Linear models in R

A linear model specifies a linear relationship between a dependent variable and n independent variables

We take height to be a variable that describes the heights (in cm) of ten people.
height <- c(176, 154, 138, 196, 132, 176, 181, 169, 150, 175)
Now let's take bodymass to be a variable that describes the masses (in kg) of the same ten people.
bodymass <- c(82, 49, 53, 112, 47, 69, 77, 71, 62, 78)
We can now create a simple plot of the two variables as follows:
plot(bodymass, height)
We can enhance this plot using various arguments within the plot() command.

plot(bodymass, height, pch = 16, cex = 1.3, col = "blue", main = "HEIGHT PLOTTED AGAINST BODY MASS", xlab = "BODY MASS (kg)", ylab = "HEIGHT (cm)")

In the above code, the syntax pch = 16 creates solid dots, while cex = 1.3 creates dots that are 1.3 times bigger than the default (where cex = 1)

Now let's perform a linear regression using lm() on the two variables by adding the following text at the command line:
lm(height ~ bodymass)

Time Series Statistical Models

White Noise

White noise is an important concept in time series forecasting.
If a time series is **white noise**, it is a sequence of random numbers and cannot be predicted. If the series of forecast errors are not white noise, it suggests improvements could be made to the predictive model.

What is a White noise time series?
A time series is white noise if the variables are independent and identically distributed with a mean of zero.
This means that all variables have the same variance (sigma^2) and each value has a zero correlation with all other values in the series.
If the variables in the series are drawn from a Gaussian(normal) distribution, the series is called **Gaussian white noise.**

Why does it matter?

It is important for two main reasons:

1. **Predictability**: If your time series is white noise, then, by definition, it is random. You cannot reasonably model it and make predictions.
2. **Model Diagnostics**: The series of errors from a time series forecast model should ideally be white noise.

Is your time series white noise?

Your time series is probably NOT white noise if one or more of the following conditions are true:

- Is the mean/level non-zero?
- Does the mean/level change over time?
- Does the variance change over time?
- Do values correlate with lag values?

Some tools that you can use to check if your time series is white noise are:

- **Create a line plot.** Check for gross features like a changing mean, variance, or obvious relationship between lagged variables.
- **Calculate summary statistics.** Check the mean and variance of the whole series against the mean and variance of meaningful contiguous blocks of values in the series (e.g. days, months, or years).
- **Create an autocorrelation plot.** Check for gross correlation between lagged variables.

Summary

- White noise time series is defined by a zero mean, constant variance, and zero correlation.
- If your time series is white noise, it cannot be predicted, and if your forecast residuals are not white noise, you may be able to improve your model.

Two ways of introducing serial correlation and more smoothness into time series model

- Moving Averages
- Autoregressions

Moving Averages

We might replace the white noise series ω_t by a moving average that smooths the series.For example, consider replacing ω_t in

$\omega_t \sim \omega_n(0, \sigma_{\omega}^{2})$ by an average of its current value and its immediate neighbors in the past and future. That is, let

$$v_t = \frac{1}{3}(\omega_{t-1} + \omega_t + \omega_{t+1})$$

INSTRUCTIONS

Gaussian white noise series and three-point moving average of the Gaussian white noise series.

```
w = rnorm(500,0,1)
# 500 N(0,1) variates
v = filter(w, sides=2, rep(1/3,3))
# moving average
par(mfrow=c(2,1))
plot.ts(w, main="white noise")
plot.ts(v, main="moving average")
```

PROBLEM 01

Use the following data as basement for this problem:

```
set.seed(98234)# Creating example series
my_series <- 1:100 + rnorm(100, 0, 10)
my_series# Printing series
#example data is a series of numeric values with a length of 100
```

Compute Moving Average (rolling average, running average) Using User-Defined Function:

```
moving_average <- function(x, n = 5) {
# Create user-defined function
  stats::filter(x, rep(1 / n, n), sides = 2)
}
my_moving_average_1 <- moving_average(my_series)    # Apply user-defined
function
my_moving_average_1 # Printing moving average
```

Compute Moving Average Using rollmean() Function of zoo Package

```
install.packages("zoo")                             # Install zoo package
library("zoo")                                      # Load zoo
my_moving_average_2 <- rollmean(my_series, k = 5)   # Apply rollmean
function
my_moving_average_2                                 # Printing moving
average
```

PROBLEM 02

Calculate the three-point moving averages using the above “my_series”.(use both methods)

PROBLEM 03

The following table gives information about the number of people staying in a hotel each quarter in 2021 and 2022.

Year	**2021**				**2022**			
Quarter	1	2	3	4	1	2	3	4
Number of people	261	353	372	290	193	309	292	202

(a) Calculate the 4-point moving averages for this information.
The first three has been done for you.
319, 302, 291, ………, ………

(b) Plot the 4-point moving averages.
(x-Year and quarter, y- Number of people)

(c) Describe what the moving averages show about the trend in the number of people staying at the hotel over this period.

PROBLEM 04

The table shows the number of laptops sold in each of the first five months of 2012.

Month	January	February	March	April	May
Number of laptops	2190	2220	2280	2250	2280

(a) Work out the 3-point moving averages for the first five months of 2012.

The 3-point moving average of the number of laptops sold in April, May and June of 2012 was 2300.

(b) Work out the number of laptops sold in June 2012.
(c) Plot the time series graph with the trend line.
(d) Describe what the moving averages show about the trend in the number of laptops sold in the first six months 2012.

PROBLEM 05

The table gives information about the number of students who enroll for a course in each term in 2011, in 2012 and in 2013.

The 3- point moving averages are given correct to 3 significant figures.

Year	**Term**	**Number of students**	**3-point moving average**
2011	Autumn	330	
	Spring	180	200
	Summer	90	220
2012	Autumn	390	
	Spring	230	
	Summer	170	
2013	Autumn	460	
	Spring	260	
	Summer	210	

(a) Calculate the 3-point moving average values and the fill the table. (Give your answer correct to 3 significant figures.
(b) Plot the 3-point moving averages on the time series graph. (x-Term and year, y-Number of students).
(c) On the time-series graph, draw a trend line for the 3-point moving averages.
(d)

i. Use your trend line to find an estimate for the mean seasonal variation in numbers enrolling for the Autumn Term.
ii. Predict the number of students who will enroll in the Autumn Term of 2014.