⊙⊙

TIME SERIES INTRODUCTION

Documents prepared for use in courses B01.1305 and C22.0101
New York University, Stern School of Business

Cover photo: Coffee beans, Kauai, Hawaii, 2005.

⊙⊙⊙

The business world provides plenty of data in the form of time series.

The simplest form for a time series is $X_1, X_2, X_3, \ldots$ in which

X_1 is the value collected at time point 1
X_2 is the value collected at time point 2
X_3 is the value collected at time point 3
and so on.

The time points are usually evenly spaced. For example, the data could be weekly financial reports or hourly temperature readings.

The data could be daily values on an equity index. These would be unevenly spaced because of weekends and holidays. For data of this type, the analyst would watch for weekend effects.

Time series are described through statistical models that specify the random and nonrandom mechanisms that create the data. Many different statistical models have been proposed for time series. You should be aware that *the data will not come to you with a label that indicates the model*. The data will, at best, provide clues as to what type of model might have created them and thus might provide a good description. A good deal of statistical work has been invested on the problem of model identification.

There are two main types of models for statistical time series.

Time-domain models describe X_t , the value obtained at time point *t*, as related to the values obtained at other time points. Most business time series are described in time domain models.

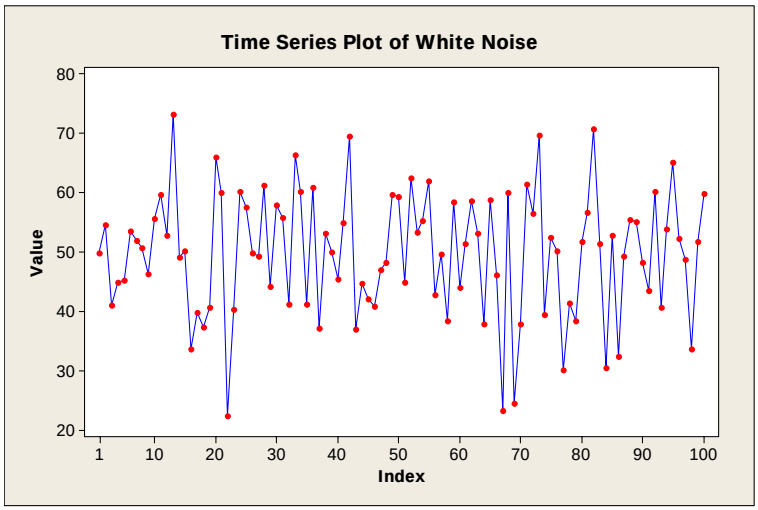
Frequency-domain models conceptualize the observations as points on a sum of cosine waves. The model $X_t = \mu + \sum_{j=1}^{5} R_j \cos\left(\omega_j t + \phi_j\right) + \varepsilon_t$ describes a sum of five cosine waves, and the statistical interest is nearly always on the wave frequencies ω_1 through ω_5. These models are especially useful in engineering, where the frequencies are interpreted as sounds or as vibrations.

Every time-domain model has an equivalent frequency-domain version, and vice versa. While most business series are analyzed through time-domain methods, there can occasionally be great benefits to considering their frequency-domain forms.

This document will consider only time-domain models. It will give definitions and examples for the most commonly used time-domain models.

1. White noise

The data series $X_1, X_2, X_3, \ldots$ consists of independent values, sampled from a population with mean μ and standard deviation σ. If the values follow a normal distribution, the series would be described as normal white noise. If the values of μ and σ are unknown, then the usual statistical interest is in estimating them.



Each value in this series was generated independently of all the others, each with a mean of 50 and a standard deviation of 10.

TECHNICAL NOTE: The white noise series has the *stationarity* property, meaning that the distribution of X_t (considered in isolation) is exactly the same for every t. An immediate consequence is that the mean of X_t and the standard deviation of X_t does not change over time.

The full definition of stationarity is that, for any positive integer k, the combined distribution of $(X_t , X_{t+1} , X_{t+2} , ..., X_{t+k})$ is exactly the same for every t.

2. Random walk

The data series $Y_1, Y_2, Y_3, \ldots$ consists of accumulated sums of white noise. If $X_1, X_2, X_3, \ldots$ is a white noise series, then

$$
\begin{aligned}
Y_1 &= X_1 \\
Y_2 = Y_1 + X_2 &= X_1 + X_2 \\
Y_3 = Y_2 + X_3 &= X_1 + X_2 + X_3 \\
Y_4 = Y_3 + X_4 &= X_1 + X_2 + X_3 + X_4 \\
&\text{and so on}
\end{aligned}
$$

This model is sometimes used (with controversy) for equity prices.

The recommended analysis for a random walk begins with differencing. Specifically, create

$\nabla Y_2 = Y_2 - Y_1$

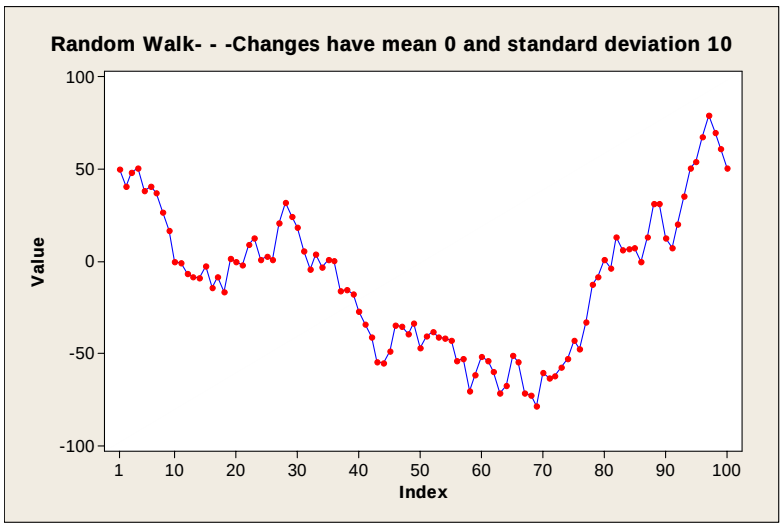
$\nabla Y_3 = Y_3 - Y_2$

$\nabla Y_4 = Y_4 - Y_3$

and so on

The series ∇Y_2 , ∇Y_3 , ∇Y_4 , ∇Y_5 , ... can then be treated as white noise. Observe these four things:

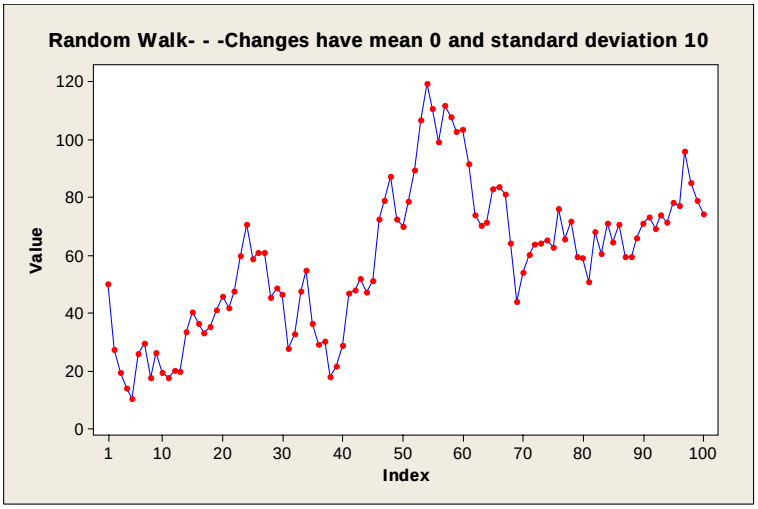
* The series ∇Y_2 , ∇Y_3 , ∇Y_4 , ∇Y_5 , ... is exactly the same as series X_2 , X_3 , X_4 , X_5 , ... That is, the differencing operation just recovers the white noise.
* The series ∇Y_2 , ∇Y_3 , ∇Y_4 , ∇Y_5 , ... has one observation fewer than the original data series. This is not a material problem, but it's an accounting nuance that one should be aware of.
* Sometime there is a nonrandom starting value Y_0 , so that the differencing can start with $\nabla Y_1 = X_1$.
* The random walk is *not* a stationary series, as the standard deviation increases with time.



The changes (here meaning X_1, X_2, X_3, X_4, ...) were generated independently with mean 0 and standard deviation 10. The mean (here 0) corresponds to a notion that most users would describe as *drift*. This series should have the property that it "goes nowhere," but this picture shows how deceptive this notion is. There are several critical points:

* Your impression depends on where you stop. This started at value 50, but ended up around 60 at time index 100, so you might call it a success (assuming that high values are good). If you had stopped your surveillance of these data at time index 70, you would have declared this a serious failure.
* Random walks can create long "waves," and you can be greatly misled by these. Notice that this series spent nearly all of its time below the starting value of 50, even though the drift was zero.
* Random walks can drop below zero, as this one did. This can be a concern for modeling equity prices, so some people prefer the log-normal random walk presented later.

Here is another result, obtained from *exactly the same model*:



3. Lognormal random walk

In the notation of the previous example, suppose that there is a series of positive values P_0 , P_1 , P_2 , P_3 , ... We assume that P_0 is nonrandom. Now form the association

$$Y_1 = \log\frac{P_0}{P_1}$$

$$Y_2 = \log\frac{P_2}{P_1}$$

$$Y_3 = \log\frac{P_3}{P_2}$$

and so on

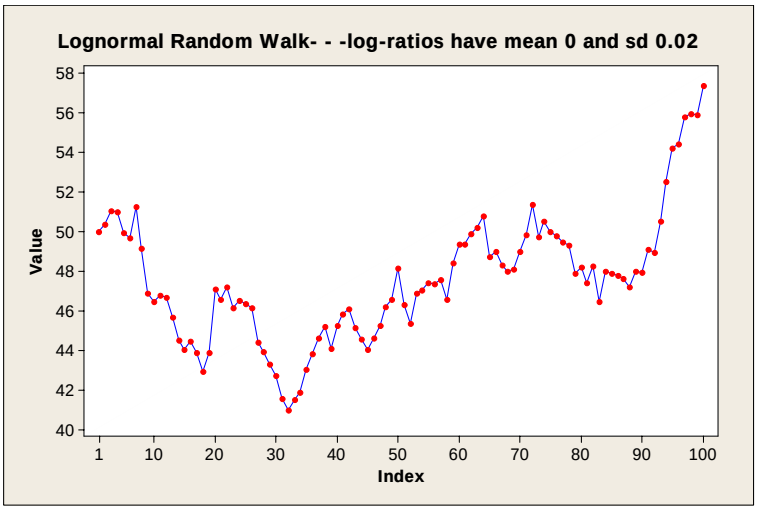
This model is often used for equity prices, with P_0 = known price on day 0, P_1 = random price on day 1, P_2 = random price on day 2, and so on. The "lognormal random walk" name applies to the price series P_0 , P_1 , P_2 , P_3 , ... Here P_n denotes the price on day n, and it can be related to the white noise X_1, X_2, X_3, ... through the equation

$$P_n = P_0 \, e^{X_1+X_2+X_3+\ldots+X_n}$$

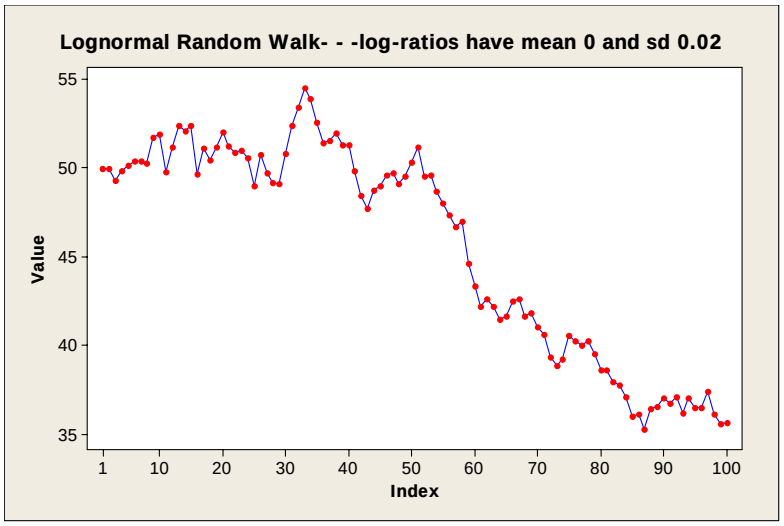
You may also see the related forms $\frac{P_t}{P_{t-1}} = e^{X_t}$, $X_t = \log\frac{P_t}{P_{t-1}}$, and $\log\frac{P_n}{P_0} = X_1 + X_2 + \ldots + X_n$.

The lognormal random walk is *not* a stationary series.

A lognormal random walk can never turn negative. The behavior of this model depends dramatically on the mean and standard deviation of the log-ratio random variables X_1, X_2, X_3, ...

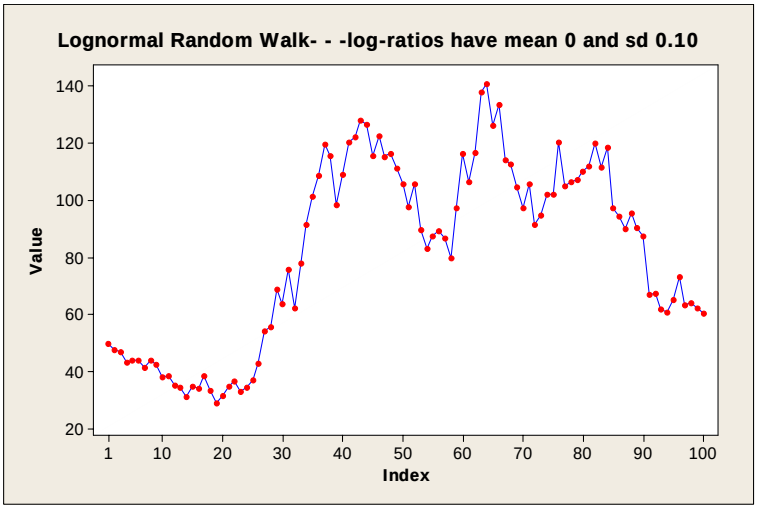


You can think of the "Value" here as being an equity price. The series above would be regarded as a success, in that the price advanced from 50 to about 58.



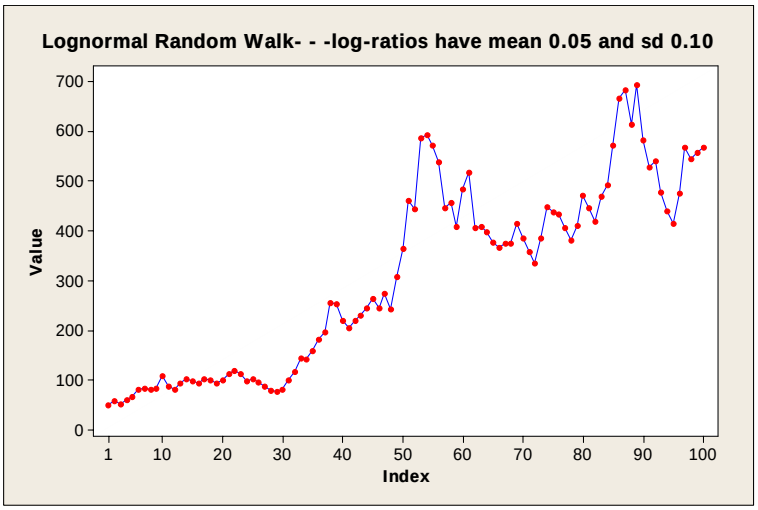
This second illustration of the lognormal random walk was created with exactly the same parameters, but it would have to be called a failure.

The standard deviation is clearly related to the volatility. The next picture shows the same model, with the standard deviation of 0.02 replaced by standard deviation 0.10.

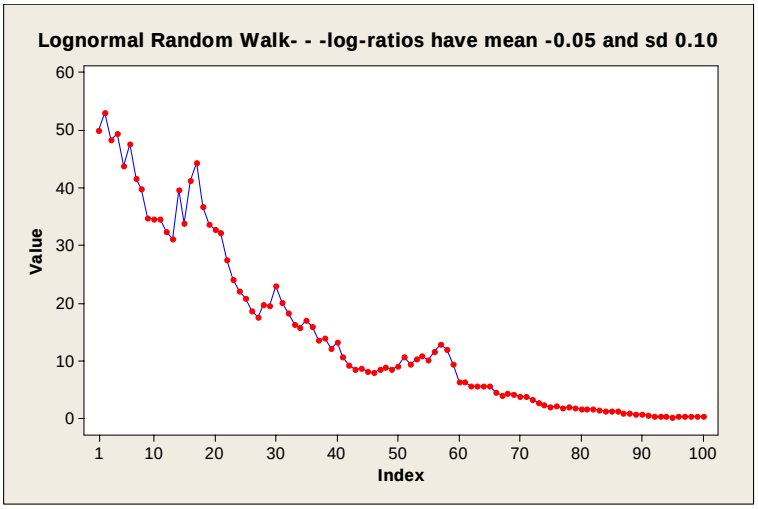


The vertical scale here is much wider than in the previous picture!

If the drift, the mean, is (very) different from zero, the results can be quite dramatic. Here is a case with positive drift:



Here is an illustration with negative drift:



4. Autoregressive, order 1 (AR1)

Examples 1, 2, and 3 are white noise or convertible to white noise. The model discussed here is an intellectual leap forward. The model starts with a nonrandom value X_0. Thereafter,

$$X_t = \rho X_{t-1} + \varepsilon_t \qquad [4a]$$

This says that the value obtained at time t is a multiple of the value at time t - 1, plus an added random noise term. The set of noise terms ε_1 , ε_2 , ε_3 , ... is assumed to be white noise, with a mean of zero. In addition, it is assumed that ε_t is also independent of X_0 , X_1 , ..., X_{t-1} . This model is only useful in the case of stationarity. For reasons of stationarity, as will be made clear below, it is necessary to assume that $-1 < \rho < 1$. Here ρ is called the autoregressive parameter.

You will often see the AR1 model given with a mean term:

$$X_t - \mu = \rho (X_{t-1} - \mu) + \varepsilon_t \qquad [4b]$$

In this form, E X_t = expected values of $X_t = \mu$ at every time point.

Since [4b] can be written as $X_t = \mu(1 - \rho) + \rho X_{t-1} + \varepsilon_t$, you may also see this model in form

$$X_t = \nu + \rho X_{t-1} + \varepsilon_t \qquad [4c]$$

TECHNICAL NOTE: With $\rho = 1$, this is a random walk. In any of [4a] or [4b] or [4c] with $\rho = 1$, the model is $X_t = X_{t-1} + \varepsilon_t$, which was discussed under point 2. Thus,

$X_1 = X_0 + \varepsilon_1$

$X_2 = X_1 + \varepsilon_2 \quad = X_0 + \varepsilon_1 + \varepsilon_2$

$X_3 = X_2 + \varepsilon_3 \quad = X_0 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3$

$X_4 = X_3 + \varepsilon_4 \quad = X_0 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$

and so on

TECHNICAL NOTE: Why can we not have $\rho > 1$? In terms of just modeling, we do have the freedom to create any model we desire, but $\rho > 1$ creates some consequences that we might wish to avoid.

Use form [4c] and investigate the variance of X_t . We will assume that $\sigma^2 = \text{Var}(\varepsilon_t)$ for every time point t.

$$\begin{aligned} \text{Var}(X_t) &= \text{Var}(\nu + \rho X_{t-1} + \varepsilon_t) \\ &= \text{Var}(\rho X_{t-1} + \varepsilon_t) && \text{since } \nu \text{ is not random} \\ &= \text{Var}(\rho X_{t-1}) + \text{Var}(\varepsilon_t) && \text{since } \varepsilon_t \text{ is independent of } X_{t-1} \\ &= \rho^2 \text{Var}(X_{t-1}) + \sigma^2 \end{aligned}$$

If we have $\rho > 1$ or $\rho < -1$, then certainly $\rho^2 > 1$. This would have $\text{Var}(X_t)$ growing to infinity at an exponential rate. This is almost certainly *not* a property that we want a model to have.

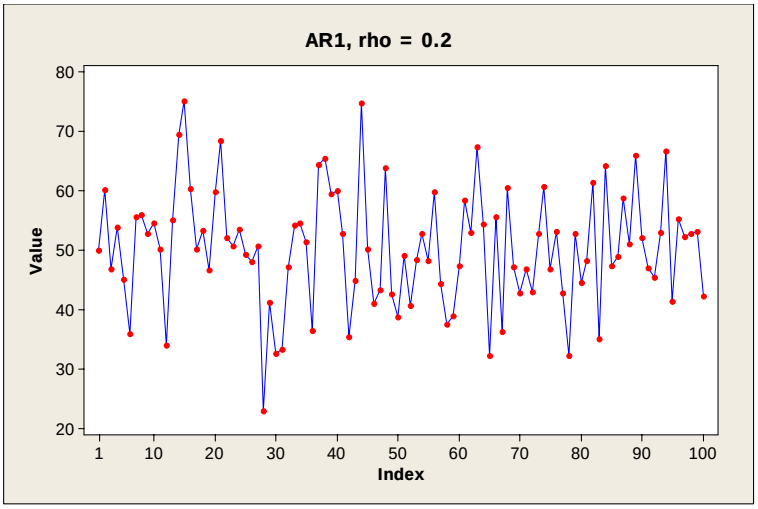
With $-1 < \rho < 1$, we can have $\text{Var}(X_t)$ the same for every value of t. Let's say $\text{Var}(X_t) = \tau^2$. Then $\tau^2 = \rho^2 \tau^2 + \sigma^2$, and

$$\tau^2 = \frac{\sigma^2}{1 - \rho^2} \qquad [4d]$$

The appeal of the AR1 model is easily grasped. It says that our statistical performance on Thursday depends on what we did Wednesday (but *not* directly on what we did Tuesday, Monday, Sunday, Saturday, ...), plus a little random noise. Models of this form are called *Markovian*, meaning that they depend on all of past history only through the most recent value.

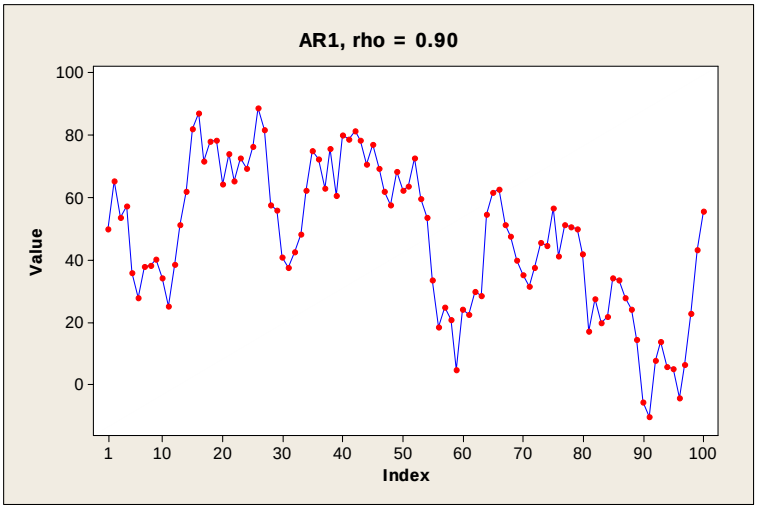
In the graph of an AR1 time series, μ and σ are just scaling parameters, while the parameter ρ dictates the appearance. A serious user will want to estimate μ and σ, but ρ is the most interesting parameter.

Here is a picture with ρ = 0.2:



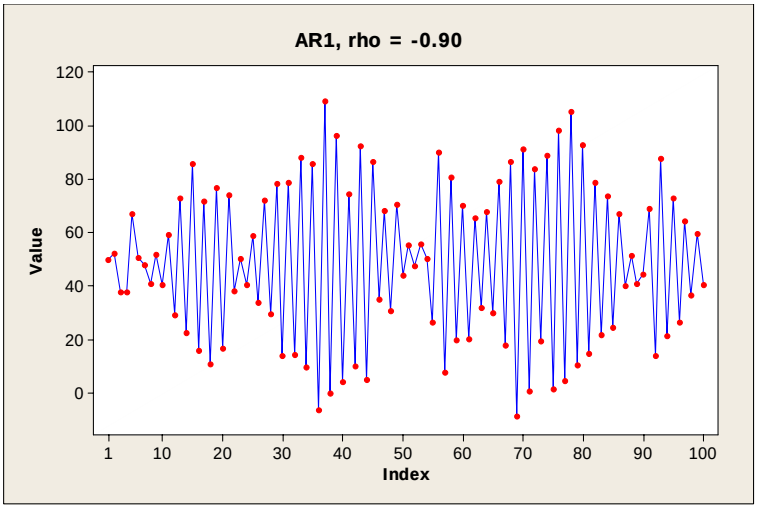
The AR1 time series will always fluctuate around the same value. In the picture above, that value is 50. Each data value is roughly similar to the previous value, but there is plenty of variability.

Here is a picture with ρ = 0.90:



Here each data value is *very* similar to the previous value. Graphs of AR1 series with ρ near 1.0 tend to produce long waves. When ρ is near 1.0, the appearance will resemble that of a random walk. Note that up to time index 50, nearly all of the data values exceed the starting value of 50. This property makes it very difficult to estimate the mean μ.

Just for the sake of amusement, here is an AR1 series with ρ = -0.90:



When ρ is close to -1.0, the data will oscillate, and the picture above is very typical.

For ρ negative, but just below zero, the picture would not be so extreme.

There are real-data AR1 series with negative ρ, and these might arise in games for which the time index refers to turns of play. The sequence of distances achieved by a golfer practicing at a driving range could be such a situation. We'd let X_1 = distance on first ball, X_2 = distance on second ball, and so on.

> TECHNICAL NOTE: If the time index is clock time, you should be suspicious of any time series that is modeled as AR1 with a negative ρ. Suppose that $X_1, X_2, X_3, \ldots$ represents a sequence of equity prices at the end of the trading day. An AR1 model with negative ρ would seem to say that the price tends to rebound from its performance on the previous day. This would, however, create a situation in which $X_1, X_3, X_5, X_7, \ldots$ is AR1 with positive ρ and $X_1, X_4, X_7, X_{10}, X_{13}, \ldots$ is again AR1 with negative ρ. The possibility that ρ is negative thus says that the apparent behavior of the series can be materially altered just by changing the time spacing of the observations.

If you believe that a time series is reasonably described as AR1, you can estimate ρ by a simple linear regression. Just regress { X_t } on { X_{t-1} }.

Here are explicit instructions for doing this in Minitab 14. Suppose that the time series appears as column C1 and that it has length *n*. Use **Stat** ⇒ **Time Series** ⇒ **Lag** and then set up the panel as follows:

The first entry of the lagged column, C2 in this example, will have the missing data code * in its first position. Now ask for the simple linear regression of C1 on C2. The slope coefficient in this regression is the estimate of ρ.

> TECHNICAL NOTE: The *X* values in an AR1 series are statistically dependent. It can be shown that Corr(X_{t+u} , X_t) = $\rho^{-|u|}$. The absolute value merely allows the use of negative *u*'s.

> TECHNICAL NOTE: The AR1 series writes each X_t in terms of the previous X_{t-1} and an independent noise term. The series can also be represented as a linear combination of *all* past noise terms. Re-examine [4b]:

$$X_t - \mu = \rho\,(X_{t-1} - \mu) + \varepsilon_t$$

> Now rewrite this stepping down the time index from *t* back to *t* – 1:

$$X_{t-1} - \mu = \rho\,(X_{t-2} - \mu) + \varepsilon_{t-1}$$

Substitute the second equation into the first to produce this:

$$X_t - \mu = \rho \{ \rho (X_{t-2} - \mu) + \varepsilon_{t-1} \} + \varepsilon_t$$

$$= \varepsilon_t + \rho \varepsilon_{t-1} + \rho^2 (X_{t-2} - \mu)$$

Express now $X_{t-2} - \mu$ in terms of $X_{t-3} - \mu$ and substitute into the equation just above. This will produce

$$X_t - \mu = \rho \{ \rho (X_{t-2} - \mu) + \varepsilon_{t-1} \} + \varepsilon_t$$

$$= \varepsilon_t + \rho \varepsilon_{t-1} + \rho^2 \varepsilon_{t-2} + \rho^3 (X_{t-3} - \mu)$$

We can extend this argument indefinitely far into the past. This shows that X_t is combination of ε_t and all the previous ε's. If you create the mathematical fiction that the series extends back to time $-\infty$, you can write $X_t - \mu = \sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j}$. In this form X_t is an infinite combination of past ε's.

TECHNICAL NOTE: Autoregressive series can be extended to higher orders. The AR2 model, in form similar to [4b], is

$$X_t - \mu = \rho_1 (X_{t-1} - \mu) + \rho_2 (X_{t-2} - \mu) + \varepsilon_t$$

The general AR*p* model is

$$X_t - \mu = \rho_1 (X_{t-1} - \mu) + \rho_2 (X_{t-2} - \mu) + \rho_3 (X_{t-3} - \mu) + \ldots$$

$$+ \rho_p (X_{t-p} - \mu) + \varepsilon_t$$

In this form X_t can be written as an infinite combination of all past ε's. The structure of the coefficients is much more complicated for the AR*p* model than for the AR1 model.

5. Moving average (MA)

Suppose that $\varepsilon_0, \varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots$ is a white noise series. (The mean does not necessarily have to be zero.) Then the series

$$X_1 = a_0\,\varepsilon_1 \;-\; a_1\,\varepsilon_0$$

$$X_2 = a_0\,\varepsilon_2 \;-\; a_1\,\varepsilon_1$$

$$X_3 = a_0\,\varepsilon_3 \;-\; a_1\,\varepsilon_2$$

$$X_4 = a_0\,\varepsilon_4 \;-\; a_1\,\varepsilon_3$$

and so on

is called a moving average of extent 1. We identify this as MA1. In what follows next, we'll assume $a_0 = 1$. (If we don't fix a_0 or a_1 then we will not be able to disentangle a_0, a_1, and $\sigma = SD(\varepsilon_t)$.) Later we will restore a_0 .

Certainly X_1 and X_2 are statistically dependent, since both depend on ε_1 . However X_1 and X_3 are independent; note that X_1 depends on ε_0 and ε_1 , while X_3 depends on ε_2 and ε_3 .

TECHNICAL NOTE: The MA1 series can be written in the form of an infinite autoregression. Start with

$$X_t = \varepsilon_t - a_1\,\varepsilon_{t\text{-}1}$$

Use the relationship $X_{t\text{-}1} = \varepsilon_{t\text{-}1} - a_1\,\varepsilon_{t\text{-}2}$ to recover

$$\varepsilon_{t\text{-}1} = X_{t\text{-}1} + a_1\,\varepsilon_{t\text{-}2}$$

Substitute this into the previous to obtain

$$X_t = \varepsilon_t - a_1\,\{\;X_{t\text{-}1} + a_1\,\varepsilon_{t\text{-}2}\;\} = \;\varepsilon_t\; - a_1\;X_{t\text{-}1}\;\;-\;\;a_1^2\;\varepsilon_{t-2}$$

Then use $X_{t\text{-}2} = \varepsilon_{t\text{-}2} - a_1\,\varepsilon_{t\text{-}3}$ to get

$$\varepsilon_{t\text{-}2} = X_{t\text{-}2} + a_1\,\varepsilon_{t\text{-}3}$$

Substitue for $\varepsilon_{t\text{-}2}$, giving

$$X_t \;\;=\;\varepsilon_t\; - a_1\;X_{t\text{-}1}\;\;-\;\;a_1^2\;\{X_{t-2} + a_1\varepsilon_{t-3}\}$$

$$=\;\;\varepsilon_t\;\;-\;\;a_1X_{t-1}\;\;-\;\;a_1^2\;X_{t-2}\;\;-\;\;a_1^3\;\varepsilon_{t-3}$$

If we create the fiction that the series times index goes all the way back to -∞, we can write $X_t = \varepsilon_t - \sum_{j=1}^{\infty} a_1^j X_{t-j}$. This creates the MA1 series as an infinite autoregression.

TECHNICAL NOTE: The MA2 series is $X_t = \varepsilon_t - a_1 \varepsilon_{t-1} - a_2 \varepsilon_{t-2}$. The general form MA$q$ is $X_t = \varepsilon_t - a_1 \varepsilon_{t-1} - a_2 \varepsilon_{t-2} - a_2 \varepsilon_{t-3} - \ldots - a_q \varepsilon_{t-q}$. This can also be written as an infinite autoregression, but the coefficients are more complicated.

The analyst who works with time series will nearly always start with nothing but the data. The analyst will have to make a decision as to what kind of time series it is (white noise? autoregressive? moving average? random walk?). He or she will also have to decide the order (ARp for what p? MAq for what q?) In addition, the various coefficients will have to be estimated.

Moving average series, on the other hand, are sometimes produced intentionally. Many government-produced data series are given as moving averages. This is done from observed ε_1 , ε_2 , … (not even necessarily white noise) and, using specified a's, produces the X 's for public consumption. For example, data that are acquired as monthly ε's can be put through a twelve-month moving average to make resulting X 's that have smoothed out monthly effects.

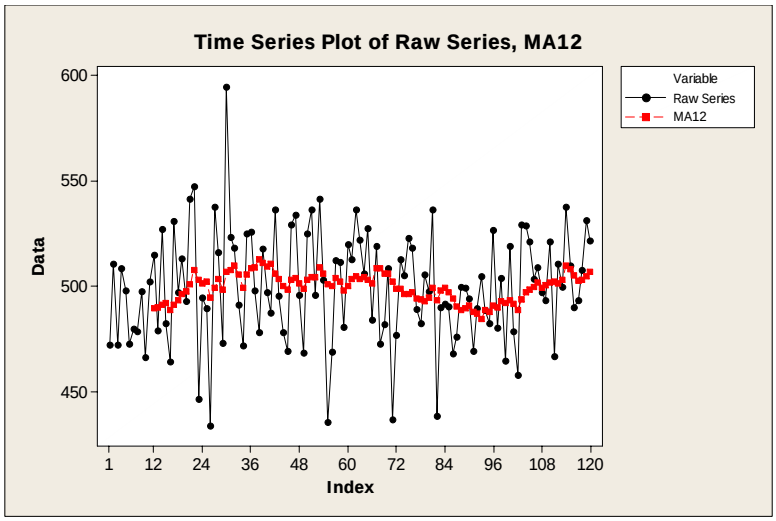
In summary, the discussion on moving average series is done on two levels:

(1) The analyst gets the X 's as an observed series. He or she never gets to see the ε's and may not even succeed in figuring out that the data are MAq (or anything else). After claiming that the series is MAq, the analyst still needs to estimate a_1 through a_q.

(2) The X series is presented as a specified moving average. The extent q will be identified, the coefficients a_1 through a_q will be available, and the original series of ε's (not necessarily white noise) will also be available. (The analyst might need to make a special request to get these.)

TECHNICAL NOTE: In a twelve-month moving average, we usually use equal weights, as in $X_t = \frac{1}{12}\varepsilon_t + \frac{1}{12}\varepsilon_{t-1} + \ldots + \frac{1}{12}\varepsilon_{t-11}$. In this form, $a_0 = \frac{1}{12} = a_1 = \ldots = a_{11}$. However, there is no requirement to do so.

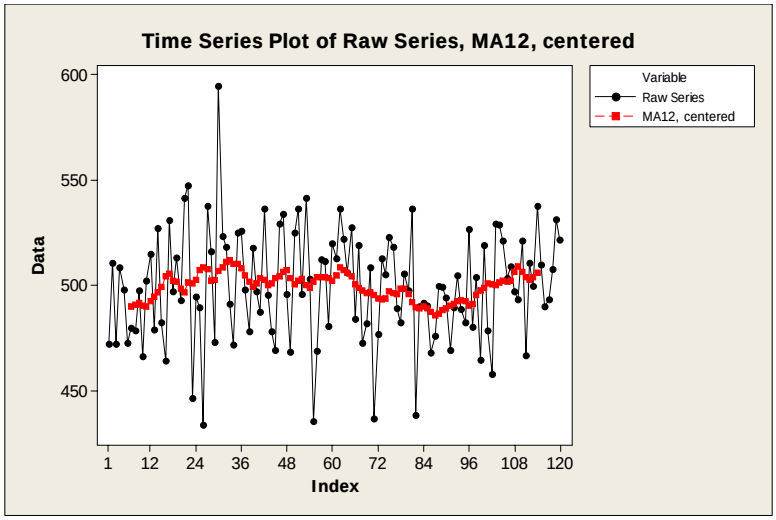
The following pictures, produced by Minitab, show that a twelve-month moving average greatly damps down erratic behavior. It's much easier to grasp the behavior of the squares than of the dots.



This picture was produced with all weights equal to $\frac{1}{12}$. The first 12 values of the raw series are averaged together to produce X_{12}. That is,

$$X_{12} = \frac{1}{12}\varepsilon_1 + \frac{1}{12}\varepsilon_2 + \ldots + \frac{1}{12}\varepsilon_{12}$$

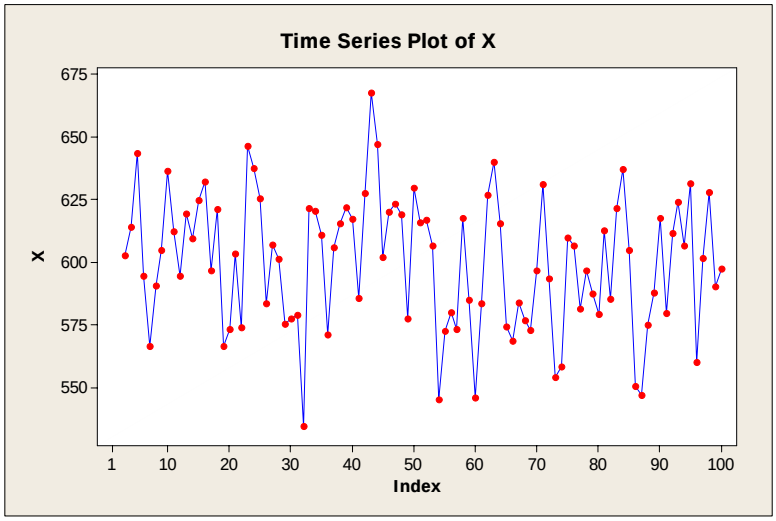
At times we like to align the X indices at the centers of the values that were averaged. This is a mere accounting nuance. Here is (almost) the same data with that feature.



For this picture, first square for the moving average is produced at time point 7. The "almost" in the sentence above the graph refers to a computing convention when the moving average is to be taken over an even number of time points. Here this means

$$X_7 = \frac{1}{24}\varepsilon_1 + \underbrace{\frac{1}{12}\varepsilon_2 + \frac{1}{12}\varepsilon_3 + \frac{1}{12}\varepsilon_4 \ldots + \frac{1}{12}\varepsilon_{12}}_{\text{Weight } \frac{1}{12} \text{ used on 11 values centered at } \varepsilon_7} + \frac{1}{24}\varepsilon_{13}$$

Consider this picture:



This plot shows an MA2 series with a_1 = -0.4 and a_2 = +0.2. Identifying time series types from their graphs is not easy.

6. Hybrid models

It is possible to form models that combine the features of autoregressive and moving average series. Consider

$$\sum_{h=0}^{p} \rho_h \left(X_{t-h} - \mu \right) = \sum_{j=0}^{q} a_j \, \varepsilon_{t-j}$$

The left side (usually with $\rho_0 = 1$) is part of an autoregression of order p, and the right side (usually with $a_0 = 1$) is part of a moving average of order q. This particular model is called ARMA (p, q), meaning autoregressive moving average of orders p and q.

The analyst with a series of unknown type will often try to identify it as an ARMA (p, q). The challenge includes the identification of all the unknown coefficients, the ρ's and the a's.

❁❁❁ MULTIPLE REGRESSION DATA COLLECTED AS TIME SERIES ❁❁❁

Time series in regression present some challenging problems. Let's suppose that we have the simple linear regression model $Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ for $t = 1, 2, ..., n$.

> We've used t as the subscript instead of the more conventional i. This was done to suggest that the data were collected in time order with (x_1, Y_1) first, then (x_2, Y_2), then (x_3, Y_3), and so on.
>
> We've used lower case x to suggest nonrandom values, along with upper case Y to suggest random values. This is a non-binding suggestion, and you will find other conventions regarding upper case and lower case symbols.
>
> The spreadsheet holding the data will have a column for x, a column for Y, and almost certainly also a column that identifies the time. The values in this column could just be the sequence 1, 2, ..., n or they could be Jan 1981, Feb 1981, Mar 1981, Apr 1981, ..., Nov 2004, Dec 2004.
>
> If the time column has real dates, it may be helpful to create a column with consecutive integers. This is used in Solution 4 below. The correspondence between the dates and numbers should be noted; in the example just above, we'd note 1 $\Leftrightarrow$ Jan 1981, 2 $\Leftrightarrow$ Feb 1981, and so on.
>
> The time series problems can occur in either simple regression (one predictor) or in multiple regression (two or more predictors). However, the statistical issues are exactly the same, and we will use a simple regression to illustrate the ideas.

In doing the regression work, we think of the values $x_1, x_2, \ldots, x_n$ as non-random, even though they are really a time series. The real problem with the time series regression is that the noise terms $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ will be a time series, instead of being statistically independent. The most plausible time series model for these noise terms is AR1, autoregressive of order 1.

The data shown on the next page are the CO_2 emissions for Australia for the years 1950 to 1997. These data clearly constitute a time series. Suppose that we do the regression of CO_2 emissions on Solid Fuels. The regression looks routine:

```
The regression equation is
CO2Emissions = - 4252 + 1.94 SolidFuels

Predictor      Coef  SE Coef      T      P
Constant      -4252     1454  -2.92  0.005
SolidFuels  1.94260  0.05240  37.08  0.000

S = 3868.89   R-Sq = 96.8%   R-Sq(adj) = 96.7%

Analysis of Variance
Source          DF           SS           MS        F      P
Regression       1  20574813518  20574813518  1374.56  0.000
Residual Error  46    688542533     14968316
Total           47  21263356051
```

(Discussion continues on the page following the data.)

YEAR	CO2 Emissions	Solid Fuels
1950	14941	12028
1951	16112	12581
1952	16432	12835
1953	16223	13163
1954	18517	13956
1955	19291	13987
1956	19934	13986
1957	20340	14090
1958	21184	14371
1959	22849	15472
1960	24052	16083
1961	24703	16368
1962	25883	16781
1963	27551	17393
1964	29719	18323
1965	32988	19394
1966	32814	19487
1967	35251	20580
1968	36712	20902
1969	38793	21282
1970	38888	20277
1971	40011	20268
1972	41238	21216
1973	43814	21756

YEAR	CO2 Emissions	Solid Fuels
1974	44170	23254
1975	45199	23729
1976	47009	24255
1977	50697	26250
1978	51490	25665
1979	52433	26468
1980	55348	28066
1981	58365	28886
1982	59536	29676
1983	56734	29173
1984	59398	30221
1985	60863	32572
1986	60909	32171
1987	64656	35130
1988	65799	35881
1989	69898	38765
1990	72601	39791
1991	69886	40344
1992	74412	42561
1993	76422	41465
1994	78886	43934
1995	79989	45281
1996	85936	48973
1997	86336	50875

The CO2Emissions data is the country's total emissions, and the SolidFules data is the emissions component from burning solid fuel. Both variables are in units of thousands of metric tons.

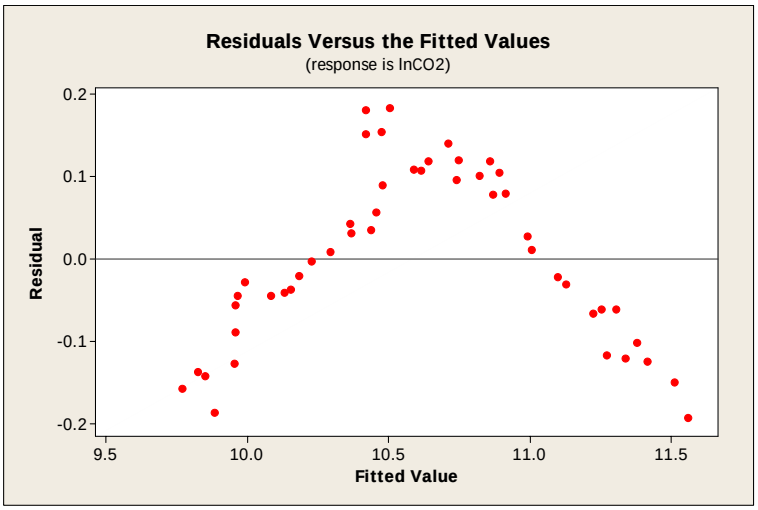
The next page shows the base-*e* logarithms of these values.

YEAR	ℓn(CO2)	ℓn(Solid)
1950	9.6119	9.3950
1951	9.6873	9.4399
1952	9.7070	9.4599
1953	9.6942	9.4852
1954	9.8264	9.5437
1955	9.8674	9.5459
1956	9.9002	9.5458
1957	9.9203	9.5532
1958	9.9610	9.5730
1959	10.0367	9.6468
1960	10.0880	9.6855
1961	10.1147	9.7031
1962	10.1613	9.7280
1963	10.2238	9.7638
1964	10.2995	9.8159
1965	10.4039	9.8727
1966	10.3986	9.8775
1967	10.4702	9.9321
1968	10.5109	9.9476
1969	10.5660	9.9656
1970	10.5684	9.9172
1971	10.5969	9.9168
1972	10.6271	9.9625
1973	10.6877	9.9876

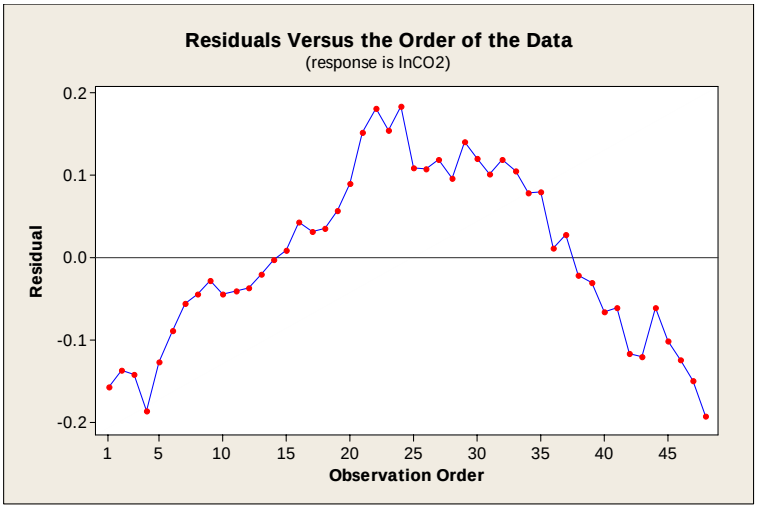
YEAR	ℓn(CO2)	ℓn(Solid)
1974	10.6958	10.0542
1975	10.7188	10.0745
1976	10.7581	10.0964
1977	10.8336	10.1754
1978	10.8491	10.1529
1979	10.8673	10.1837
1980	10.9214	10.2423
1981	10.9745	10.2711
1982	10.9943	10.2981
1983	10.9461	10.2810
1984	10.9920	10.3163
1985	11.0164	10.3912
1986	11.0171	10.3788
1987	11.0768	10.4668
1988	11.0944	10.4880
1989	11.1548	10.5653
1990	11.1927	10.5914
1991	11.1546	10.6052
1992	11.2174	10.6587
1993	11.2440	10.6326
1994	11.2758	10.6904
1995	11.2896	10.7206
1996	11.3614	10.7990
1997	11.3660	10.8371

These data will be discussed in logarithm terms. The original values showed variability proportional to size; the values for CO_2 and Solid moved around by hundreds in the early years and by thousands in the later years. This is a firm indication of the need for logarithms.

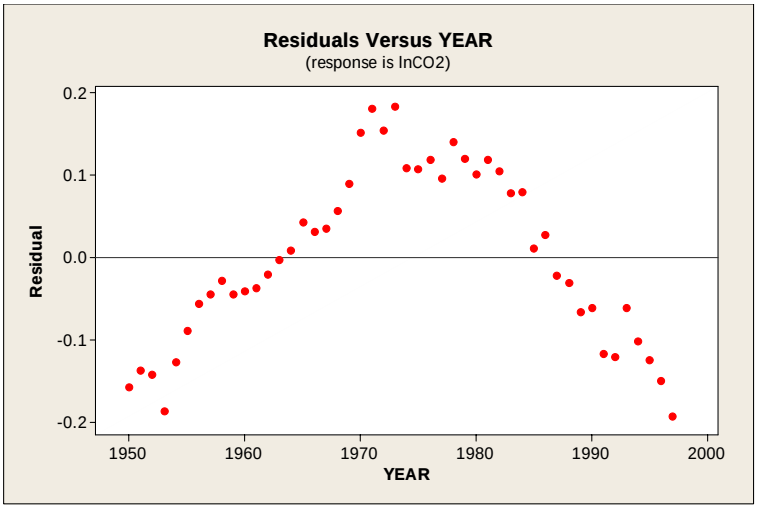
The regression of ℓnCO2 on ℓnSolid seems to have problems with the residuals. Here is the residual versus fitted plot:



It's tempting to just say that this is a problem of curvature. Curvature could be cured by using also $(\ell\text{nSolid})^2$ as a predictor. However there are other clues. Suppose that we ask for the residuals in time sequence. This is available in Minitab through **Stat** ⇒ **Regression** ⇒ **Regression** ⇒ **Graphs** ⇒ **Residuals versus order**. The result is this:



Since the data file has a column for Year, you could also do this as **Stat ⇒ Regression ⇒ Regression ⇒ Graphs ⇒ Residuals versus the variables**, naming Year in the selection box. This would produce



The information is identical, but the "versus order" option connects the dots. The "versus the variables" option has better labels on the horizontal axis.

In any case, we see that e_t = residual at time t very closely resembles e_{t-1} = residual at time t-1. This of course violates our regression assumptions. It's not a simple case of curvature! In the regression context, this is almost certainly a case of autocorrelated errors.

Minitab provides a routine test for this problem, the Durbin-Watson statistic. This statistic should ***always*** be requested with time series data. This is available through **Stat ⇒ Regression ⇒ Regression ⇒ Options ⇒ Durbin-Watson statistic**. For these data, you get the result is

```
Durbin-Watson statistic = 0.0947725
```

The target value for the statistic is 2. That is, a value near 2 suggests the *absence* of an autocorrelation problem. Lower values indicate serious autocorrelation. Minitab does not provide a *p*-value for this statistic, so that you will need to consult a statistical table. A plausible approximate cutoff for concern is 1.2, meaning that you should worry about autocorrelation when the Durbin-Watson statistic is below 1.2. Certainly the value obtained here, 0.0947725, suggests that the autocorrelation problem is very serious.

> Tables of *DW*, the Durbin-Watson statistic, will provide two cutoffs, c_{lower} and c_{upper}. If $DW > c_{\text{upper}}$, then there is no problem related to autocorrelation. If $DW < c_{\text{lower}}$, then there is significant autocorrelation. The intermediate story $c_{\text{lower}} \le DW \le c_{\text{lower}}$ is inconclusive.

It is possible to get values of *DW* noticeably larger than 2. The theoretic upper limit is 4, but you will probably never see a value of *DW* as large as 3. This would suggest an AR1 process for the noise terms with negative autocorrelation, and this is logically implausible.

Regression data with a low Durbin-Watson statistic requires a repair. There are several possible solutions.

SOLUTION 1: Difference the data. Just let

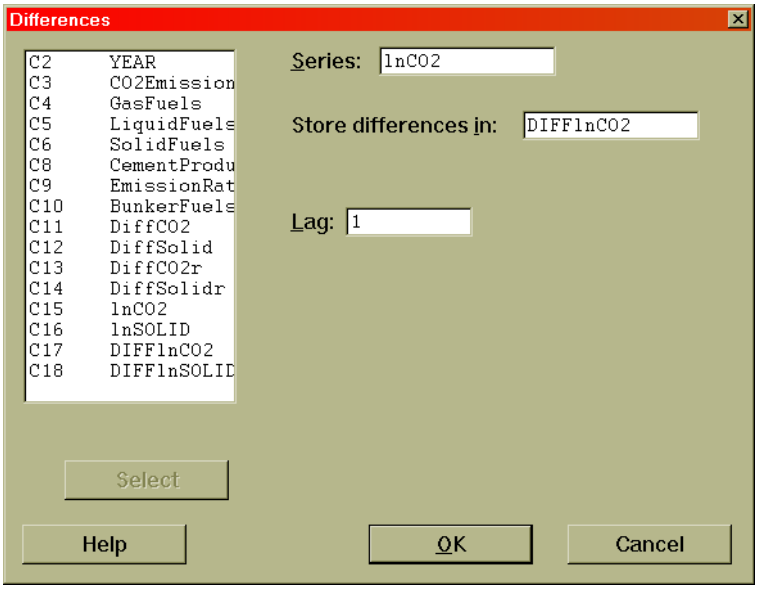
$$\begin{cases} Y_t^* = Y_t - Y_{t-1} \\ x_t^* = x_t - x_{t-1} \end{cases}$$

Some people would write this as

$$\begin{cases} Y_t^* = \nabla Y_t \\ x_t^* = \nabla x_t \end{cases}$$

This use the "del" symbol ∇ to denote differences.

In the multiple regression context, this differencing would be done to the dependent variable and to *all* the independent variables. In Minitab, this operation can be done as **Stat** ⇒ **Time Series** ⇒ **Differences**. You might fill in the information panel as follows:



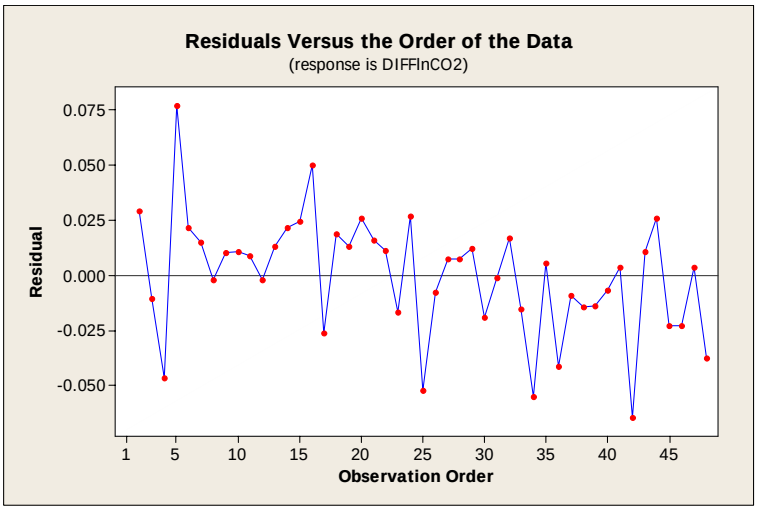
This should be done for each of the variables in the regression. The first data point, year 1950 in this example, will be noted as missing.

The regression should now be done as Y^* on X^*, meaning DiffℓnCO2 on DiffℓnSolid. In this regression, you must still check the plot of the residuals in time order, and you must compute the Durbin-Watson statistic. The output is this:

Regression Analysis: DIFFlnCO2 versus DIFFlnSOLID

```
The regression equation is
DIFFlnCO2 = 0.0177 + 0.641 DIFFlnSOLID

47 cases used, 1 cases contain missing values


Predictor        Coef   SE Coef     T      P
Constant     0.017656  0.005722  3.09  0.003
DIFFlnSOLID    0.6409    0.1334  4.80  0.000

S = 0.0274007   R-Sq = 33.9%   R-Sq(adj) = 32.4%

Analysis of Variance

Source          DF        SS        MS      F      P
Regression       1  0.017323  0.017323  23.07  0.000
Residual Error  45  0.033786  0.000751
Total           46  0.051109

Unusual Observations

Obs  DIFFlnSOLID  DIFFlnCO2       Fit   SE Fit  Residual  St Resid
  5       0.0585    0.13226   0.05515  0.00545   0.07711      2.87R
 21      -0.0484    0.00245  -0.01335  0.01128   0.01579      0.63 X
 34      -0.0171   -0.04821   0.00670  0.00752  -0.05491     -2.08R
 42       0.0138   -0.03811   0.02650  0.00459  -0.06462     -2.39R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 2.03940
```

Here is the plot of the residuals in time order:



This is an excellent outcome. The Durbin-Watson statistic is close to 2.0, showing that the autocorrelation problem has been cured. The regression fits very well. The slope coefficient of 0.641 indicates that proportional changes in Solid are associated with smaller proportional changes in CO_2 , and in the same direction.

The solution by differencing is sometimes called pre-whitening. It's a clear attempt to convert random walks back to white noise.

It's important to understand what is happening when we take differences of logarithms in a time series. Since $\nabla Y_t = Y_t - Y_{t-1}$, it follows that

$$\nabla (\ell\text{n } CO2_t) = (\ell\text{n } CO2_t) - (\ell\text{n } CO2_{t-1}) = \ell\text{n} \frac{CO2_t}{CO2_{t-1}}$$

$$= \ell\text{n} \left(1 + \frac{CO2_t - CO2_{t-1}}{CO2_{t-1}} \right) = \ell\text{n} \left(1 + \left[\begin{array}{c} \text{proportional change} \\ \text{from time } t-1 \text{ to time } t \end{array} \right] \right)$$

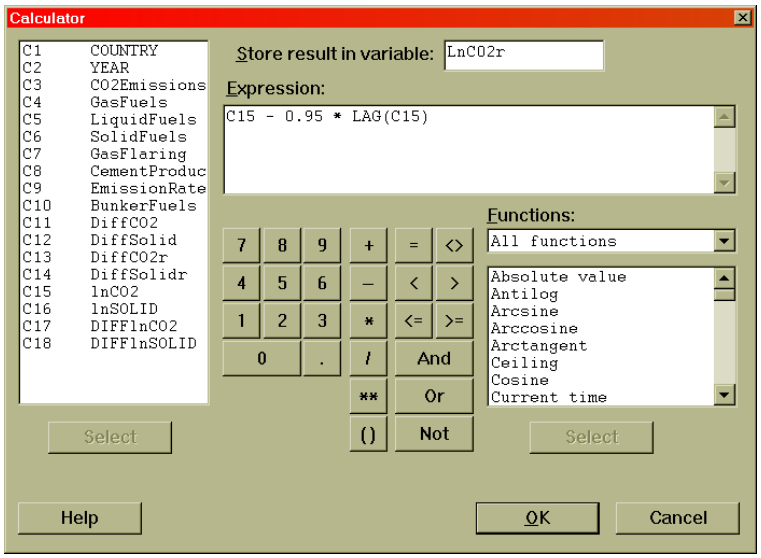
As a plausible approximation, $\ell\text{n}(1 + q) \approx q$. This works for q near zero, say for $-0.10 < q < +0.10$. When the consecutive changes tend to stay within ± 10%, then analyzing the proportional changes will give pretty much the same result as analyzing the differences of the logarithms.

SOLUTION 2: Estimate the autocorrelation coefficient and adjust it away. You can use complicated methods to estimate, but a quick simple estimate is $\hat{\rho} = 1 - \frac{DW}{2}$. Then compute

$$\begin{cases} Y_t^{\hat{\rho}} = Y_t - \hat{\rho}\, Y_{t-1} \\ x_t^{\hat{\rho}} = x_t - \hat{\rho}\, x_{t-1} \end{cases}$$

Then regress $Y^{\hat{\rho}}$ on $x^{\hat{\rho}}$. In our example, $\hat{\rho} = 1 - \frac{0.0947725}{2} \approx 0.95$. For this particular example, $\hat{\rho}$ is rather close to 1, so the end result will be very similar to simple differencing.

Let's use DiffLnCO2r as the name for $Y_t^{\hat{\rho}} = Y_t - \hat{\rho}\, Y_{t-1}$. Minitab can get this through **Calc ⇒ Calculator**. Set up the panel like this:



Perform a similar operation to create LnSOLIDr.

The regression of LnCO2r on LnSOLIDr produces this:

Regression Analysis: LnC02r versus LnSOLIDr

```
The regression equation is
LnC02r = 0.202 + 0.685 LnSOLIDr


47 cases used, 1 cases contain missing values

Predictor     Coef  SE Coef     T      P
Constant   0.20164  0.05199  3.88  0.000
LnSOLIDr   0.68453  0.09727  7.04  0.000


S = 0.0249415   R-Sq = 52.4%   R-Sq(adj) = 51.3%

Analysis of Variance

Source          DF        SS        MS      F      P
Regression       1  0.030810  0.030810  49.53  0.000
Residual Error  45  0.027993  0.000622
Total           46  0.058804

Unusual Observations

Obs  LnSOLIDr   LnC02r      Fit   SE Fit  Residual  St Resid
  4     0.498  0.47255  0.54270  0.00498  -0.07015     -2.87R
  5     0.533  0.61697  0.56633  0.00364   0.05064      2.05R
 42     0.543  0.52152  0.57360  0.00377  -0.05207     -2.11R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 2.32547
```

The graph of the residuals in time order is similar to that of SOLUTION 1, and it will not be shown.

You might observe that the R^2 in SOLUTION 2 was 52.4%, substantially better than the 33.9% of SOLUTION 1.

> The R^2 value in the original regression was 96.8%. We cannot use that original regression as the assumptions of the regression model were violated.

SOLUTION 3: Convert the problem to generalized least squares. This is a high-power method, and it requires the construction of the model in vector-matrix notation. This goes under a number of names, like Cochrane-Orcutt. It will not be discussed here.

SOLUTION 4: Use time itself as an additional independent variable. This solution is very simple to implement, but it's only occasionally successful. Here we'll just regress CO2Emissions on (SolidFuel, Year).

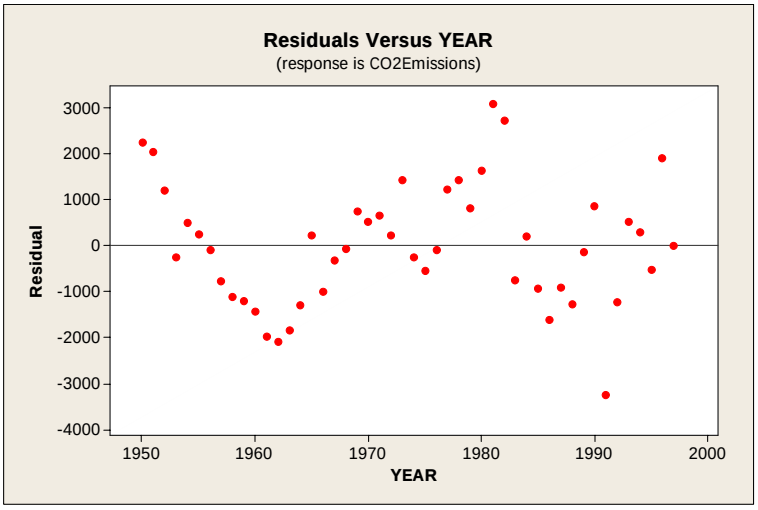
The regression output looks very pleasing.

Regression Analysis: CO2Emissions versus SolidFuels, YEAR

```
The regression equation is
CO2Emissions = - 1950264 + 0.683 SolidFuels + 1002 YEAR


Predictor       Coef  SE Coef       T      P
Constant    -1950264   106602  -18.29  0.000
SolidFuels   0.68302  0.07138    9.57  0.000
YEAR         1002.43    54.91   18.26  0.000

S = 1349.20   R-Sq = 99.6%   R-Sq(adj) = 99.6%

Analysis of Variance

Source          DF           SS           MS        F      P
Regression       2  21181440596  10590720298  5817.98  0.000
Residual Error  45     81915455      1820343
Total           47  21263356051

Source      DF       Seq SS
SolidFuels   1  20574813518
YEAR         1    606627077

Unusual Observations

Obs  SolidFuels  CO2Emissions    Fit  SE Fit  Residual  St Resid
 32       28886         58365  55271     276      3094      2.34R
 33       29676         59536  56813     280      2723      2.06R
 42       40344         69886  73122     337     -3236     -2.48R
 47       48973         85936  84027     601      1909      1.58 X
 48       50875         86336  86329     675         7      0.01 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 0.872753
```

This gives a wonderful R^2, it has significant coefficients on both SolidFuel and YEAR, but the Durbin-Watson statistic is too low.

In addition, the plot of residuals in time order tells us that the method has failed:



Please be aware that our decisions cannot be guided by the R^2 value alone! If the model assumptions are flawed, no value of R^2 will save the analysis.