Time Series Analysis with Applications in R

J. Hamski

2/11/2017

Notes from: Time Series Analysis with Applications in R

2nd Addition, Cryer and Chan, 2008

These notes are mixture of code from the book, my own code, summaries of the text and math shown, and links to other useful resources. The goal is to get a comprehensive understanding of time series analysis.

Ch.1 Introduction

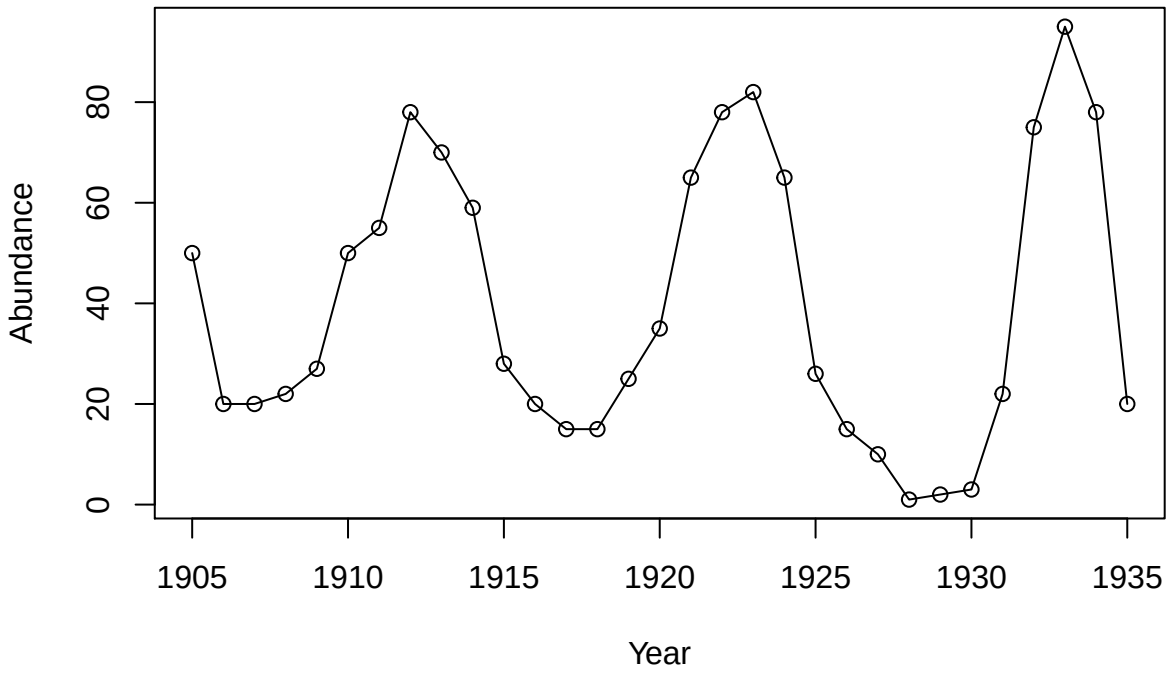
Generally time series analysis has two goals: (1) understand or model the stochastic mechanism that gives rise to an observed series, and (2) predict or forcast the future values of a series based on the history of that series and perhaps other related series or factors.

A somewhat unique feature of time series: we usually cannot assume that the observations arise independently froma common population. Therefore, models that incoporate dependence are a key concept in time series analysis.

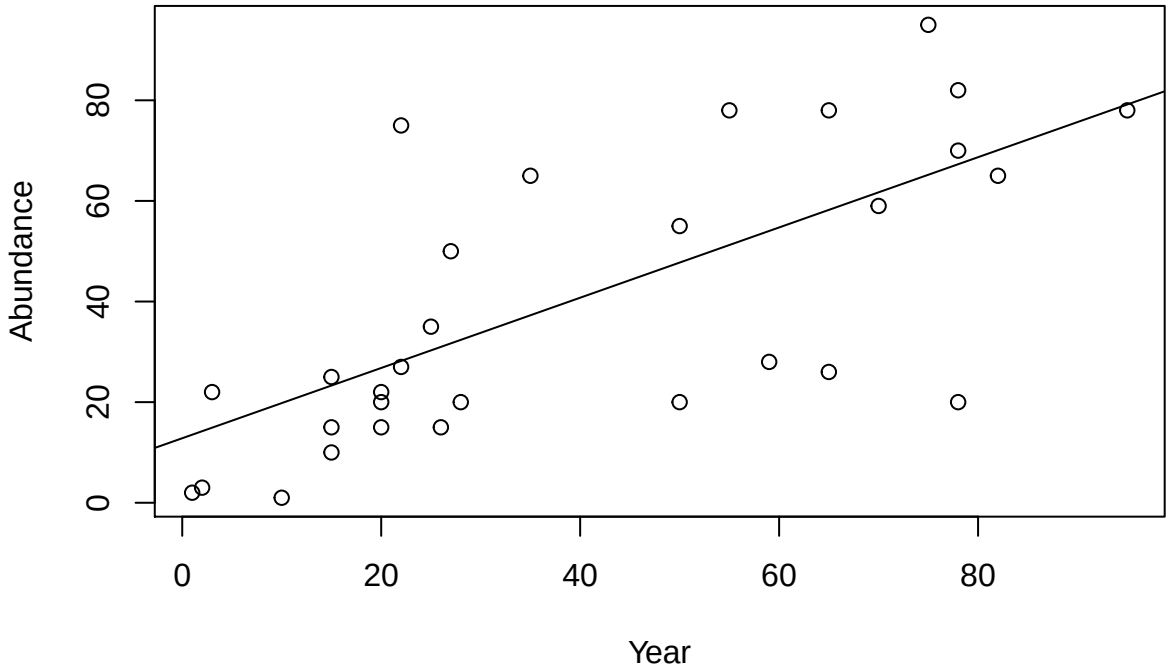
Common technique: plot previous period vs. current period for entire series to see if there is dependency from one year to the next. The rainfall example shows no trend, indicating independence in rainfall from one year to the next. But look at Hare populations:

```
library(ggplot2)
library(magrittr)
library(TSA)

data(hare)
plot(hare, ylab = "Abundance", xlab = "Year", type = "o")
```




```
plot(y = hare, x = zlag(hare), ylab = "Abundance", xlab = "Year") + abline(lm(zlag(hare) ~ hare))
```

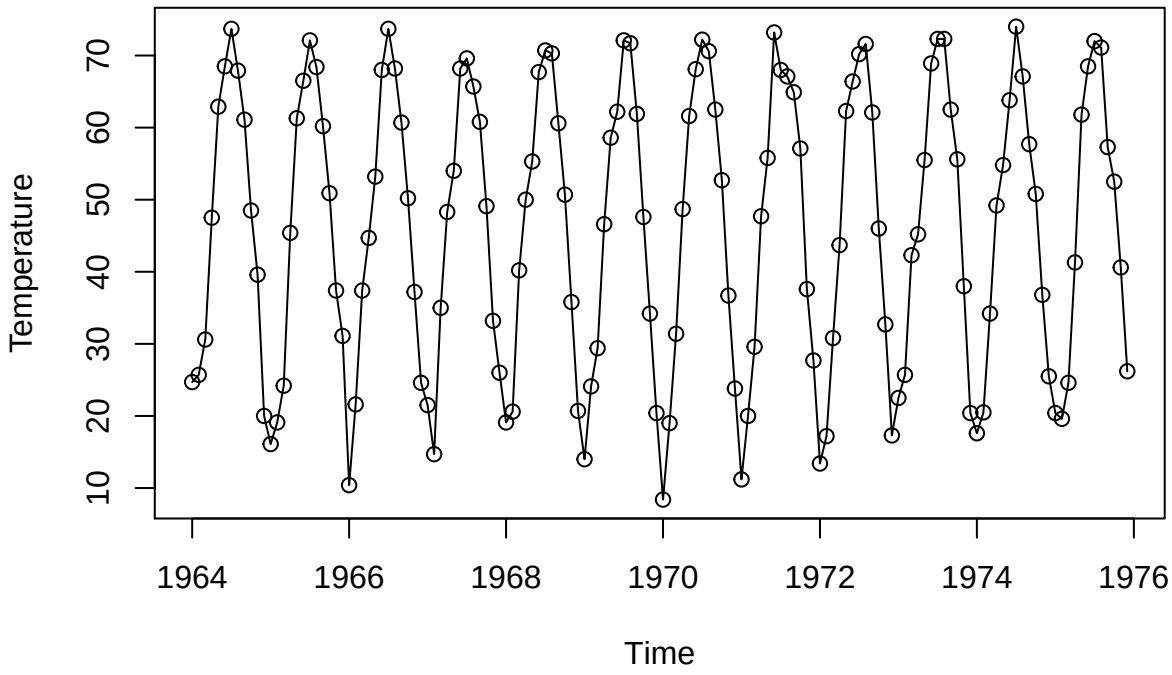



```
## numeric(0)
```

Neighboring values are very closely related. Low values tend to be followed by low values the next year, and high followed by high.

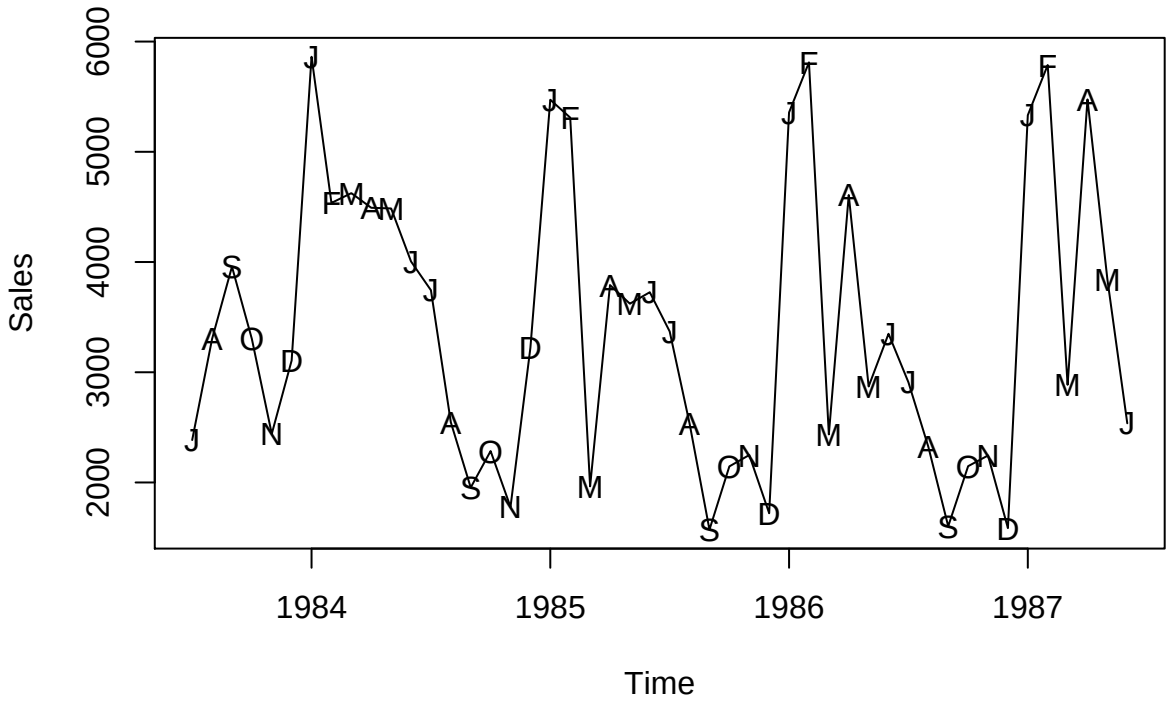
Another common time series analysis is looking for seasonality. Modeling a seasonal time series must account for the variation among same month temperature values (January to January), but still maintain the seasonality (January to July).

```
data(tempdub)
plot(tempdub, ylab = "Temperature", type = "o")
```



Sometimes it helps to identify seasonality by labeling the timeseries graphic.

```
data(oilfilters)
plot(oilfilters, type = "l", ylab = "Sales")
points(y = oilfilters, x = time(oilfilters), pch = as.vector(season(oilfilters)))
```



This makes it apparent that January and Feburary are peaks, while Autumn is a trough.

A Model-Building Strategy Box and Jenkins (1976) is the seminal time series analysis citation. They outline a three step process: (1) model specification (i.e. picking what might be an appropriate class of models for the problem) (2) model fitting (ie least squares, max likelihood) (3) model diagnosis. Attempt to ahere to

the *principle of parsimony* wherein the model should require the smallest number of parameters that will adequately represent the time series.

Ch. 2 Fundamental Concepts

Time Series and Stochastic Processes

From Wikipedia: *Autocorrelation, also known as serial correlation, is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them. The analysis of autocorrelation is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies. It is often used in signal processing for analyzing functions or series of values, such as time domain signals.*

Covariance and correlation are measures of the linear dependence between random variables but the unitless correlation is somewhat easier to interpret. Note there are several important relationships between correlation, covariance, variance, etc.. and we're primarily lookint at auto-covariance/correlation, meaning the relationships of a series within is self varying by time, however the statistical properties are the same as if it was two different variables. **Eq. 2.2.5**

$$\begin{aligned} \gamma_{t,t} = Var(Y_t) \quad & \rho_{t,t} = 1 \\ \gamma_{t,s} = \gamma_{s,t} \quad & \rho_{t,s} = \rho_{s,t} \\ |\gamma_{t,s}| \leq \sqrt{\gamma_{t,t}\gamma_{s,s},} \quad & |rho_{t,s} \leq 1| \end{aligned}$$

Equation 2.2.6 To Investigate covariance properties of time series models, the following equation is used: If $c_1, c_2, ..., c_m$ and $d_1, d_2, ..., d_n$ are constants and $t_1, t_2, ..., t_m$ and $s_1, 2_2, ..., s_n$ are time points, then:

$$Cov[\sum_{i=1}^{m} c_i Y_{t_i}, \sum_{j=1}^{n} d_j Y_{s_i}] = \sum_{i=1}^{m}\sum_{j=1}^{n} c_i d_j Cov(Y_{t_i}, Y_{s_j})$$

In English - the covariance of the two series is equal to the sum of their multiple times the covariance of the series.
NOTE: there's also a "special case" equation 2.2.7 for the variance of the sum of $c_i Y_{t_i}$, come back to it if important.
NOTE: need to confirm this in English translation.

The Random Walk

Let $e_1, e_2, ...$ be an i.i.d. sequence, each RV with $\mu = 0$ and variance σ_e^2. The observed time series $\{Y_t : t = 1, 2, ...\}$ is constructed:

$$Y_1 = e_1$$

$$Y_t = e_1 + e_2 + ... + e_t$$

also expressed as $Y_t = Y_{t-1} + e_t$ with intial condition $Y_1 = e_1$ with $e's$ being the "size of steps" taken along a number line, forward (positive) or backward (negative).

$$\mu_t = E(Y_t) = E(e_1 + e_2 + ... + e_t) = 0 + 0 + 0 = 0$$

Which is pretty obvious, the series is going to dance around 0 randomly. For variance:

$$Var(Y_t) = Var(e_1 + e_2 + ... + e_t) = \sigma_e^2 + \sigma_e^2 + ... + \sigma_e^2$$

Therefore, $Var(Y_t) = t\sigma_e^2$ - in English, *process variance increases linearly with time.*

This is called a random walk for a reason - we can expect zero covariance using equation 2.2.6 unless we're comparing the covariance $i = j$, where the covariance function is just equal to the variance of the series.

$$\gamma_{t,s} = Cov(Y_t, Y_s) = Cov(e_1 + e_2 + ... + e_t, e_1 + 2 + e_2 + ... + e_t + e_{t+1} + ... + e_s)$$

The autocovariance function for all time points t and s is: $\gamma_{t,s} = t\sigma_e^2$

The autocorrelation function for a random walk is:

$$\rho_{t,s} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} = \sqrt{\frac{t}{s}}$$

Let's check out some values for ρ to see how the random walk behaves:

```
find.rho <- function(t,s){return(sqrt(t/s))}
find.rho(1,2) %>% print()
```

```
## [1] 0.7071068
```

```
find.rho(8,9) %>% print()
```

```
## [1] 0.942809
```

```
find.rho(24,25) %>% print()
```

```
## [1] 0.9797959
```

```
find.rho(1,25) %>% print()
```

```
## [1] 0.2
```

Conclusion: values of Y are more and more strongly correlated at t increases (time goes by), but values of Y are less and less correlated the farther apart they are (see $1{,}25 = 0.2$).

Note: in the book they use a dataset from TSA to show a random walk, but I think it is a lot more instructive to generate one using base R.

A normal distribution is used here to obtain e, but that isn't required - its iid that's important.

```
e.series <- rnorm(60, mean = 0, sd = 1)

Y <- 0 #initial condition

for (i in 1:length(e.series)){
  Y <- c(Y, Y[i-1] + e.series[i])
}

plot(Y, type = "o", ylab = "Random Walk")
```

A Moving Average

Suppose:

$$Y_t = \frac{e_t + e_{t-1}}{2}$$

Again assume e is iid with mean of zero and variance is sigma squared.

$$\mu_t = E(Y_t) = \frac{E(e_t) + E(e_{t-1})}{2} = 0$$

$$Var(Y_t) = Var[\frac{e_t + e_{t-1}}{2}] = \frac{Var(e_t) + Var(e_{t-1})}{4}$$

$$Cov(Y_t, Y_{t-1}) = Cov[\frac{e_t + e_{t-1}}{2}, \frac{e_{t-1} + e_{t-2}}{2}] = \frac{Cov(e_{t-1}, e_{t-2})}{4}$$

as all other covariances are zero, giving:

$$\gamma_{t,t-1} = 0.25\sigma_e^2$$

for all t.

Covariance

$$\gamma_{t,s} = \begin{cases} 0.5\sigma_e^2 & \text{for} \quad |t-s| = 0 \\ 0.25\sigma_e^2 & \text{for} \quad |t-s| = 1 \\ 0 & \text{for} \quad |t-s| > 1 \end{cases}$$

Autocorrelation

$$\rho_{t,s} = \begin{cases} 1 & \text{for} \quad |t-s| = 0 \\ 0.5 & \text{for} \quad |t-s| = 1 \\ 0 & \text{for} \quad |t-s| > 1 \end{cases}$$

Notice that $\rho_{2,1} = \rho_{3,2} = \rho_{9,8} = 0.5$ - values of Y precisely one time unit apart have the same correlatio no matter where they occur in time.

Stationarity

Stationarity is an assumption that the probability laws that govern the behavior of the process observed by the time series do not change over time. "Statistical equilibrium". Even if you chose an arbitrary lag time k, the joint distribution of process $\{Y_t\}$ doesn't change. The mean and variance of Y_t and Y_{t-k} are constant over time.

See book for following proof: $\gamma_{t,s} = \gamma_{0,|t-s|}$ meaning the covariance between Y_t and Y_s is depends on time only through the time difference $|t-s|$, no on the actual placement of times t and s!
Thus, for a stationary process:

$$\gamma_k = Cov(Y_t, Y_{t-k})$$
$$\rho_k = Corr(Y_t, Y_{t-k})$$
$$\rho_k = \frac{\gamma_k}{\gamma_0}$$

When stationary, general properties in eq. 2.2.5 become:

$$\begin{array}{ll} \gamma_0 = Var(Y_t) & \rho_{t,t} = 1 \\ \gamma_k = \gamma_{-k} & \rho_k = \rho_{-k} \\ |\gamma_k| \leq \gamma_0 & |\rho_k| \leq 1 \end{array} \}$$

If a process is strictly stationary and has a fininte variance, then the covariance function must depend only on the time lag.

Weak / Second-order Stationary
1. The mean function is constant over time
2. $\gamma_{t,t-k} = \gamma_{0,k}$ for all time t and lag k

In practice and in book examples, this is the form of stationarity used.

White Noise

White noise process is a sequence of i.i.d. random variables $\{e_t\}$, which is strictly stationary.

Note this is an iteration on the moving average example. It is useful for constructing processes and I think it will show up again later on.

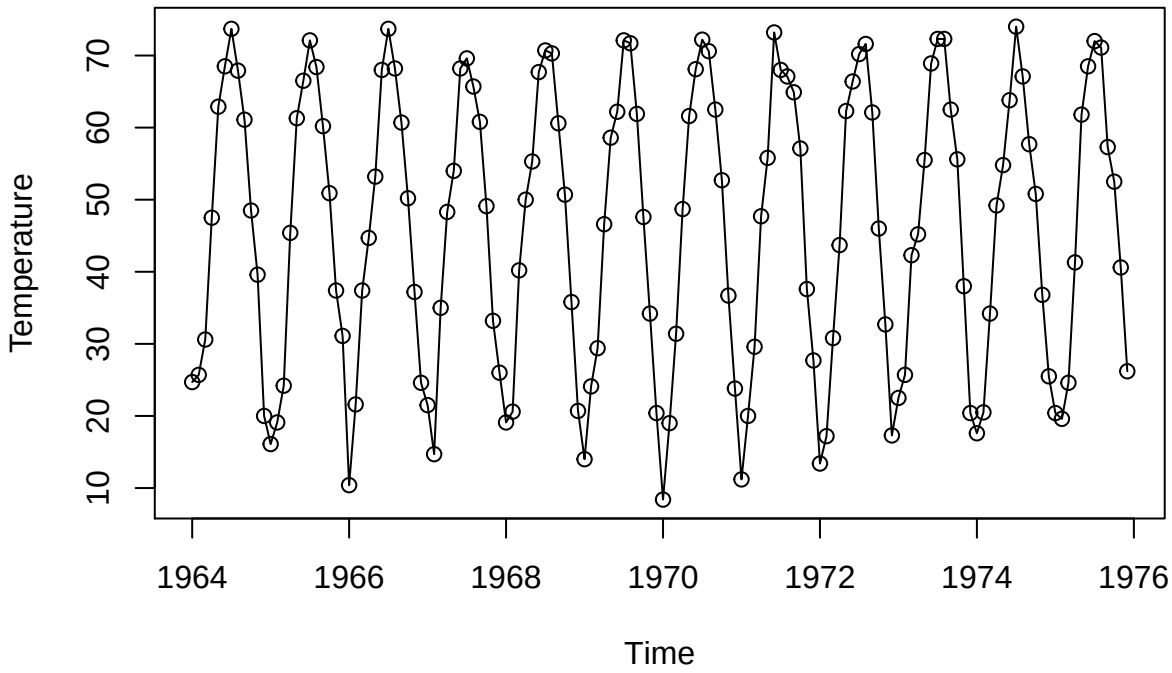
Ch. 3 Trends

General time series: mean function is a totally arbitrary funciton of time
Stationary time series: mean function is constant with time
This chapter considers simple, but not constant, mean functions of time.

Sometimes you see a trend in a random walk - however that percieved trend is just due to strong positive correlation between the series values at nearby time points and the increasing variance in the process as time goes by. These are often called **stochastic trends**.

However, also think about a trend like average monthly temperature:

```
data(tempdub)
plot(tempdub, ylab = "Temperature", type = "o")
```



The Earth's inclination towards the sun causes this cycle. We might model this as $Y_t = \mu_t + X_t$ where μ_t is a deterministic function that has a period of 12, hence

$$\mu_t = \mu_{t-12} \; for \; all \; t$$

. This model could be described as having a *deterministic trend*.

Estimation of a Constant Mean

For review: Mean and Variance of Sample Mean

Consider a constant mean function (trend): $Y_t = \mu + X_t$ where $E(X_t) = 0$ for all t. The most common estimate of μ is the sample mean. Suppose $\{Y_t\}$ is a stationary time series with autocorrelation function ρ_k. Then:

$$Var(\bar{Y}) = \frac{\gamma_0}{n}[1 + 2\sum_{k=1}^{n-1}(1 - \frac{k}{n})\rho_k$$

$\frac{\gamma_0}{n}$ is the process (population) variance divided b the sample size. If the series is just white noise, the $\rho_k = 0$ and the variance of the mean $Var(\bar{Y}) = \frac{\gamma_0}{n}$.

However, if $\rho_k \geq 0$ (i.e. the autocorrelation for a given lag time isn't 0) we see that $Var(\bar{Y}) > \frac{\gamma_0}{n}$. (the variance of the sampling mean will be greater than the variance of the population variance) Poistive correlations make the mean *more* difficult to estimate than in the white noise case. For many stationary processes, the autocorrelation function decays quickly enough with increasing lags that:

$$\sum_{k=0}^{\infty} |\rho_k| < \infty$$

The random cosine wave is an exception.

Under this assumption of a quickly decreasing autocorrelation function decay, the following is a useful approximation:

$$Var(\bar{Y}) \approx \frac{\gamma_0}{n}[\sum_{k=-\infty}^{\infty} \rho_k] \; for \; large \; n$$

The variance is inversely proportional to the sample size n.

```
#using my own random walk simulation to show that the variance of the estimate of the mean increases wi

var(Y[1:10])
```

```
## [1] 4.004905
```

```
var(Y[1:30])
```

```
## [1] 14.77816
```

```
var(Y[1:60])
```

```
## [1] 9.798834
```

Regression Methods