Time Series

Lesson 2

Grant Foster

TS process which is white noise, but not iid?

Consider this probability function:

$$P_1(x) = \begin{cases} \frac{2}{3} & x = -1 \\ \frac{1}{3} & x = 2 \end{cases}$$

white noise, not iid (cont.)

Consider this probability function:

$$P_1(x) = \begin{cases} \frac{2}{3} & x = -1 \\ \frac{1}{3} & x = 2 \end{cases}$$

Expected value $\langle x \rangle = 0$, expected square $\langle x^2 \rangle = 2$, variance is

$$\langle x^2 \rangle - \langle x \rangle^2 = 2 - 0 = 2.$$

white noise, not iid (cont.)

Now consider a different probability:

$$P_2(x) = \begin{cases} \frac{1}{3} & x = -2 \\ \\ \frac{2}{3} & x = 1 \end{cases}$$

Again, its expectated value is zero, $\langle x \rangle = 0$,
its expected square is $\langle x^2 \rangle = 2$,
so its variance is

$$\langle x^2 \rangle - \langle x \rangle^2 = 2 - 0 = 2.$$

white noise, not iid (cont.)

TS process: random, probability function alternating between P_1 and P_2.

For each value $\langle x \rangle = 0$, and $\langle x^2 \rangle - \langle x \rangle^2 = 2$ – same for each; for different values x_j and x_k $(j \neq k)$, $\langle x_j x_k \rangle = \langle x_j \rangle \langle x_k \rangle = 0$ so $\mathrm{cov}(x_j, x_k) = 0$. Therefore it *is white noise*.

But: even and odd values follow a different distribution, so *not* identically distributed (and therefore not iid).

Weakly stationary but not strongly

Easy answer: same time series as for previous problem.

It's white noise, therefore weakly stationary.

pdf is not time-translation invariant (different between evens and odds), therefore *not* strongly stationary.

Show correlation between random variables x and y cannot be > 1

- Already know that correlation between x and y is equal to correlation between $x - const.$ and $y - const.$ Subtract mean value from each, so x and y have mean value zero.

Show correlation between random variables x and y cannot be > 1

- Already know that correlation between x and y is equal to correlation between $x - const.$ and $y - const.$ Subtract mean value from each, so x and y have mean value zero.

- In that case, covariance (not correlation) is simply

$$\mathrm{cov}(x, y) = \langle xy \rangle = \gamma,$$

(only because we imposed $\langle x \rangle = 0 = \langle y \rangle$). I've simply given the name γ to the covariance.

Correlation ≤ 1 (cont.)

Define a new variable

$$z = y - \frac{\gamma x}{\langle x^2 \rangle},$$

(keep in mind, $\langle x^2 \rangle$ is *not* a random variable, it's just a number, a property of the probability distribution).

Correlation ≤ 1 (cont.)

Define a new variable

$$z = y - \frac{\gamma x}{\langle x^2 \rangle},$$

(keep in mind, $\langle x^2 \rangle$ is *not* a random variable, it's just a number, a property of the probability distribution).

Like x and y, it has mean value zero

$$\langle z \rangle = \langle y \rangle - \frac{\gamma}{\langle x^2 \rangle} \langle x \rangle = 0.$$

Correlation ≤ 1 (cont.)

Note that covariance of x and z is

$$\langle xz \rangle = \langle xy - \frac{\gamma x^2}{\langle x^2 \rangle} \rangle$$

$$= \langle xy \rangle - \frac{\gamma \langle x^2 \rangle}{\langle x^2 \rangle} = \gamma - \gamma = 0,$$

(again, only because $\langle x \rangle = 0 = \langle z \rangle$).

Correlation ≤ 1 (cont.)

We can express y as

$$y = \frac{\gamma x}{\langle x^2 \rangle} + z.$$

Correlation ≤ 1 (cont.)

Variance of y is (using $\langle y \rangle = 0$)

$$\langle y^2 \rangle = \Big\langle \frac{\gamma^2}{\langle x^2 \rangle^2} x^2 + 2\frac{\gamma}{\langle x^2 \rangle} xz + z^2 \Big\rangle$$

$$= \frac{\gamma^2}{\langle x^2 \rangle^2} \langle x^2 \rangle + 2\frac{\gamma}{\langle x^2 \rangle} \langle xz \rangle + \langle z^2 \rangle$$

$$= \frac{\gamma^2}{\langle x^2 \rangle} + 0 + \langle z^2 \rangle = \frac{\gamma^2}{\langle x^2 \rangle} + \langle z^2 \rangle.$$

Correlation ≤ 1 (cont.)

Correlation of x, y is (again using $\langle x \rangle = 0 = \langle y \rangle$)

$$\mathrm{corr}(x, y) = \frac{\langle xy \rangle}{\sqrt{\langle x^2 \rangle \langle y^2 \rangle}}$$

$$= \frac{\gamma}{\sqrt{\langle x^2 \rangle (\gamma^2 / \langle x^2 \rangle + \langle z^2 \rangle)}}$$

$$= \frac{\gamma}{\sqrt{\gamma^2 + \langle x^2 \rangle \langle z^2 \rangle}}$$

Correlation ≤ 1 (cont.)

Note that $\langle x^2 \rangle \geq 0$ and $\langle z^2 \rangle \geq 0$ (they're both squares!), so

$$\sqrt{\gamma^2 + \langle x^2 \rangle \langle z^2 \rangle} \geq |\gamma|.$$

Correlation ≤ 1 (cont.)

Note that $\langle x^2 \geq 0$ and $\langle z^2 \rangle \geq 0$ (they're both squares!), so

$$\sqrt{\gamma^2 + \langle x^2 \rangle \langle z^2 \rangle} \geq |\gamma|.$$

Hence when $\gamma \geq 0$,

$$\mathrm{corr}(x, y) \leq \frac{\gamma}{|\gamma|} \leq 1.$$

Q.E.D.

Time Series Process

Ultimately boils down to a *joint probability function* for x at all moments of time t.

If x continuous, pdf = probability density function.
If x discrete, pmf = probability mass function.

Deterministic: $x(t) = f(t)$ so $p(x) = \delta(x - f(t))$,

$$p(x_1, x_2, ..., x_n) = \prod_{j=1}^{n} \delta(x_j - f(t_j)).$$

Cov, Corr of values from the same TS

Covariance between two different values of a time series is

$$\gamma(j,k) = cov(x_j, x_k) = \langle x_j x_k \rangle - \langle x_j \rangle \langle x_k \rangle.$$

Likewise the correlation between two different values is

$$\rho(j,k) = corr(x_j, x_k)$$

$$= \frac{cov(x_j, x_k)}{\sqrt{cov(x_j, x_j) cov(x_k, x_k)}} = \frac{\gamma(j,k)}{\sqrt{\gamma(j,j)\gamma(k,k)}}.$$

Because these are the covariance and correlation between different values of the *same* time series, we call them *autocovariance* and *autocorrelation*

AutoCoVariance Function (ACVF)

Focus on evenly sampled time series so that the time spacing between observations is everywhere equal.

In that case we can think of the index we attach to a value (the "j" in x_j) as a perfectly good "time index."

AutoCoVariance Function (ACVF)

TS *stationary* $\Rightarrow$ expected value constant over time, i.e., $\langle x_j \rangle = \mu$. Autocovariance obeys

$$\gamma(j,k) = \langle x_j x_k \rangle - \langle x_j \rangle \langle x_k \rangle = \langle x_j x_k \rangle - \mu^2$$

$$= \langle x_{j+s} x_{k+s} \rangle - \mu^2 = \gamma(j+s, k+s),$$

for any index offset s. Let $h = k - j$ be the *lag* between the two values, then

$$\gamma(j, j+h) = \gamma(j+s, j+s+h).$$

AutoCoVariance Function (ACVF)

This means that

$$\gamma(j, j+h) = \gamma(n, n+h),$$

for any two index values j and n.

Hence stationary time series $\Rightarrow$ autocovariance depends only on the *lag* between the two values. Evenly sampled time series $\Rightarrow$ time lag is determined by the index lag. Define the *autocovariance function* (ACVF) as a function of the index lag h

$$\gamma(h) = \gamma(j, j+h) = cov(x_j, x_{j+h}).$$

AutoCoVariance Function (ACVF)

For a TS *not* evenly sampled, define ACVF as

$$\gamma(\tau) = cov(x(t), x(t+\tau)).$$

AutoCoVariance Function (ACVF)

Commutative property of multiplication $\Rightarrow$

$$\gamma(h) = \langle x_j x_{j+h} \rangle - \mu^2 = \langle x_{j+h} x_j \rangle - \mu^2 = \gamma(-h),$$

AutoCoVariance Function (ACVF)

Commutative property of multiplication $\Rightarrow$

$$\gamma(h) = \langle x_j x_{j+h} \rangle - \mu^2 = \langle x_{j+h} x_j \rangle - \mu^2 = \gamma(-h),$$

so for a stationary TS the ACVF is an *even* function, i.e., for negative lag is equal to its value for the same-size positive lag.

AutoCoVariance Function (ACVF)

Commutative property of multiplication $\Rightarrow$

$$\gamma(h) = \langle x_j x_{j+h} \rangle - \mu^2 = \langle x_{j+h} x_j \rangle - \mu^2 = \gamma(-h),$$

so for a stationary TS the ACVF is an *even* function, i.e., for negative lag is equal to its value for the same-size positive lag.

ACVF at lag zero is just the variance of the data series

$$\gamma(0) = \langle x_j^2 \rangle - \langle x_j \rangle^2 = \sigma^2.$$

AutoCorrelation Function (ACF)

Stationary TS $\Rightarrow$ like the ACVF, it depends only on the *lag* between the times of the two time series values. For an evenly sampled time series

$$\rho(h) = corr(x_j, x_{j+h}),$$

for an *un*evenly sampled time series

$$\rho(\tau) = corr(x(t), x(t+\tau)).$$

AutoCorrelation Function (ACF)

Note from the definition of correlation

$$\rho(j,k) = \frac{\gamma(j,k)}{\sqrt{\gamma(j,j)\gamma(k,k)}}.$$

TS *stationary* $\Rightarrow$ even simpler relation

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}.$$

Hence for a stationary time series the ACF, like the ACVF, is an even function, i.e., $\rho(-h) = \rho(h)$.

ACVF, ACF of White Noise

Definition of white noise is a stationary TS for which

$$\langle x_j x_k \rangle = \mu^2 + \sigma^2 \delta(t_k - t_j),$$

where $\delta(h)$ = Dirac δ-function, μ = expected value (*mean*) of the TS, σ^2 = its variance. Using index values

$$\langle x_j x_k \rangle = \mu^2 + \sigma^2 \delta(k - j),$$

where δ = discrete Dirac delta-function.

ACVF, ACF of White Noise

ACVF is nonzero only at lag zero

$$\gamma(h) = \begin{cases} \sigma^2 & h = 0 \\ 0 & h \neq 0 \end{cases}$$

It follows, ACF of white noise has the especially simple form

$$\rho(h) = \begin{cases} 1 & h = 0 \\ 0 & h \neq 0 \end{cases}$$

ACVF, ACF of White Noise

More compactly, $\rho(h) = \delta(h)$.

This simple behavior of the ACF and ACVF gives us a clue whether a time series might be white noise. Suppose we had an *estimate* of the ACF, given by $\hat{\rho}(h)$, at any arbitrary lag h. If the series is white noise, the true ACF is $\rho(h) = \delta(h)$. Therefore the estimated (or *sample*) ACF should be *approximately* equal to the Dirac δ-function.

Yule-Walker Estimate

Given N data points in an evenly sampled time series (with index values ranging from 1 to N), one useful estimate of the sample ACVF is the *Yule-Walker* estimate

$$\hat{\gamma}(h) = \frac{1}{N} \sum_{j=1}^{N-h} (x_j - \bar{x})(x_{j+h} - \bar{x}),$$

where $\bar{x}$ is the sample mean (average)

$$\bar{x} = \frac{1}{N} \sum_{j=1}^{N} x_j.$$

Yule-Walker Estimate

Yule-Walker estimate is a *biased* estimate, i.e., its expected value is *not* the true value!

$$\langle \hat{\gamma}(h) \rangle \neq \gamma(h).$$

Despite this drawback, the expected value of the Yule-Walker sample ACVF is *approximately* equal to the true value.

Yule-Walker estimate of the ACF is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\sum_{j=1}^{N-h}(x_j - \bar{x})(x_{j+h} - \bar{x})}{\sum_{j=1}^{N}(x_j - \bar{x})^2}.$$

Yule-Walker Estimate

For white noise, *variance* of the Y-W estimate is approximately

$$var(\hat{\rho}(h)) \approx \frac{1}{N} \quad \text{(unless } h = 0\text{)}.$$

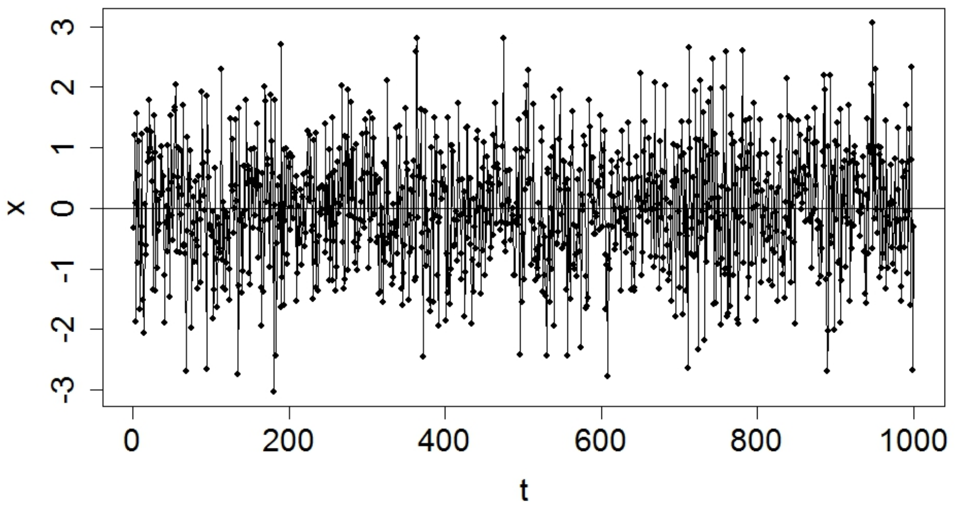
The Y-W sample ACF, like the Y-W sample ACVF, is a biased estimate, but again, the bias is small, the Y-W estimate is good, and $\hat{\rho} \to \rho$ as $N \to \infty$.

“Eyeball” test for white noise

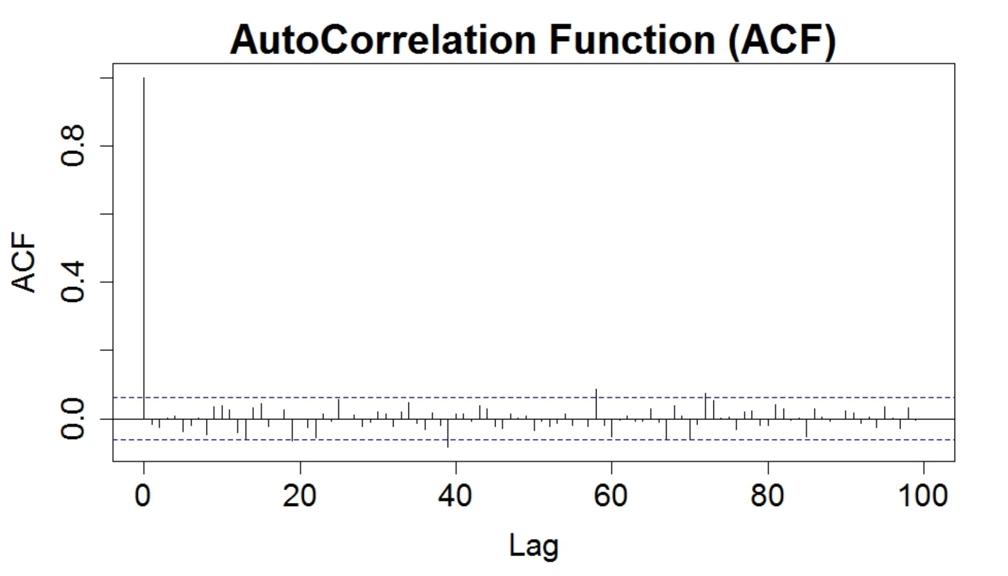
Given a sample, compute the sample ACF $\hat{\rho}(h)$. Lag zero ($h = 0$): true *and* sample ACF are equal to 1. Nonzero lag: sample ACF should equal zero within its error limits, which at 95% confidence is within about two standard deviations of zero. Variance of the sample ACF for white noise is approximately $1/N$, standard deviation approximately $\sqrt{1/N}$, sample ACF should be between $-2\sqrt{1/N}$ and $+2\sqrt{1/N}$ for most (about 95%) of lags which we test.

“Eyeball” test for white noise

Sample ACF for white noise will approximately follow the normal distribution. All of this is only approximate, but for a decent sample size the approximation is usually a good one, and as the sample gets bigger it gets better (in fact the Y-W estimates are *asymptotically* normal).



x
3
2
1
0
-1
-2
-3
0
200
400
600
800
1000
t


AutoCorrelation Function (ACF)
ACF
0.0
0.4
0.8
0
20
40
60
80
100
Lag

AR(1) Noise

A type of noise which is stationary but *not* white noise. Rather than no autocovariance or autocorrelation at nonzero lags, it has autocovariance and autocorrelation at all lags.

AR(1) Noise

A type of noise which is stationary but *not* white noise. Rather than no autocovariance or autocorrelation at nonzero lags, it has autocovariance and autocorrelation at all lags.

It's the first example of a whole class of random processes which we'll study in detail later. Let's take a look at a *first-order autoregressive process*, also known as AR(1) noise.

AR(1) Noise

To generate AR(1) noise:

- Multiply present the noise value by some constant ϕ
- Add a white-noise value to get the *next* AR(1) noise value

Hence AR(1) noise is defined by

$$x_n = \phi x_{n-1} + w_n.$$

Here w_n is white noise, so its expected value never changes and different values of w_n are uncorrelated. Usually insist that w_n is a *zero-mean* white noise process so that $\langle w_n \rangle = 0$.

AR(1) Noise

Apply the definition recursively to note that

$$x_n = \phi x_{n-1} + w_n$$

$$= \phi[\phi x_{n-2} + w_{n-1}] + w_n$$

$$= \phi\Big[\phi[\phi x_{n-3} + w_{n-2}] + w_{n-1}\Big] + w_n = ...$$

AR(1) Noise

Therefore

$$x_n = \phi x_{n-1} + w_n$$

$$= \phi^2 x_{n-2} + \phi w_{n-1} + w_n$$

$$= \phi^3 x_{n-2} + \phi^2 w_{n-2} + \phi w_{n-1} + w_n = ...$$

AR(1) Noise

Therefore

$$\begin{aligned} x_n &= \phi x_{n-1} + w_n \\ &= \phi^2 x_{n-2} + \phi w_{n-1} + w_n \\ &= \phi^3 x_{n-2} + \phi^2 w_{n-2} + \phi w_{n-1} + w_n = ... \end{aligned}$$

We can even recurse the process *infinitely* backward to see that

$$\begin{aligned} x_n &= w_n + \phi w_{n-1} + \phi^2 w_{n-2} + \phi^3 w_{n-3} + ... \\ &= \sum_{j=0}^{\infty} \phi^j w_{n-j}. \end{aligned}$$

AR(1) Noise

Use the fact that $\langle w_n \rangle = 0$ for all w_n to compute the expected value of x_n as

$$\langle x_n \rangle = \sum_{j=1}^{\infty} \phi^j \langle w_{n-j} \rangle = 0.$$

This AR(1) noise process is a zero-mean noise process.

AR(1) Noise

We can also compute the variance of an AR(1) process. We have

$$\sigma^2 = \langle x_n^2 \rangle = \langle \Big(\sum_{j=0}^{\infty} \phi^j w_{n-j} \Big)^2 \rangle$$

$$= \langle \Big(\sum_{j=0}^{\infty} \phi^j w_{n-j} \Big) \Big(\sum_{k=0}^{\infty} \phi^k w_{n-k} \Big) \rangle$$

$$= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \phi^{j+k} \langle w_{n-j} w_{n-k} \rangle .$$

AR(1) Noise

Zero-mean white noise $\Rightarrow \langle w_j w_k \rangle = \sigma_w^2 \delta(j-k)$, so that

$$\langle w_{n-j} w_{n-k} \rangle = \sigma_w^2 \delta(n-j-n+k)$$

$$= \sigma_w^2 \delta(k-j) = \sigma_w^2 \delta(j-k).$$

We end up with

$$\sigma^2 = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \phi^{j+k} \sigma_w^2 \delta(j-k) = \sigma_w^2 \sum_{j=0}^{\infty} \phi^{2j}.$$

AR(1) Noise

If the parameter ϕ satisfies $|\phi| < 1$, then it's a standard algebraic result that

$$\sum_{j=0}^{\infty} \phi^{2j} = \frac{1}{1 - \phi^2},$$

and we have

$$\sigma^2 = \frac{\sigma_w^2}{1 - \phi^2}.$$

AR(1) Noise

If ϕ satisfies $|\phi| \geq 1$, then the sum is infinite, i.e., it *diverges*, in which case the variance σ^2 of our AR(1) process is infinite. This is often undesirable behavior; such an AR(1) process is called *explosive*.

AR(1) Noise

The covariance between different x values for lag $h > 0$ is

$$\gamma(n, n+h) = \langle x_n x_{n+h} \rangle - \langle x_n \rangle \langle x_{n+h} \rangle.$$

AR(1) Noise

The covariance between different x values for lag $h > 0$ is

$$\gamma(n, n+h) = \langle x_n x_{n+h} \rangle - \langle x_n \rangle \langle x_{n+h} \rangle.$$

Since AR(1) noise is zero-mean, this reduces to

$$\gamma(n, n+h) = \langle x_n x_{n+h} \rangle.$$

AR(1) Noise

The covariance between different x values for lag $h > 0$ is

$$\gamma(n, n+h) = \langle x_n x_{n+h} \rangle - \langle x_n \rangle \langle x_{n+h} \rangle.$$

Since AR(1) noise is zero-mean, this reduces to

$$\gamma(n, n+h) = \langle x_n x_{n+h} \rangle.$$

We can use the recursive form to compute this, just as we did for the variance

$$\gamma(n, n+h) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \phi^{j+k} \langle w_{n-j} w_{n+h-k} \rangle.$$

AR(1) Noise

Again apply $\langle w_j w_k \rangle = \sigma_w^2 \delta(j-k)$ to get

$$\gamma(n, n+h) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \phi^{j+k} \sigma_w^2 \delta(k-j-h)$$

$$= \sigma_w^2 \sum_{j=0}^{\infty} \phi^{2j+h} = \sigma_w^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j}.$$

By the same algebraic relation we applied before, this is

$$\gamma(n, n+h) = \phi^h \frac{\sigma_w^2}{1-\phi^2} = \phi^h \sigma^2.$$

AR(1) Noise

This doesn't depend on the particular index value n, so it's time-translation invariant. Since AR(1) noise has constant (zero) mean and time-translation-invariant autocovariance, it's a *stationary* process. We can sum up its first two moments by saying

$$\langle x_n \rangle = 0,$$

and for any lag h (positive, negative, or zero)

$$\gamma(h) = \langle x_n x_{n+h} \rangle = \phi^{|h|} \sigma^2.$$

AR(1) Noise

ACF is therefore

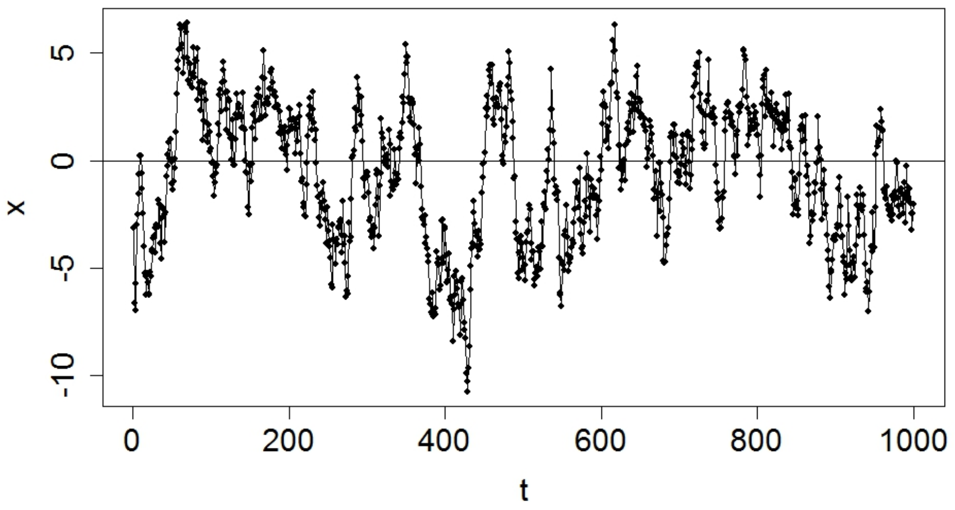
$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^{|h|}.$$

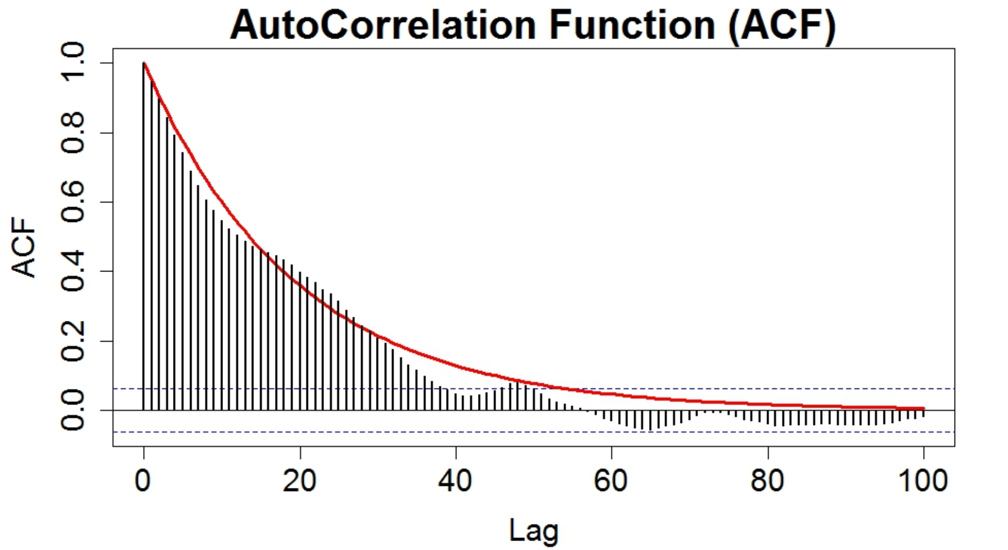
AR(1) Noise

ACF is therefore

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^{|h|}.$$

We've introduced AR(1) noise, long before we consider the class of noise processes of which it's a member, simply to give an example of a purely random process which shows autocorrelation.


5
0
-5
-10
x
0
200
400
600
800
1000
t



AutoCorrelation Function (ACF)
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Lag

Unit Vector

For a stationary time series, expected value is constant over time and autocovariance depends only on the lag, i.e., the difference between the time values of the observations. We can say

$$\langle x_n \rangle = \mu,$$

where μ is the mean value. Holds true for all n values, i.e., there is an identical copy of this equation for every n.

Unit Vector

We can also write this as a *vector* equation by using a Greek letter subscript, which according to the convention introduced previously means that the subscripted quantity is a vector rather than an individual value. But we *cannot* say

$$\langle x_\alpha \rangle = \mu,$$

because such an equation is nonsense. The left-hand side is a vector (because it's the expected value of a vector), but the right-hand side is *not* a vector, it's a scalar.

Unit Vector

Therefore let's introduce a remarkably useful quantity, the *unit vector* $\mathbf{1}_\alpha$. It's a vector for which all the individual components are equal to 1. Hence we can say that

$$\mathbf{1}_n = 1,$$

where we've used a Latin index to indicate that this equation refers to the individual values, so we have a copy of this equation for each possible index value n.

Unit Vector

With the unit vector in hand, we can express the constancy of the *vector* of expected values of the TS by saying

$$\langle x_\alpha \rangle = \mu \mathbf{1}_\alpha.$$

This single equation is an equality between two vectors rather than a *set* of equations expressing equality between scalars. The distinction may not seem important or useful at this time, but its value will become clear later.

Variance-Covariance Matrix

We can express the time-translation invariance of the covariances between different time series values by saying

$$cov(x_j, x_k) = \gamma(|k - j|),$$

which expresses the fact that the covariance depends only on the *difference* between the time indexes, i.e., the *lag* between the values. For a time series which is not evenly sampled we would say

$$cov(x(t_j), x(t_k)) = \gamma(|t_j - t_k|).$$

Variance-Covariance Matrix

Whether a TS is stationary or not, we can arrange the covariances of the values into the *variance-covariance matrix*

$$V_{jk} = cov(x_j, x_k).$$

Note the *diagonal* elements are the variances of the individual TS values while the off-diagonal elements are the covariances between different values, hence the name "variance-covariance matrix." Many authors use the symbol Γ to represent the variance-covariance matrix, but I prefer the symbol V, reserving Γ for other uses.

Variance-Covariance Matrix

The previous expression is a whole *set* of equations, one for each pair of TS values x_j and x_k. We can write it as a single *matrix* equation

$$V_{\alpha\beta} = \langle x_\alpha x_\beta \rangle - \langle x_\alpha \rangle \langle x_\beta \rangle.$$

The variance-covariance matrix is fundamental in time series analysis. Its importance can hardly be overstated.

Variance-Covariance Matrix

Because multiplication is commutative, the variance-covariance matrix is *symmetric*, i.e.,

$$V_{jk} = V_{kj}.$$

We can write this as a genuine matrix equation by saying

$$V_{\alpha\beta} = V_{\beta\alpha},$$

which is *not* an equation about individual values; the quantity $V_{\beta\alpha}$ is not a particular entry of the matrix, it's a matrix which is the *transpose* of the matrix $V_{\alpha\beta}$.

Variance-Covariance Matrix

When the TS is evenly sampled and stationary we have

$$V_{jk} = \gamma(|j - k|),$$

so its value depends only on the difference between the index values. This means that the values are unchanged when one moves up-and-to-the-left or down-and-to-the-right around the matrix. Any matrix which has this property is called a *Toeplitz* matrix.

Variance-Covariance Matrix

Therefore, for an evenly sampled TS the property of stationarity is equivalent to the two requirements that the mean is time-independent, and that the variance-covariance matrix is a symmetric Toeplitz matrix.

Variance-Covariance Matrix of White Noise

For white noise, the variance-covariance matrix takes the especially simple form

$$V_{jk} = \sigma^2 \delta_{jk},$$

where δ_{jk} is the *Kronecker delta*. It's like the Dirac δ-function except that for the Kronecker delta, being nonzero only for a limited set of values applies to the indices, i.e.

$$\delta_{jk} = \delta(j-k) = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}$$

Variance-Covariance Matrix of White Noise

We can turn the *set* of equations into a single matrix equation just by writing

$$V_{\alpha\beta} = \sigma^2 \delta_{\alpha\beta},$$

for white noise. Hence the symbol δ_{jk} denotes a set of values which are zero except when $j = k$, while the symbol $\delta_{\alpha\beta}$ denotes a *matrix* (which happens to be the *identity matrix*).

Variance-Covariance Matrix, Stationary TS

For a stationary TS, the mean μ is constant, the variance-covariance matrix is a symmetric Toeplitz matrix, so

$$\langle x_j x_k \rangle = \mu^2 + V_{jk},$$

or in matrix form

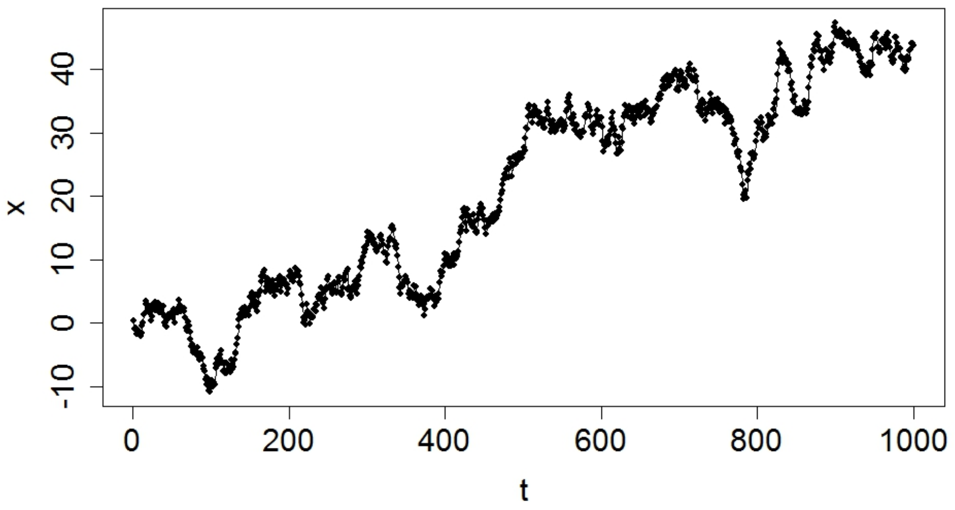
$$\langle x_\alpha x_\beta \rangle = \mu^2 \mathbf{1}_\alpha \mathbf{1}_\beta + V_{\alpha\beta}.$$

This serves to emphasize just how fundamental is the variance-covariance matrix.

Making Data Stationary

Many methods and models which can be applied to a stationary TS. But if a time series is not stationary, those methods don't apply. We can, however, sometimes find a convenient way to make it so.

The data in the following TS is not stationary, but is still the result of a purely random noise process.



x
-10
0
10
20
30
40
0
200
400
600
800
1000
t

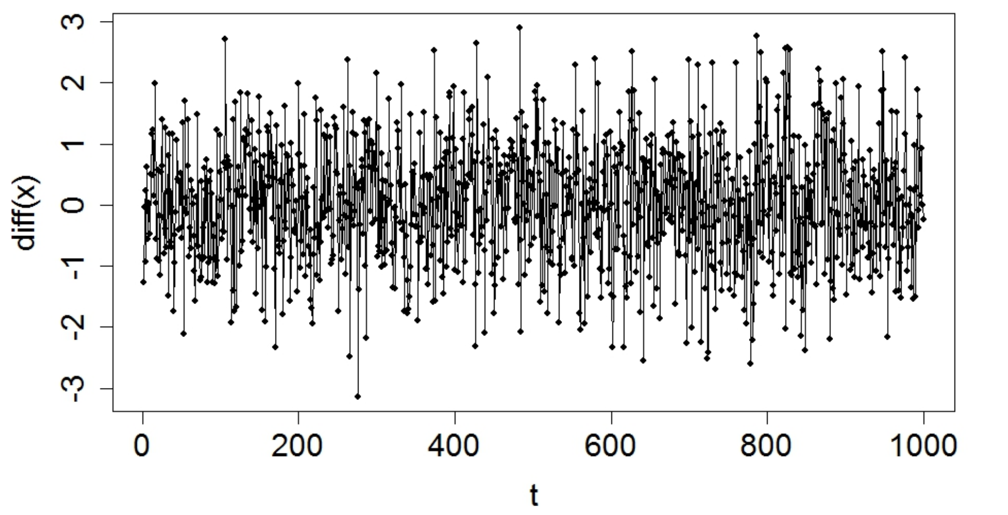
Making Data Stationary

In some cases (and even when there is a signal present) we can eliminate the drift (the changing mean value over time) by computing the *first-differenced* time series. The first differences are defined as

$$\Delta x_j = x_j - x_{j-1}.$$

Making Data Stationary

We can think of Δ as an *operator*, the *first-difference operator*, which transforms a TS into its first differences. Note that the first difference for the initial data value is undefined because we don't know the value of its predecessor, so the first-difference TS has one data point fewer than the series from which it's derived. The first differences of this set of data are shown in the following figure.



diff(x)
t
3
2
1
0
-1
-2
-3
0
200
400
600
800
1000

Making Data Stationary

Be aware that if a time series is the sum of signal and noise, first differencing will alter the signal as well as the noise, and in some cases will eliminate it. Suppose a TS is the sum of a perfectly linear trend and stationary noise

$$x_j = \beta_o + \beta_1 t_j + \varepsilon_j.$$

We can directly compute the first-difference values as

$$\Delta x_j = x_j - x_{j-1} = \beta_1 (t_j - t_{j-1}) + \varepsilon_j - \varepsilon_{j-1}.$$

Making Data Stationary

When the data are evenly spaced with spacing τ so $t_j = j\tau$

$$\Delta x_j = \beta_1 \tau + \varepsilon_j - \varepsilon_{j-1}.$$

This happens to be a stationary noise process. It's not zero-mean noise because its mean value is $\beta_1 \tau$ (unless the slope β_1 is equal to zero). It's not white noise because it shows autocorrelation at lags other than zero (its lag-1 autocorrelation is $-\frac{1}{2}$). But it is a pure noise process, there's no signal to extract. The signal has been eliminated by the first-difference operator.

Making Data Stationary

Suppose the trend is a quadratic function of time

$$x_j = \beta_o + \beta_1 t_j + \beta_2 t_j^2 + \varepsilon_j.$$

Applying the first-difference operator gives (assuming even sampling with time spacing τ)

$$\Delta x_j = \beta_1 \tau + 2\beta_2 \tau t_j - \beta_2 \tau^2 + \varepsilon_j - \varepsilon_{j-1}.$$

Not yet stationary because a linear time trend is still present.

Making Data Stationary

Applying the first-difference operator *again* gives

$$\Delta^2 x_j = 2\beta_2\tau^2 + \varepsilon_j - 2\varepsilon_{j-1} + \varepsilon_{j-2}.$$

We now have a stationary time series to work with. The twice-applied difference operator Δ^2 is called the *second-difference operator*.

In general, a signal which is a polynomial of degree p will be eliminated by applying the difference operator p times.

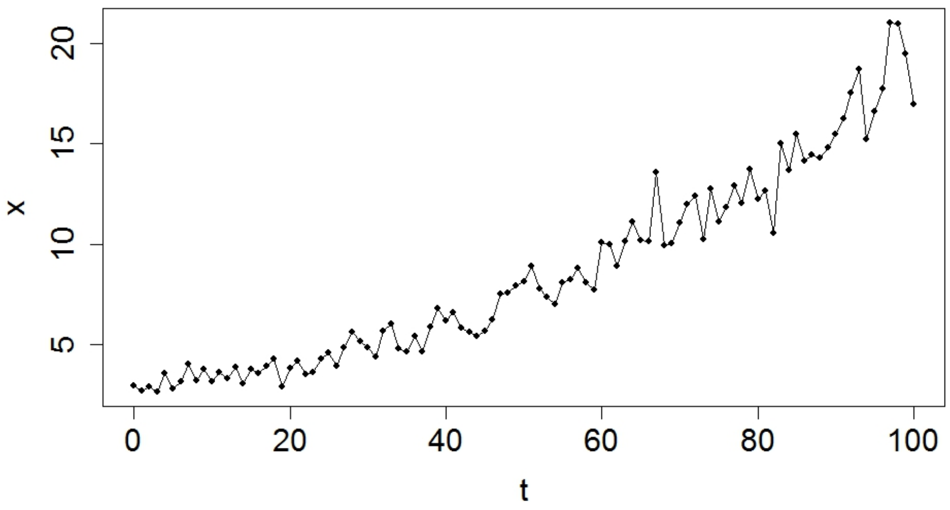
Making Data Stationary

In some fields, notably economics, it is customary to remove trends by applying the difference operator enough times to make the data stationary. But in the physical sciences, it is usually the signal which we're most interested in studying. Removing it by repeated differencing eliminates exactly what we want to study. One of our focal points is not to rely on the viewpoint that differencing is always the way to deal with non-stationary time series. Sometimes it is! But in the physical sciences it is sometimes counterproductive.

Heteroskedasticity

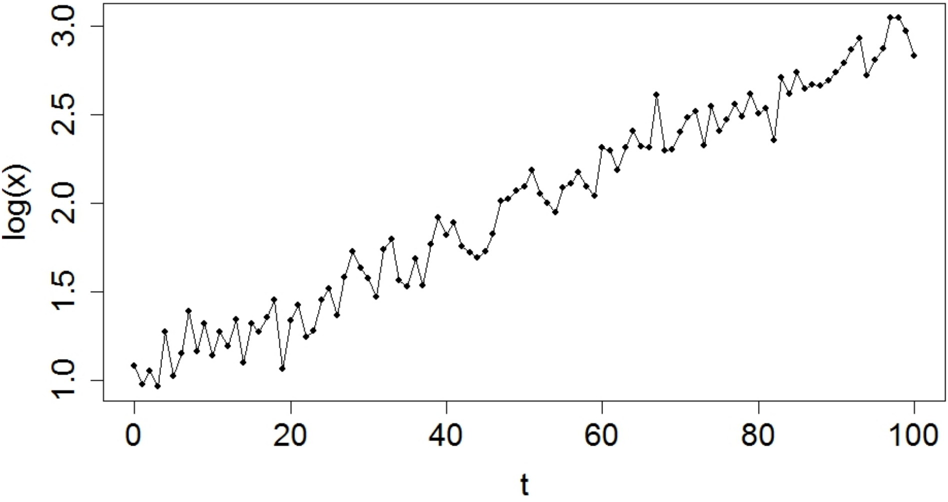
Time series can be non-stationary for reasons other than trend. The following figure shows data which exhibit a trend, but also show another kind of non-stationarity, the fact that the variance of the data shows notable changes. Such behavior is called *heteroskedasticity*.

It often happens when the variance of the data is larger for larger data values, as is the case in this example.


x
t
0
20
40
60
80
100
5
10
15
20

Heteroskedasticity

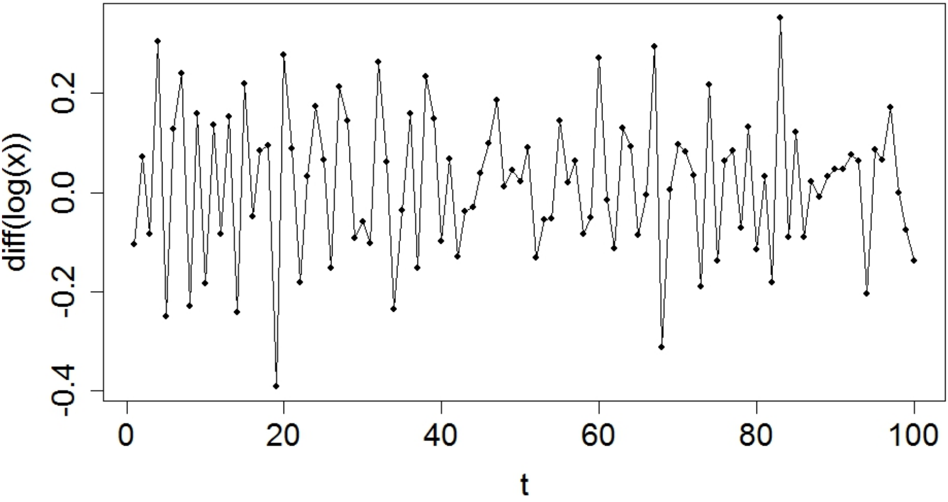
When that happens, we can sometimes eliminate heteroskedasticity by transforming the data. One common approach is to log-transform the data. When applied to these data, it gives the following (the varying degree of data variance has been eliminated).



log(x)
1.0
1.5
2.0
2.5
3.0
0
20
40
60
80
100
t

Heteroskedasticity

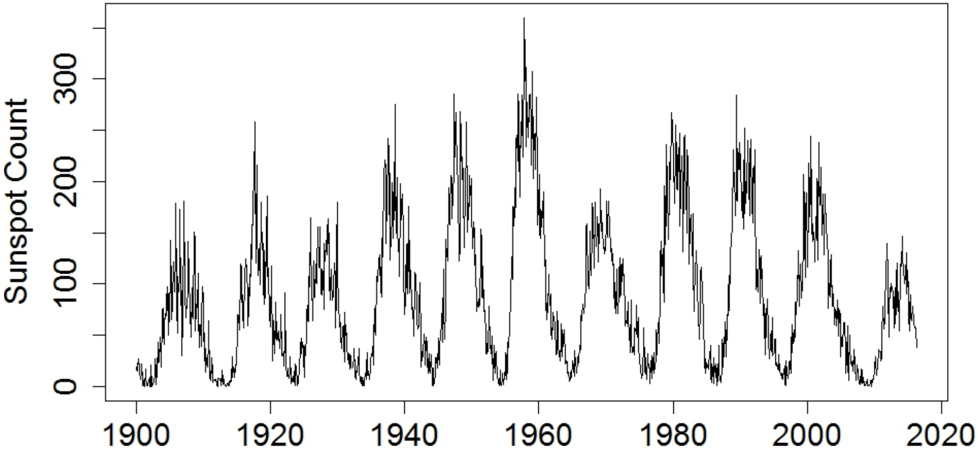
Of course the trend still remains, but that can be eliminated by first-differencing as in the following. The first-differenced log-transformed data are in fact a stationary time series.



diff(log(x))
-0.4
-0.2
0.0
0.2
0
20
40
60
80
100
t

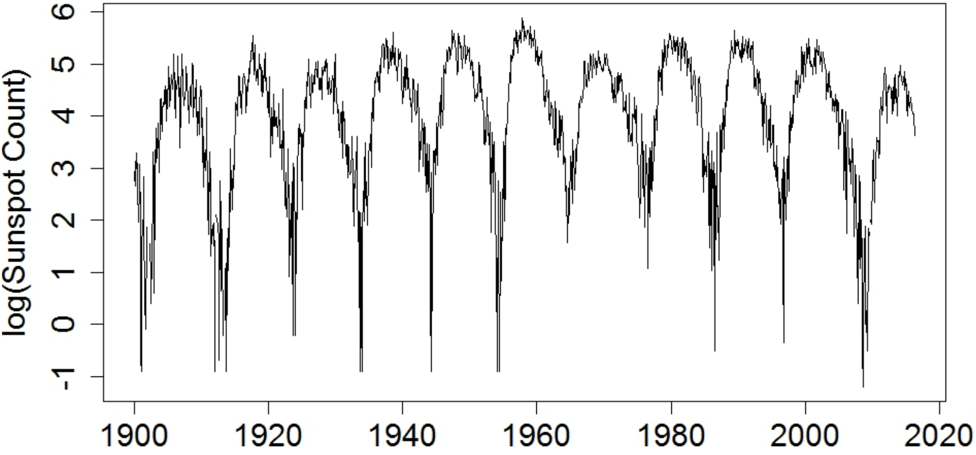
Heteroskedasticity

Sometimes log-transforming doesn't eliminate heteroskedasticity, it only changes it. An example is monthly mean sunspot numbers.


Sunspot Count
0
100
200
300
1900
1920
1940
1960
1980
2000
2020

Heteroskedasticity

There is greater variance when sunspot numbers are large than when they are small. But log-transforming the data doesn't solve the problem, only reverses it so that there is greater variance when sunspot numbers are small than when they are large.


log(Sunspot Count)
6
5
4
3
2
1
0
-1
1900
1920
1940
1960
1980
2000
2020

Heteroskedasticity

Another common strategy is to subject the data to a *power transform*. This is defined for some well-chosen exponent λ as

$$y_j = \frac{x_j^\lambda - 1}{\lambda g^{\lambda-1}},$$

where g is the geometric mean of the x values

$$g = \left(\prod_{j=1}^{N} x_j \right)^{1/N}$$

Heteroskedasticity

The factor $g^{\lambda-1}$ is included in the denominator so that the units of measurement will remain unchanged.

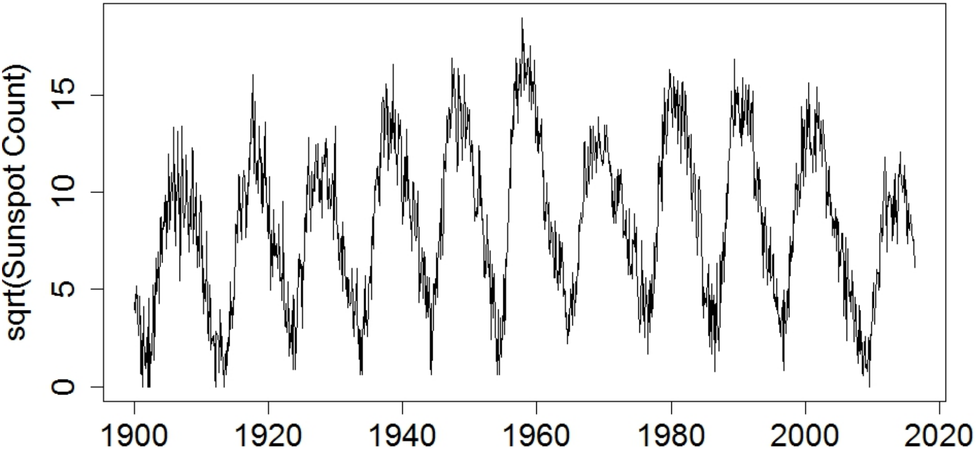
Sometimes the factor g is ignored, which gives the very similar *Box-Cox transform*

$$y_j = \frac{x^\lambda - 1}{\lambda}.$$

With this definition, the log-transform of the data is the limit as $\lambda \to 0$.

Heteroskedasticity

The following shows the result of applying a Box-Cox transform to sunspot counts with exponent $\lambda = \frac{1}{2}$.


sqrt(Sunspot Count)
0
5
10
15
1900
1920
1940
1960
1980
2000
2020