptable forecast error for next week's gasoline prices. Second, consider the residual plot shown in Figure(8). There is a clear autocorrelation structure in the residuals. With considerable structure in the residuals, the residual standard error is likely not a good estimate of the forecast error. A model appropriate for forecasting should have residuals that come from white noise. However, it seems reasonable to assume the mean and variance in the residual plot are stationary. Hence the sample acf and sample



,

Figure 7: Average gasoline prices first quarter 2005 and trend



,

Figure 8: Residuals to linear model and sample acf

pacf estimate the population ACF and PACF, respectively, and their significance levels can be used to gain insight into the structure of the residuals.

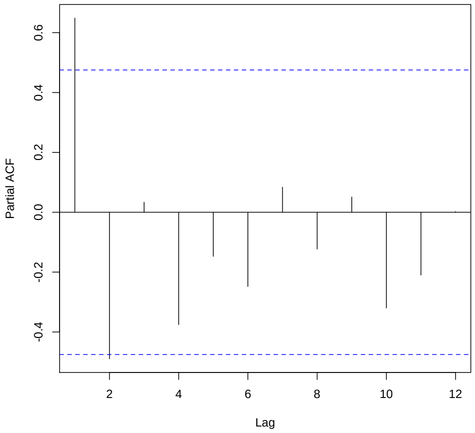
In Figure (8), the sample *acf* shows a periodic dampening sequence of autocorrelations in the lags. Some of these lags exceed the level of significance. This pattern is common for data from an AR(2) process, but the *pacf* can be used to provide more information on the order of the autoregression model to be used. Indeed, the sample *pacf*, Figure(9) indicates correlation coefficients at lags 1 and 2 are likely significant, although the coefficient at lag 2 is just touching the level of significance line.

Based on this information we choose an AR(2) model for the residuals,

$$\epsilon_t = \alpha_1 \epsilon_{t-1} + \alpha_2 \epsilon_{t-2} + \nu(t)$$

Using ordinary least squares we obtain the fit,

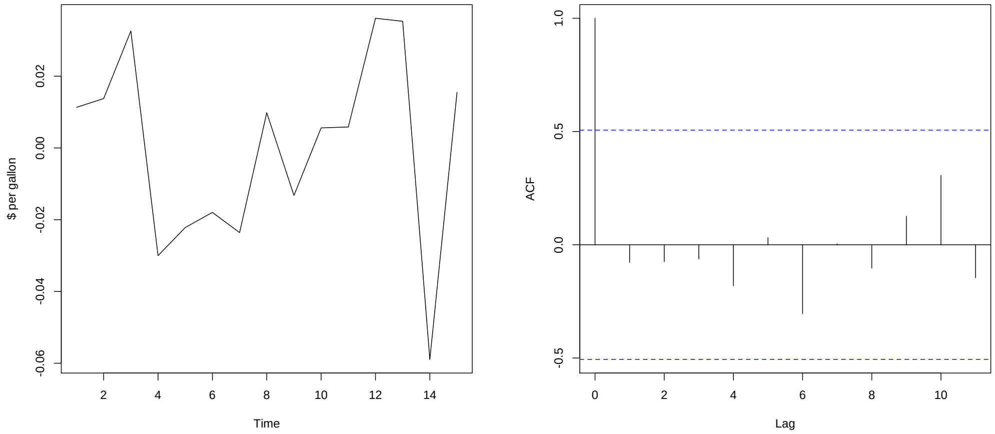
$$\hat{\epsilon}_t = 1.1043\epsilon_{t-1} - 0.6033\epsilon_{t-2} + \nu(t) \tag{20}$$



,

Figure 9: The sample $pacf$ to the residuals from linear model

Finally we examine the residuals, disturbances, for this AR(2) model. From the information



,

Figure 10: Disturbances to Residual model and sample acf

contained in the graphs in Figure(10) it is likely disturbances for the AR(2) model for the residuals come from a white noise process. Hence, the residual standard error of about 0.0306 approximates the one period ahead forecast error. The final model has the form,

$$
\begin{aligned}
&y(t) = \beta_1 t + \beta_2 + \epsilon_t \\
&\epsilon_t = \alpha_1 \epsilon_{t-1} + \alpha_2 \epsilon_{t-2} + \nu(t) \\
&\nu(t) \sim WN
\end{aligned}
$$

The estimates for the parameters now have been determined. We are now ready to forecast using the fit to our model. We will use the average weekly price of gasoline for weeks 16 through 22 is shown in Table (16). For notation, if $y(t)$ denotes the process from which our data has been generated and $\hat{y}(t)$ is the fitted model, then the one period ahead forecast at time T is denoted ,$\hat{y}_{T,1}$. As an example, if we wish to use the fitted model developed above to forecast the average price of gasoline for week T in 2005, we need to calculate

$$\hat{y}_{T,1} = \hat{\beta}_1(T+1) + \hat{\beta}_2 + \hat{\epsilon}_{T,1}$$

where the one period ahead forecast in the residual is

$$\hat{\epsilon}_{T,1} = \hat{\alpha}_1 \epsilon_T + \hat{\alpha}_2 \epsilon_{T-1}$$

and ϵ_T and ϵ_{T-1} are the observed errors from the linear model at times T and $T-1$. Therefore,

$$\begin{aligned} \hat{\epsilon}_{17,1} &= 1.1043 \cdot -0.041549 + -0.6032 \cdot -0.008451 = -0.040783 \\ \hat{y}_{17,1} &= 0.032098 \cdot 18 + 1.69288 - 0.40783 = 2.22986 \end{aligned}$$

We summarize the calculations of the one week ahead forecasts in Table (14).

Week	Observed Price	Linear Fit	Residual	Residual Fit	Forecast	Forecast Error
16	2.198	2.2065	-0.0085			
17	2.197	2.2386	.0.0415			
18	2.191	2.2706	-0.0796	-0.0408	2.2299	-0.0389
19	2.137	2.3027	-0.1657	-0.0629	2.2399	-0.1029
20	2.166	2.3348	-0.2188	-0.1350	2.1999	-0.0839
21	2.077	2.3669	-0.2899	-0.1417	2.2253	-01483
22	2.051	2.3990	-0.3480	-0.1881	2.2109	-0.1599

Table 14: One week ahead gasoline price forecasts May 2005

Student Activity: Models for forecasting gasoline prices

The previous discussion demonstrates model construction for a model used to predict one week ahead gasoline prices. In this activity, the reader will inlcude an explanatory variable in the model. Consider the US weekly average crude oil and and gasoline prices in Table (15)

Week	$Oil(t)$	$Gasoline(t)$	Week	$Oil(t)$	$Gasoline(t)$
1	42.52	1.745	9	51.75	1.904
2	44.07	1.771	10	52.74	1.979
3	46.79	1.802	11	54.22	2.039
4	47.85	1.839	12	55.93	2.095
5	48.56	1.896	13	52.95	2.137
6	46.97	1.89	14	54.97	2.196
7	46.08	1.873	15	55.24	2.251
8	47.82	1.878	16	51.44	2.198
			17	52.39	2.197

Table 15: Sample Weekly Oil and Gasoline Prices Jan-Apr 2005

1. Construct a model to predict 1 - week ahead average gas prices in the US using the previous week's price of crude oil as an explanatory variable.

   (a) Plot on the same axis, $Oil(t)$ and $25 \times Gasoline(t)$ from the Table (15). Can you identify places where the movement in the price of gasoline seems to lag the price of oil by a week?

   (b) Use linear regression to estimate the model parameters for

   $$Gasoline(t) = \beta_1 Oil(t-1) + \epsilon_t \tag{21}$$

Week	$Oil(t)$	$Gasoline(t)$	Week	$Oil(t)$	$Gasoline(t)$
17	52.39	2.197	22	50.15	2.051
18	52.00	2.191	23	53.76	2.078
19	50.64	2.137	24	53.74	2.099
20	50.33	2.116	25	56.18	2.128
21	47.77	2.077	26	59.04	2.186

Table 16: Average Weekly Oil and Gasoline Prices May-June 2005

   (c) Assess the model fit. Identify Adjusted R^2, residual standard error, and plot the residuals.

   (d) Examine the residuals. Construct the sample acf and $pacf$. Do the residuals appear to have come from white noise?

   (e) Use the estimated model to provide the one week ahead forecast for gasoline prices in May 2005 and June 2005 given the data in Table(16).Calculate the one week ahead forecast errors.

2. Construct a model to predict 1 - week ahead average gas prices in the US using an explanatory variable and a stochastic trend model.

   (a) Add the previous week's gasoline price to the model,

   $$Gasoline(t) = \beta_1 Oil(t-1) + \beta_2 Gasoline(t-1) + \epsilon_t \tag{22}$$

   (b) Assess the model fit. Identify Adjusted R^2, residual standard error, and plot the residuals.

   (c) Examine the residuals. Construct the sample acf and $pacf$. Do the residuals appear to have come from white noise?

   (d) Use the estimated model to provide the one week ahead forecast for gasoline prices in May 2005 and June 2005 given the data in Table(16). Calculate the one week ahead forecast errors.

3. Compare the three models for forecasting the one week ahead average US gasoline prices. Which one of the three ended up providing the best forecasts? Why do you think this model was superior to the others? Which model did the Adjusted R^2, residual standard error, and other statistics imply would be the best to use for forecasting?

Notes to Instructor. *The discussion to the exercises follows.*

1. *Solutions to using the previous week's oil prices as an explanatory variable in the modeling the average weekly gasoline prices.*

   (a) *The oil vs scaled gasoline is shown in Figure (11) with the scaled gasoline time series formatted with dashed lines. Although it is not always clear that the change in gasoline prices is preceded by the change in oil prices, near the end of week 5, 2005 (2005.10) the upturn in oil prices clearly precedes the upturn in gasoline prices.*

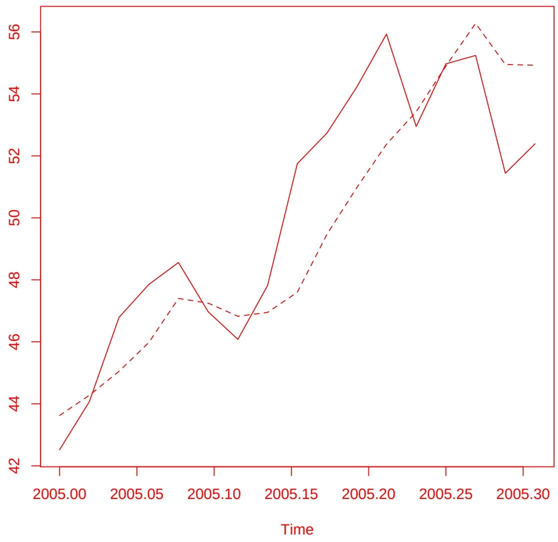


Figure 11: Price of oil vs 25x price of gasoline

   (b) *The estimate for* β_1 *is* $\hat{\beta}_1 = .0399052$.

   (c) *In assessing the model fit, we identify Adjusted* $R^2 = 0.9989$, *residual standard error of* 0.06767, *and plot the residuals in Figure(12).*

   (d) *The residual plot appears random or at least there is no reason to believe that it is not second order stationary. The sample* acf *and* $pacf$ *are shown in Figure (13). Both graphs indicate that the residuals could have come from white noise.*

   (e) *The one week ahead forecasts are shown in Table(17).*

Week	*forecasts*	*error*	*Week*	*forecasts*	*error*
18	*2.09*	*0.10*	*23*	*2.00*	*0.08*
19	*2.08*	*0.06*	*24*	*2.15*	*-0.05*
20	*2.02*	*0.10*	*25*	*2.14*	*-0.02*
21	*2.01*	*0.07*	*26*	*2.24*	*-0.06*
22	*1.91*	*.14*			

Table 17: 1-week ahead forecasts and forecast errors May-June 2005

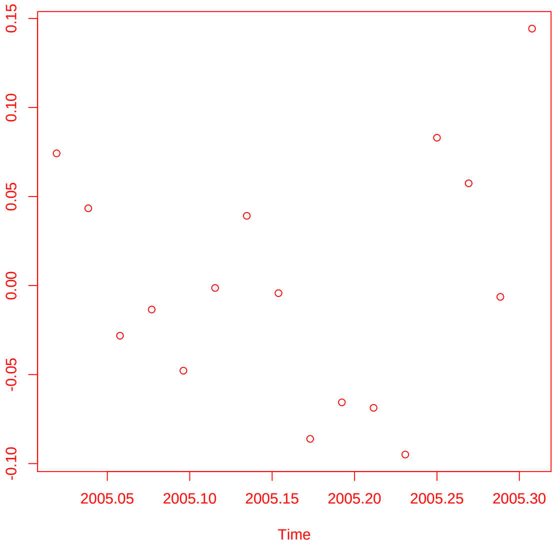


Figure 12: Residual plot

2. *Solutions to using the previous week's oil prices as an explanatory variable and modeling the stochastic trend in the model of the average weekly gasoline prices.*

   (a) *The estimate for β_1 and β_2, respectively, is $\hat{\beta}_1 = 0.010459$ and $\hat{\beta}_2 = 0.748583$.*

   (b) *In assessing the model fit, we identify Adjusted $R^2 = 0.9998$, residual standard error of 0.0283, and plot the residuals in Figure(14).*

   (c) *The residual plot appears random or at least there is no reason to believe that it is not second order stationary. The sample acf and $pacf$ are shown in Figure (15). Both graphs indicate that the residuals could have come from white noise.*

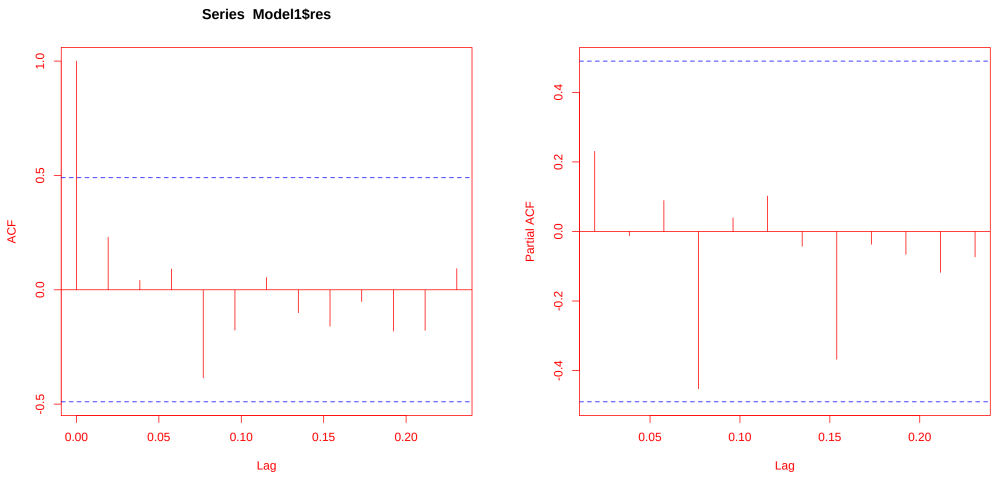
   (d) *The one week ahead forecasts are shown in Table(18).*

Week	forecasts	error	Week	forecasts	error
18	2.19	0.00	23	2.06	0.02
19	2.18	-0.05	24	2.12	-0.02
20	2.13	-0.01	25	2.13	-0.01
21	2.11	-0.03	26	2.18	0.01
22	2.05	0.00			

Table 18: 1-week ahead forecasts and forecast errors May-June 2005

3. *Discussion on comparing the three models for forecasting the one week ahead average US gasoline prices.*

   The model using the previous week's oil price and the stochastic trend is clearly the best model for forecasting the one week ahead prices. The model with the deterministic trend and AR(2) residuals has a very good fit with a small residual standard error but performs poorly as a forecasting model. This suggests that the actual trend is stochastic.



,

Figure 13: The sample ACF and PACF for the residuals

This is an example of a model that has a very good in-sample performance but poor out of sample performance. The model using only the previous week's oil price as an explanatory variable clearly needed to capture more of the gasoline price dynamics. The adjusted R^2 was very good, but the residual standard error indicated a poor out of sample performance was likely. Finally, by including the stochastic trend term, we arrived a very good forecasting model. The residual standard error seems to approximate the forecasting errors well.

We should note that the sample acf and pacf are not quite as useful with our small sample. The small sample of 17 weeks causes a wide 95% confidence band and makes it difficult to identify nonzero autocorrelation coefficients and partial autocorrelation coefficients.

Topic 4: Forecasting Peak Production - Differential Equation Models

In 1956, a geoscientist named M. King Hubbert, predicted that peak oil production in the continental US would peak between 1965 and 1970. To make this prediction, Hubbert used the logistic model and historical US oil production data. To see how Hubbert made this prediction consider the logistic differential equation,

$$\frac{dy}{dt} = ky(1 - \frac{y}{Q}) \tag{23}$$

The solution to this differential equation is the well known logistic function and can be obtained by separation of variables

$$y(t) = \frac{Q}{1 + Ae^{-kt}}, \qquad A = \frac{Q - y(0)}{y(0)} \tag{24}$$

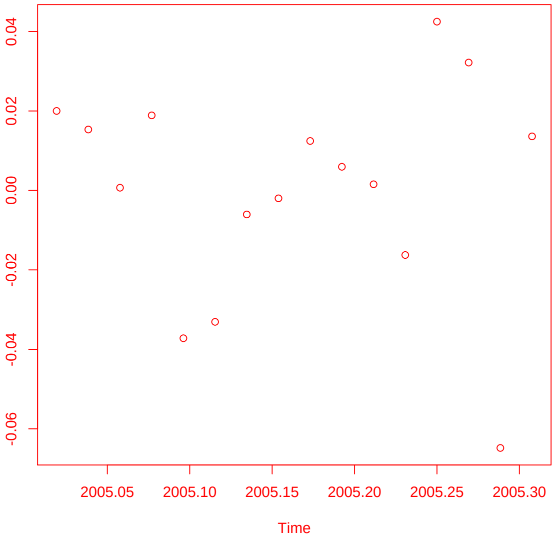


Figure 14: Residual plot

In the context of modeling a population, the parameter Q is called the carrying capacity. To understand the assumptions to which this model would apply, one might consider the rabbit population in a ten acre field. Being a specific size the field has finite resources for food, water, and shelter. Initially, the rabbits have plenty of each and reproduce at a rate nearly proportional to their population, the ratio $\frac{y}{Q}$ in equation (23) is close to zero for small values of y. As the population size increases, competition for each of the available resources drives the growth rate down. This effect can be seen in the model. As the population size, y, approaches the carrying capacity Q, the ratio $\frac{y}{Q}$ approaches 1, and the growth rate $\frac{dy}{dt}$ approaches zero.

For modeling the rate of extraction of a scarce resource, the parameter Q represents the total amount that can be extracted or recovered with current field knowledge and technology. Initially, extraction of the resource occurs at nearly an exponential rate. Eventually, new discoveries of resource deposits become more rare and the older deposits yield less in the extraction process. Hence, the rate of extraction declines sharply.

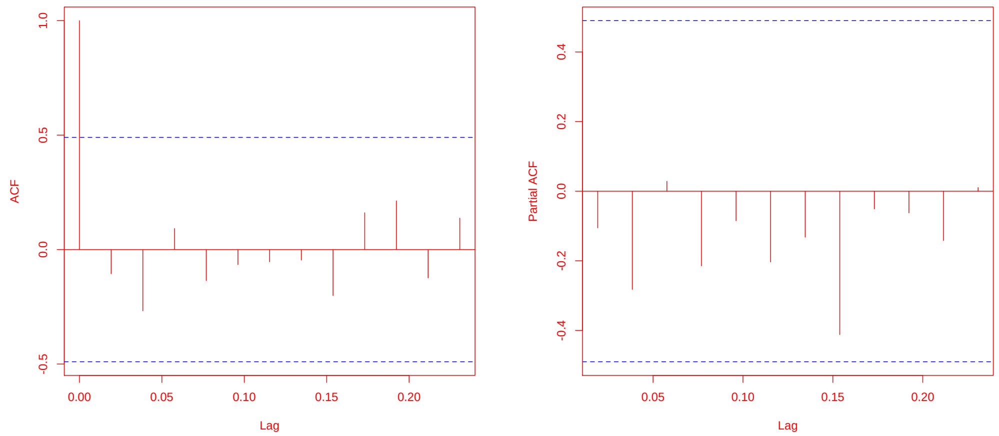
To fit the model to oil production data, we rearrange (23)

$$\frac{1}{y}\frac{dy}{dt} = k - \frac{k}{Q}y \tag{25}$$

to show the equation for the relative rate of change. We let the average annual production for a year, Δy, approximate $\frac{dy}{dt}$ and apply linear regression to fit the model parameters. The regression model for the relative rate of change that we use has the form

$$\frac{\Delta y}{y} = \beta_1 + \beta_2 y + \epsilon \tag{26}$$

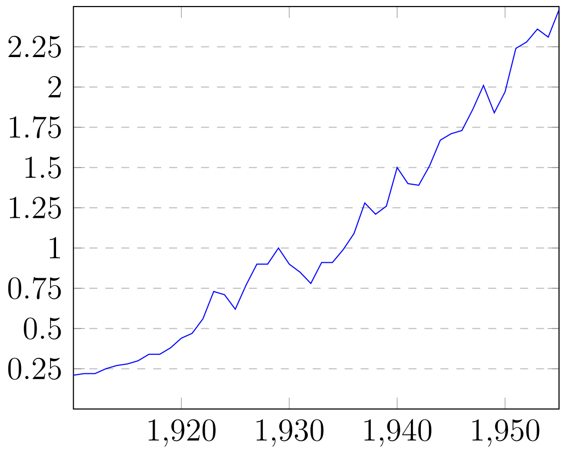
In 1956, the data available is shown in the plot. The numerical values can be seen in column Δy in Table (19).



,

Figure 15: The sample ACF and PACF for the residuals

US Oil Annual Production 1910-1955



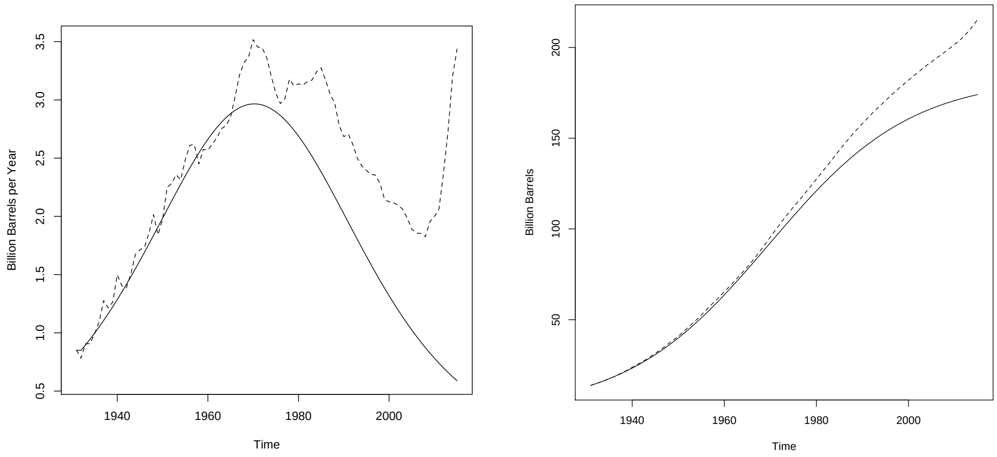
Using the data from Table 19 from the years 1931 - 1950, form the column $\frac{\Delta y}{y}$. Least square regression yields,

$$\frac{\Delta y}{y} = 0.06472 - 0.000353y \tag{27}$$

Assessing the fit, we obtained an $r^2 = 0.47$. Though not a strong fit, about 50% of the variation in the data can be explained by the linear relationship in the data. Comparing (27) to (25) we obtain the estimates $k = 0.06472$ and $Q = 183.34$. With these estimates the model solution becomes,

$$y(t) = \frac{183.34}{1 + 12.26e^{-0.06472t}}, \qquad Year(t) = 1931 + t \tag{28}$$

Now that we have fitted the model, we can interpret the results. First, (23) is quadratic in y. Using the equation for the vertex of a parabola, we find that $\frac{dy}{dt}$ has a maximum at $y = \frac{Q}{2}$. Therefore, our regression model, $\frac{dy}{dt}$ has a maximum at $y = 91.67$ billion barrels of total production. Substituting this result into equation (28) yields $t = 38.72$. Hence, maximum annual production occurs between 1969 and 1970. Finally, $Q = 183.34$ billion barrels that can be extracted with current (1950's) field knowledge and technology. Hubbert provided a range of total recoverable oil resources from the lower 48 states between 150 and 200 billion barrels. This information came for the geological surveys of the time. The model's estimate of peak production $\frac{dy}{dt} = 2.97$. The data show that actual peak production in 1970 was about 3.5175, or about 18% more than the value predicted by the model.



,

Figure 16: The Annual and Cummulative Oil Production Model Solutions

Notice that we obtain different estimates and different predictions for peak production if we change the time interval over which we perform the regression. The interval of estimates obtained in this way may also be useful.

Year	Δy	y	Year	Δy	y
1930	0.8979	12.9794	1951	2.2477	43.1211
1931	0.8512	13.8306	1952	2.2834	45.4045
1932	0.7829	14.6135	1953	2.3572	47.7617
1933	0.9056	15.5191	1954	2.3148	50.0765
1934	0.9081	16.4272	1955	2.4846	52.5611
1935	0.9939	17.4211	1956	2.6101	55.1712
1936	1.0954	18.5165	1957	2.6171	57.7883
1937	1.2775	19.7940	1958	2.4492	60.2374
1938	1.2133	21.0072	1959	2.5747	62.8121
1939	1.2644	22.2716	1960	2.5678	65.3799
1940	1.4991	23.7706	1961	2.6218	68.0017
1941	1.4042	25.1748	1962	2.6762	70.6779
1942	1.3855	26.5603	1963	2.7528	73.4307
1943	1.5056	28.0659	1964	2.7791	76.2098
1944	1.6732	29.7391	1965	2.8485	79.0583
1945	1.7137	31.4528	1966	3.0277	82.0859
1946	1.7334	33.1826	1967	3.2157	85.3016
1947	1.8571	35.0433	1968	3.3200	88.6216
1948	2.0148	37.0581	1969	3.3719	91.9935
1949	1.8418	38.8999	1970	3.5175	95.5110
1950	1.9736	40.8734	1971	3.4540	98.9650

Table 19: US Oil Production 1930-1971

Class Activity: Applying the Logistic Model

1. The values of the model parameters determined by regression, k and Q, are sensitive to the choice of the time interval. In the discussion above we chose $\frac{\Delta y}{y}$ and y from the years 1931 to 1950.

   (a) Perform linear regression using the model in (26) and the data from years 1930 to 1950. What is r^2? What are the estimates for k and Q? In what year does the model predict peak production? What is the predicted peak production and how does it compare with the data point for that year?

   (b) Repeat the preceding regression problem and answer the same questions, but use data from years 1932 to 1950.

Notes to Instructor. *The discussion of the class activity.*

1. Solutions

   (a) $r^2 = 0.5124$, $k = 0.0669$ *and* $Q = 158.004$. *The model predicts peak production in the year 1966. The model predicts peak production to occur when 79 billion barrels have been extracted and it predicts a value of* $\frac{dy}{dt} = 2.64$.

(b) *$r^2 = 0.4277$, $k = 0.06419$ and $Q = 190.99$. The model predicts peak production in the year 1971. The model predicts peak production to occur when 95.49 billion barrels have been extracted and it predicts a value of $\frac{dy}{dt} = 3.06$.*

Student Activity: Applying the Logistic Model

1. In 2016, coal accounted for 30% of the electricity generation in the US. However, from 1980 to 2008 it consistently accounted for 50% or more of electricity generation in the US. The use of coal has dropped off precipitously in just the past two years, thus affecting any models that may have been used to predict coal production and demand produced from historical data. In Table (20), Δy is the US annual coal production in units of million short tons from 1971 to 1990[12]. y represents the cumulative coal production through the year indicated. We will assume that most current mines began after 1948, and take the total coal production from current mines to be 0 at the beginning of 1949.

Year	Δy	y	Year	Δy	y
1971	560.9	11,571.4	1981	823.8	18,261.1
1972	602.5	12,132.3	1982	838.1	19,084.9
1973	598.6	12734.8	1983	782.1	19,923.0
1974	610.0	13,333.4	1984	895.9	20,705.1
1975	654.6	13,943.4	1985	883.6	21,601.0
1976	684.9	14,598.0	1986	890.3	22,484.6
1977	697.2	15,282.9	1987	918.8	23,374.9
1978	670.2	15,980.1	1988	950.3	24,293.7
1979	781.1	16,650.3	1989	980.7	25,244.0
1980	829.7	17,431.4	1990	1,029.1	26,224.7

Table 20: US Coal Production 1971-1990

   (a) Perform linear regression using the model in (26) and the data from coal mining production in Table (20) years 1971 to 1990. What is r^2? What are the estimates for k and Q? Taking $year = 1971 + t$, in what year does the model predict peak production? What is the predicted peak production? Using on line data, find the year of peak production and the amount of peak production and compare these to your model result.

   (b) Our model is for production from current mines (production after 1948). In the online discussion[13], geologists estimate that there is 18.3 billion short tons of coal recoverable from current mines. From 1949 through 2016, $52,996$ million short tons had been mined from coal mines in the US. How much recoverable coal from current mines is predicted by our model?

[12] *Independent Statistics & Analysis*, U.S. Energy Information Administration, https://www.eia.gov/totalenergy/

[13] *Independent Statistics & Analysis*, U.S. Energy Information Administration, https://www.eia.gov/energyexplained/

Notes to Instructor. *The discussion of the student activity: Applying the Logistic Model.*

1. *Solutions*

   (a) $r^2 = 0.8258$, $k = 0.0564$ *and* $Q = 78,529.3$. *The model predicts peak production in the year 2002. The model predicts peak production to occur when 39,264 million short tons have been extracted. The model predicts peak production will take place in approximately 2002 and have a value of* $\frac{dy}{dt} = 1,107.8$. *The data shows a bimodal distribution for annual production. The first local maximum in production, 1127.7 million short tons, occurs in year 2001. The second local maximum in production, 1,162.7 million short tons, occurs in 2006.*

   (b) *Since* $52,996$ *million short tons has been mined through 2016 and our model predicts that* $78,529.3$ *will be the total amount recovered, then* $78,529.3 - 52,996 = 25,533.3$ *remains to be recoverd.*

List of Tables

A R Instructions

To assist in using R for this module, we include R instruction used to produce selected parts of the discussion. These instructions can then be applied to work the class and students activities.

R in Topic 3. For the Class Activity: Using Exponential Trends, we assume the *gasoil48_13.csv* file containing the sample average annual oil and gasoline prices from 1945-2015 as given in Table(12) is stored in the folder *Reconnect 2017* where row 1 contains column labels, column 1 lists the dates 1948 through 2013 in increments of five years, column 2 has the annual oil prices, column 3 has the annual gasoline prices for the years listed, we read the data and form the time series for gasoline prices using R:

```
> GasOil4813<-read.table("/Reconnect 2017/gasoil48_13.csv",header = TRUE,sep = ",")
> Gas4813.ts<-ts(GasOil4813[,3], st=c(1948,1), end=c(2013,1), fr=.2)
```

We take time $T = 0$ in the year 1945. Since the first value in the table is from 1948, we form the time sequence, $T = 3, 8, ...68$ corresponding to the five year time increments in the data and transform the data using the $log()$ function. Finally, the linear regression code used to produce and store the result in Model2 uses the R function lm().

```
> T <- seq(3,68,5) ##we assume T=0 in 1945.
> LogGas4813 <- log(Gas4813.ts)
> Model2 <- lm(LogGas4813~T)
> summary(Model2)

Call:
lm(formula = LogGas4813 ~ T)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37732 -0.18631 -0.02477  0.22001  0.41805

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.751687   0.146587  -11.95 5.07e-08 ***
T            0.040544   0.003591   11.29 9.50e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2708 on 12 degrees of freedom
Multiple R-squared:  0.914,Adjusted R-squared:  0.9068
F-statistic: 127.5 on 1 and 12 DF,  p-value: 9.504e-08
```

To produce Figure(3), we first assume the weely gasoline and oil prices file, Gaswkly1995.csv, is stored in the folder *Reconnect 2017* where row 1 contains column labels, column 1 has the average weekly gasoline prices, and column 2 has the weekly oil prices from the first week in 1995. After forming the time series we use the plot() and decompose() functions in R to produce the figure showing the time series, trend, weekly adjustments and random component from 1995 through 2000. The code is

```
> GASOILWK1995 <- read.table("/Reconnect 2017/Gaswkly1995.csv",header = TRUE,sep = ",")
> GASWK1995.ts <- ts(GASOILWK1995[,1], st=c(1995,1), end=c(2000,52), fr=52)
> plot(decompose(GASWK1995.ts))
```

To produce the model estimate in (16), we first produce the seasonally adjusted data and then use regress the model variable against the first and second time lag of the variable. These operations are completed with the following code.

```
> GASWK1995.decom <- decompose(GASWK1995.ts)
> GASWK1995.seas <- window(GASWK1995.decom$seasonal, st=c(1995,27), end=c(2000,25))
> GASWK1995.mod <- window(GASWK1995.ts, st=c(1995,27), end=c(2000,25))
> GASWK1995.seasadj <- GASWK1995.mod - GASWK1995.seas
> library(dynlm)
> Model1 <- dynlm(GASWK1995.seasadj ~ L(GASWK1995.seasadj,1) + L(GASWK1995.seasadj,2))
> summary(Model1)

Time series regression with "ts" data:
Start = 1995(29), End = 2000(25)

Call:
dynlm(formula = GASWK1995.seasadj ~ L(GASWK1995.seasadj, 1) +
    L(GASWK1995.seasadj, 2))

Residuals:
      Min        1Q    Median        3Q       Max
-0.031476 -0.006856 -0.001277  0.005045  0.056301

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.004049   0.007258  -0.558    0.577
L(GASWK1995.seasadj, 1)  1.505480   0.057165  26.336  < 2e-16 ***
L(GASWK1995.seasadj, 2) -0.501126   0.058054  -8.632 6.72e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.01229 on 254 degrees of freedom
Multiple R-squared:  0.9907,Adjusted R-squared:  0.9907
F-statistic: 1.36e+04 on 2 and 254 DF,  p-value: < 2.2e-16
```

Notice how we used the window() function in R to adjust the original time series data to have the same dimension as the trend and random components produced by decompose(). The trend and random components are missing 26 weeks of data at the beginning and end since decompose() uses the centered moving average to produce the trend. However, we could have gotten away with not redimensioning if all we were interested in was producing the seasonally adjusted data.

For the shorter 2005 seventeen week time series discussion used for forecasting, we assume the weely gasoline and oil prices file, Gasoilwkly2005.csv, is stored in the folder *Reconnect*

2017 where row 1 contains column labels, column 1 has dates, column 2 has the average weekly gasoline prices, and column 3 has the weekly oil prices from the first week in 2005. The following code produces the parameter estimates in equation (19) as well as the Figure(7),Figure(8),and Figure(9).

```
> Gasoilwkly2005 <- read.table("/Reconnect 2017/Gasoilwkly2005.csv",header = TRUE,sep =
> Gaswkly2005.ts<-ts(Gasoilwkly2005[,2], st=c(2005,1), end=c(2005,52), fr=52)
> Gaswkly05Q1<-window(Gaswkly2005.ts, st=c(2005,1), end=c(2005,17))
> T <- seq(1,17,1) #T is the week number
> Model <- lm(Gaswkly05Q1 ~ T)
> summary(Model)

Call:
lm(formula = Gaswkly05Q1 ~ T)

Residuals:
     Min       1Q   Median       3Q      Max
-0.07777 -0.03486  0.01282  0.02002  0.07665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.692882   0.022272   76.01  < 2e-16 ***
T           0.032098   0.002174   14.77 2.42e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.0439 on 15 degrees of freedom
Multiple R-squared:  0.9356,Adjusted R-squared:  0.9314
F-statistic: 218.1 on 1 and 15 DF,  p-value: 2.42e-10

> Trend <- predict(Model)
> ts.plot(cbind(Trend, Gaswkly05Q1), lty = 1:2)
> Model.res <- Model$res
> ts.plot(Model.res) #Left part of Figure
> acf(Model.res) #Right part of Figure
> pacf(Model.res)
```

The parameters estimates obtained in equation (20) were produced by least squares linear regression in EXCEL. In R, one could also use the ar() function.

```
> ar(Model.res,2)
Call:
ar(x = Model.res, aic = 2)
Coefficients:
      1        2
 0.9663  -0.4897
```

The result gives slightly different parameter estimates. Since the lag 1 parameter is so close to unity, the statistics showing the accuracy of the estimate may not themselves be accurate. In many texts, it is suggested that the first difference be taken $\bigtriangledown y_t = (1 - B)y_t = y_t - y_{t-1}$ and then modeled. In our case we would model $\epsilon_t - \epsilon_{t-1}$. The R-code for taking the first difference and naming the resulting time series FirstDiff is,

```
> FirstDiff <- diff(Model.res, 1).
```

In any case, the exercises show a different and more accurate forecasting model.

For Topic Four Assuming the .csv file containing US oil production data from 1859 to 2016 is stored in the folder *Reconnect 2017* where row 1 contains column labels, column 3 has annual production data, column 4 has cumulative production data, and column 5 has the relative rate, $\frac{\Delta y}{y}$, data, we read the data and form the three time series using R:

```
> OilProdAnn<-read.table("/Reconnect 2017/US_Oil_Prod.csv",header = TRUE,sep = ",")

> OilProdAnn1859.ts<-ts(OilProdAnn[,3], st=c(1859,1), end=c(2016,1), fr=1)
> OilProdCum1859.ts<-ts(OilProdAnn[,4], st=c(1859,1), end=c(2016,1), fr=1)
> OilProdCumRelRate1859.ts<-ts(OilProdAnn[,5], st=c(1859,1), end=c(2016,1), fr=1)
```

We use the *window* function to produce subsets of the time series from 1931 to 1950.

```
> OilProdAnn1931<-window(OilProdAnn1859.ts, st=c(1931,1), end=c(1950,1))
> OilProdCum1931<-window(OilProdCum1859.ts, st=c(1931,1), end=c(1950,1))
> OilProdRelRate1931<-window(OilProdCumRelRate1859.ts, st=c(1931,1), end=c(1950,1))
```

The linear regression code that produces the result in 27 is given by:

```
> Model<-lm(OilProdRelRate1931~OilProdCum1931)
> summary(Model)

Call:
lm(formula = OilProdRelRate1931 ~ OilProdCum1931)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0059862 -0.0016413 -0.0000727  0.0011557  0.0068070

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.472e-02  2.345e-03  27.605 3.48e-16 ***
OilProdCum1931 -3.530e-04  8.755e-05  -4.032 0.000782 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 0.00325 on 18 degrees of freedom
Multiple R-squared:  0.4746,Adjusted R-squared:  0.4454
F-statistic: 16.26 on 1 and 18 DF,  p-value: 0.000782
```

One way to produce the plots shown in Figure (16) is to store the predicted model values in a .csv file. In our case, we store the predicted values and data values from 1931 to 2015 in the file where columns 2 and 3 contain the annual and cumulative production, respectively, and columns 4 and 5 contain the predicted cumulative and annual production, respectively.

```
US_Oil_Model.csv

>OilProdModel<-read.table("/Reconnect 2017/US_Oil_Model.csv",header = TRUE,
sep = ",")
> OilProdAnn31to15.ts<-ts(OilProdModel[,2], st=c(1931,1), end=c(2015,1), fr=1)
>OilProdCum31to15.ts<-ts(OilProdModel[,3], st=c(1931,1), end=c(2015,1), fr=1)
> OilProdModelCum.ts<-ts(OilProdModel[,4], st=c(1931,1), end=c(2015,1), fr=1)
> OilProdModelAnn.ts<-ts(OilProdModel[,5], st=c(1931,1), end=c(2015,1), fr=1)
> ts.plot(cbind(OilProdModelAnn.ts, OilProdAnn31to15.ts),lty=1:2,
ylab="Billion Barrels per Year")
>ts.plot(cbind(OilProdModelCum.ts, OilProdCum31to15.ts),lty=1:2,
ylab="Billion Barrels")
```

B Excel Instructions

It might be helpful to know how to use Excel to create some of the descriptive representations as well as the time series models. The following begins with a set of instructions for creating box-and-whisker plots, dot plots, bar graphs, scatter plots (regression), and time series plots. This is followed by instructions for constructing time series models with Excel.

Creating Box-and-Whisker Plots in Excel

(a) In a vertical column of data, scroll down to the end of the data set. Click on the cell two rows below and one row to the left of the data set. In that cell and the four cells below it, create a vertical column with the labels: **Minimum, 1st Quartile, Median, 3rd Quartile, and Maximum**.

(b) Now click on the cell two rows directly below the end of the data set (it will be one column to the right of the cell labeled Minimum. Type the Excel command

$$= min($$

leaving the left parenthesis open for now. In the cell, Excel will display the set of arguments that are to be placed within the parentheses; in this case it is waiting for one argument the data set over which the minimum is to be taken. With the cell dialogue open, click and drag from the first data entry to the last data entry, and then release the mouse (or you can type something like B2:B20, if the data set appears in the cells from B2 through B20). Press ENTER and the cell will produce the minimum value of the data set as its answer.

(c) The highlighted cell should now be the one directly under the minimum (to the right of the cell labeled **1st Quartile**). In this cell, type the command

$$= quartile(dataarray, quartile);$$

that is $= quartile(B2 : B20, 1)$ if the data appears in cells B2 through B20 (dragging the cursor through the data set will work as well). Press ENTER and the cell will produce the first quartile value of the data set as its answer.

(d) Repeat this process for each of the remaining values typing in the respective commands for the median, third quartile, and maximum:

$$= median(dataarray)$$

$$= quartile(dataarray, 3)$$

$$= max(dataarray)$$

(e) Repeat this process for each data set. If the data sets are vertically side-by-side, rather than typing each command over and over again, Excel has the ability to easily copy and paste commands and scripts with each copied command carrying out the same instructions on cells proportionally situated relative to the initial cells outlined in the original command. Therefore, highlight the five descriptive values under the first data set. In the lower right hand corner will appear a small, solid square. Move the cursor over this square until it changes from a wide cross to a stick cross. Click and drag the mouse to the right (or left) of the original column. Each of the original commands will be copied in each additional column highlighted, doing the same calculations on each respective data set relative to the first data sets position.

(f) Now that the five descriptive values are calculated, we will use the Excel graphing tools to create box-and-whisker plots. There is no direct Excel command that does this, so we will use our imaginations. Two rows below the five descriptive statistics, in the column that contains the first data set, type an equals sign (=); then click on the cell that contains the minimum value. Note that when you do this the column-row identification of that cell appears in the new cell. Press ENTER and the minimum value will again appear.

(g) The cell below it should now be highlighted. In that cell, type the command =B23-B22; that is, subtract the minimum value from the first quartile value, whatever the cell labels happen to be. Press ENTER.

(h) Repeat this process subsequently subtracting each of the following in each respective cell:

Median - First quartile value

Third quartile value Median

Maximum Third quartile value

(i) Copy and paste each of these values in each respective column which contains another data set.

(j) Highlight each of these newly calculated values over several (3 or 4) data sets.

(k) In the **Insert** menu, click the **Column** option in the **Charts** submenu. Then choose the **Stacked Column** option (not the **100% Stacked Column** option).

(l) Several stacked columns should appear in an Excel graphic. There should be the same number of columns as data sets. If not, the rows and columns have to be switched. This can be done easily by choosing the **Switch Row/Column** option in the **Design** submenu in the **Chart Tools** menu.

(m) Click on the top row of rectangles in the stacked columns. To make these invisible in the **Chart Tools** menu, select the **Format** submenu and choose the **No Fill** and **No Outline** options in the **Shape Fill** and **Shape Outline** menus.

(n) Repeat this for the bottom rectangles.

(o) Now click on the remaining top visible rectangles. In the **Chart Tools** menu, select the **Layout** submenu. In the **Layout** submenu. In the **Error Bars** menu, select the **More error bars** options.

(p) Choose the **Plus** and **Cap** options. Then select the **Custom** option and click on the **Specify Value** icon. In the **Positive Error** Value bar select the last row of values representing **Maximum Third Quartile**. Leave the **Negative Error Value** bar alone.

(q) Now click on the remaining bottom visible rectangles. In the **Chart Tools** menu, select the **Layout** submenu. In the **Layout** submenu. In the **Error Bars** menu, select the **More error bars** options.

(r) Choose the **Negative** and **Cap** options. Then select the **Custom** option and click on the **Specify Value** icon. In the **Negative Error** Value bar select the row of values representing **First Quartile- Minimum**. This will be the second calculated row. Leave the **Positive Error Value** bar alone.

(s) All that remains is to make the inside rectangles of uniform color (usually the median is not displayed in box-and-whisker plots). Highlight each row of rectangles, choose the **No Outline** option and choose an appropriately light color. Repeat this for the remaining set of rectangles, choosing the same color.

(t) You should now have side-by-side box-and-whisker plots. If a median line is desired, simply choose the appropriate border option for one of the rows of rectangles.

Creating Dot Plots in Excel

This is a much easier process. Dot plots are most useful with discrete data sets. Again, in a blank area of the worksheet, vertically list each distinct value that appears in the data set. In each cell to the right of the value use either of the nested commands:

$$= rept(., countif(dataarray, value))$$

$$= rept(., countifs(dataarray1, value1, dataarray2, value2,))$$

The $rept(textcriteria, numberoftimes)$ command converts a numerical value to the symbol given that number of times. The $countif(dataarray, value)$ counts the number of times that value appears in the given data array.

A bar can be drawn between the values and the dots by placing the appropriate borders between the two columns. A side-by-side dot plot is easily created by placing the nested commands on the opposite side.

Creating bar charts and bar graphs in Excel

The Charts menu in Excel has the capacity to directly create bar charts and bar graphs. Highlighting the appropriate columns, going to the Charts menu within the Insert menu, and selecting the appropriate options is all that is needed.

Creating Scatter Plots and Time Series Plots in Excel

To create a standard two-dimensional (regression) plot in Excel, we use scatterplots. This is a relatively easy process in excel. Although there are several ways this can be done, these instructions will highlight the most direct way (without getting into too much chart manipulation).

(a) Place the x values (independent variable, horizontal axis) and y values (dependent variable, vertical axis) side by side in two columns with the x values in the left column and the y values in the right column. (This is the default mode for Excel scatter plots.)

(b) Highlight both columns. Then go into the Charts menu within the Insert menu, and select the **Scatter** drop down menu. In this menu select the **Scatter** with **Scatter with only Markers** option. (You will add a trendline later.) If you highlight the title or caption, this will automatically appear as the title in the chart.

    Note: To create a time series plot in Excel, in the **Scatter** drop down menu select the **Scatter with Straight Lines and Markers** option. The result will have the appearance of a time series plot.

(c) Now it is a matter of adjusting the scales to make the graph a bit more pleasing to the eye. To do this, right click on appropriate parts of the graph and select the Format option in the menu that appears. You can change plot shapes, sizes, scales, etc. Click on the axes to adjust the axis scale so the dots are well proportioned in the graph area.

(d) To add a trendline, right click on any data point. Select **Add trendline** from the menu. Click on **Add equation** as well. A trendline along with the corresponding equation will appear.