MATH50010 coursework 2023-24

30/11/2023

This coursework is due at 1pm on Friday 9th December. Please submit it via the turnitin link on blackboard. Your submission should contain your CID but not your name.

The Task

In this coursework, we will analyse the amount of Nitrogen Dioxide (NO_2) in the environment. The dataset 'tower_bridge.csv' contains the daily average NO_2 level on Tower Bridge Road in London for every day in 2022. It contains the following columns:

- Site: the location, should be SK8 for all entries.
- Species: the particle being measured, should be NO2 for all entries.
- ReadingDateTime: the date and time the reading was taken (note that time will be 0 as it is a daily average).
- Value: the average amount of NO_2 recorded.
- Units: the units of the NO_2 reading, should be ug m-3 ($\mu g/m^3$, micorgrams per cubic meter) for all entries.

The dataset is available to download from blackboard. In this coursework we are interested in determining how frequently we have high pollution periods.

The following is a step-by-step workflow to guide you through the task. Your coursework submission should be written using RMarkdown, and compiled to a PDF for submission. All code should be commented clearly. For the highest marks, you should communicate to the marker clearly what you are trying to do, and justify any arbitrary choices. There are a total of 50 marks available for this coursework. 6/50 marks are available for an extension question, you can still get a good mark overall without attempting this question.

```
# What follows is an example solution, to help the GTAs.
# It contains all necessary code, though not necessarily all the comments and interpretation.
# Other approaches are certainly possible.
```

(3 marks) Loading and exploration

1. Read the data in to R.

```
library(stringr)
dat <- read.csv("~/Downloads/tower_bridge.csv")
```

2. (1 marks) We want to split the data into low and high pollution levels. Typically it is assumed that the pollution level is high if the NO_2 level exceeds $40\mu g/m^3$. Create a new variable called 'state' indicating whether the pollution is high (1) or low (0) on each day.

```
t=40
dat$state = as.numeric(dat$Value >=t) #>t is also acceptable
#state_string<-paste(dat$state,collapse="")
```

3. (2 marks) Calculate the proportion of days in each of the two states defined above.

```
state_counts<-c(nrow(dat)-sum(dat$state), sum(dat$state))
#state_counts<-str_count(state_string,pattern=c("0","1"))
```

```r
n_times<-sum(state_counts)
state_props<-state_counts/n_times
print(state_props) #(0,1)
```

```
## [1] 0.8027397 0.1972603
```

(18 marks) A Markov Chain Model

We will now model the data as a Markov Chain.

4. (3 marks) We look at the transitions between states. Count the number of pairs in each of the possible pairs of successive states (0,0), (0,1), (1,0),(1,1). Overlaps are OK, e.g. the sequence 0100 corresponds to one (0,1) transition, one (1,0) transition and one (0,0) transition.

```r
state_string<-paste(dat$state,collapse="")
pair_counts<-str_count(state_string,paste0("(?=",c("00","01","10","11"),")"))
print(pair_counts)
```

```
## [1] 257  35  35  37
```

5. (5 marks) Assume that the high/low pollution state forms a two-state time-homogeneous Markov chain. Use the data to estimate the transition matrix of the chain.

```r
pair_mat<-matrix(pair_counts,nrow=2,byrow=TRUE)
phat<-pair_mat/rowSums(pair_mat)
phat
```

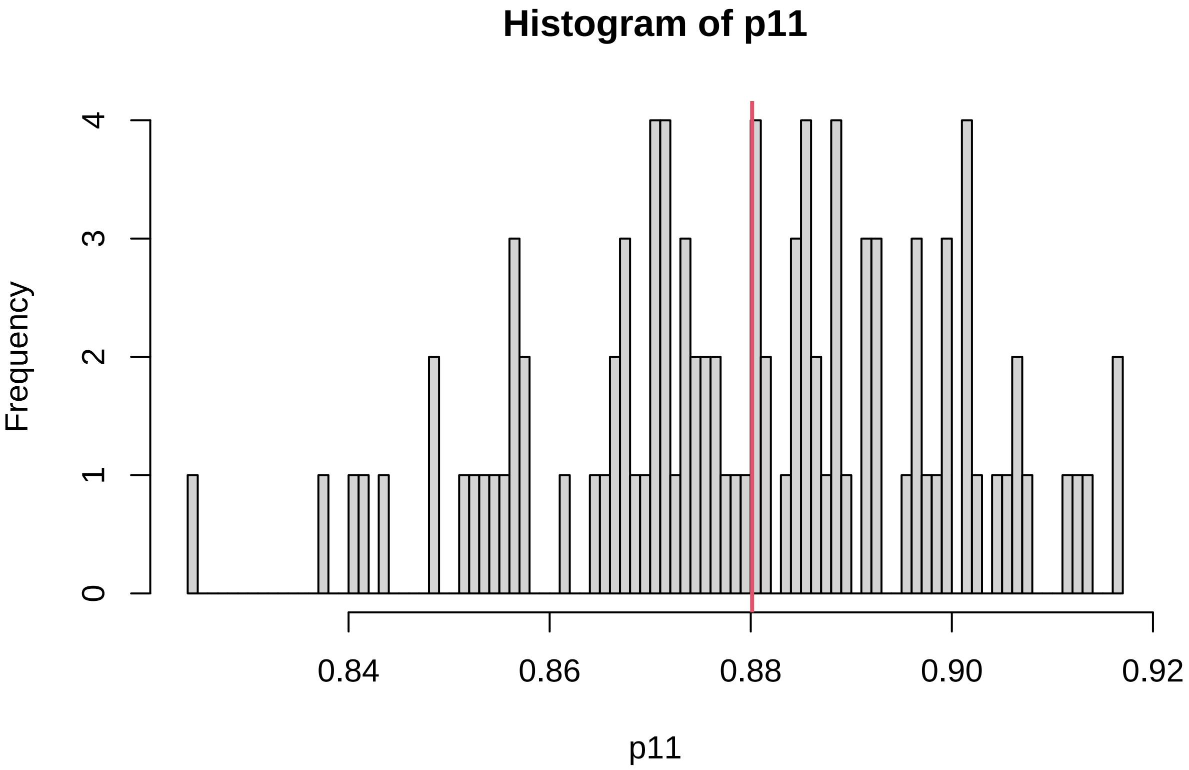
```
##           [,1]      [,2]
## [1,] 0.8801370 0.1198630
## [2,] 0.4861111 0.5138889
```

6. (3 marks) Write a function that simulates draws of length `m` from a two state Markov chain with states 0 and 1.

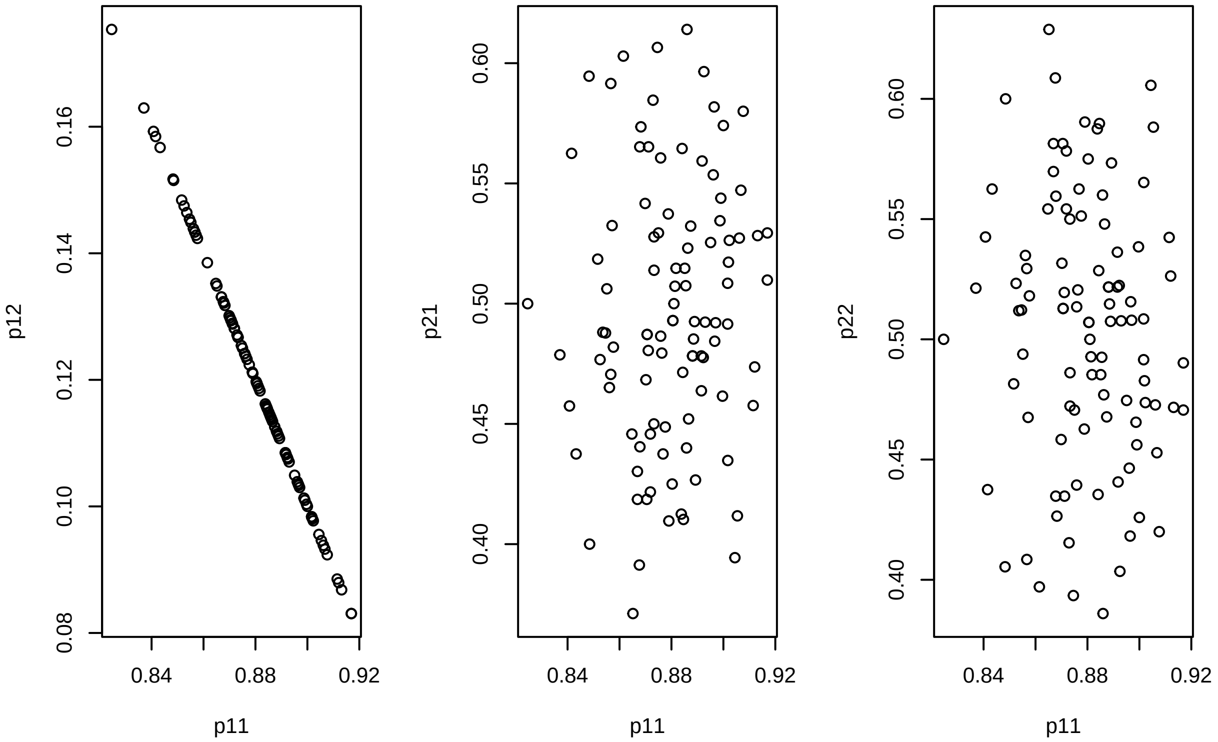
```r
sim<-function(m, initial, transition){
  store<-rep(0,times=m)
  store[1]<-sample(1:2,1,prob=initial)
  for (i in 2:m){
    store[i]<-sample(1:2,1,prob=transition[store[i-1],])
    }
c(0,1)[store]
}
```

7. (7 marks) Use your function to simulate n independent 'years' of daily high/low classifications of NO_2 using the transition probabilities from the data. For each of the n realizations of the chain, compute the estimates of the transition probabilities. Show that the estimators are approximately unbiased. Are the estimates of different transition probabilities correlated?

```r
n<-100
p11<-rep(0,times=n)
p12<-rep(0,times=n)
p21<-rep(0,times=n)
p22<-rep(0,times=n)

for(i in 1:n){
  dat_new<-sim(365,state_props,phat)
  str_new<- paste(dat_new,collapse="")
  pair_counts_new <-str_count(str_new,paste0("(?=",c("00","01","10","11"),")"))
  pair_mat_new<-matrix(pair_counts_new,nrow=2,byrow=TRUE)
  phat_new<-pair_mat_new/rowSums(pair_mat_new)
  p11[i]<-phat_new[1,1]
  p12[i]<-phat_new[1,2]
  p21[i]<-phat_new[2,1]
  p22[i]<-phat_new[2,2]
}

# plot just one
hist(p11,breaks=100)
abline(v=phat[1,1],col=2,lwd=2)
```



```
par(mfrow=c(1,3))
plot(p11,p12)
plot(p11,p21)
plot(p11,p22)
```



```r
cor.test(p11,p12)
```

```
##
##  Pearson's product-moment correlation
##
## data:  p11 and p12
## t = -664343859, df = 98, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -1 -1
## sample estimates:
## cor
##  -1
```

```r
cor.test(p11,p21)
```

```
##
##  Pearson's product-moment correlation
##
## data:  p11 and p21
## t = 0.85041, df = 98, p-value = 0.3972
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1127244  0.2773444
## sample estimates:
##        cor
## 0.08558879
```

```r
cor.test(p11,p22)
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  p11 and p22
## t = -0.85041, df = 98, p-value = 0.3972
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2773444  0.1127244
## sample estimates:
##         cor
## -0.08558879
```

```
# Estimates of probabilities that sum to 1 (e.g. P[1,1] and P[1,2]) are of course correlated.
# However, estimates for states that are unconstrained do not appear to show much correlation
```

(14 marks) Testing the Markov Model

We now need to test whether the pollution level does in fact depend on the pollution level of the previous day.

8. (3 marks) Write down a formal hypothesis test in terms of the transition probabilities to test whether the probability of the pollution level being high is independent of the pollution level of the previous day.

```
# If the probability of reaching a high state is independent of the pollution level of
# the previous day, then $p_{11}=P(X_1=1|X_0=1)=P(X_1=1)=P(X_1=1|X_0=0)=p_{01}$.
# Hence we want to test
#H_0: p_{11}=p_{01} vs H_1: p_{11} \neq p_{01}.
```

If we want to test whether two sampled data sets x_A and x_B of sizes n_A,n_B come from the same Bernoulli distribution, we can perform a hypothesis test using a Hypergeometric distribution. In particular, let p_A and p_B be the population positivity probabilities of the two data sets, then we want to test $H_0 : p_A = p_B$ vs $H_1 : p_A \neq p_B$. Note that under H_0, both data sets x_A and x_B are from the same distribution. Therefore the number of positive samples in x_A follows a Hypergeometric$(n_A + n_B, n_A, s_A + s_B)$ distribution where s_A and s_B are the number of positive samples in x_A and x_B respectively (see e.g. Section 8.6.1 of Ross (2020)). The Hypergeometric(N, K, n) distribution for $n \leq N$ has the density:

$$P(X = x|N, K, n) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \qquad \text{for } x \in \{0, 1, \dots, K\}, \text{ and } n - x \in \{0, 1, \dots, N - K\}.$$

To perform the hypothesis test, we use the p-value given by,

$$p = 2 \min\{P(X \leq s_A), P(X \geq s_A)\} \qquad \text{for } X \sim \text{Hypergeometric}(n_A + n_B, n_A, s_A + s_B).$$

This test is sometimes known as the Fisher-Irwin test.

9. (6 marks) Calculate the p-value for your hypothesis test.

```
# We first need to create the two data sets, the first containing all states encountered
# after a 1, and the second containing all states encountered after a 0.
# We can calculate this manually or observe that from `pair_counts` we have 115 samples
# from 10, 145 from 11 and 99 from 00 and 116
# from 01. From this we can create our datasets.
xA<- c(rep(0,pair_counts[3]),rep(1,pair_counts[4]))
xB<-c(rep(0, pair_counts[1]),rep(1,pair_counts[2]))

# we similarly extract the other statistics of the data that we need
nA<-pair_counts[3]+pair_counts[4]
nB<-pair_counts[1]+pair_counts[2]
sA<-pair_counts[4] # number of positives - 11's
sB<-pair_counts[2] # number of positives - 01's
```

```
# calculate the p value
# note the hypergeometric function in R uses a different parameterization
pleq<-phyper(sA,nA,nB,sA+sB)
pgeq<- 1- phyper(sA-1,nA,nB,sA+sB)  # use s_A-1
p=2*min(pleq,pgeq)
print(p)
```

```
## [1] 1.008571e-11
```

10. (2 marks) Conduct the hypothesis test at the 5% significance level.

```
# We should reject $H_0$ if $p<\alpha$ where $\alpha=0.05$ is the significance level.
# We see here that $p<\alpha$. Hence we reject $H_0$.
# There is insufficient evidence to conclude that $p_{01}$ is equal to $p_{11}$.
```

11. (3 marks) Perform a similar hypothesis test to determine whether the probability of the pollution level being low is independent of the previous days level.

```
# We want to test $H_0: p_{10}=p_{00} vs H_1:p_{10} \neq p_{00}$
# However note that $p_{00}=1-p_{01}$ and $p_{10}=1-p_{11}$ so the event $p_{10}=p_{00}$ is
# equivalent to the event $p_{01}=p_{11}$.
# Therefore the hypotheses tested will be equivalent to those in questions 10-12.
# Our conclusions thus remain the same.
# There is insufficient evidence to conclude that $p_{00}$ is equal to $p_{10}$.
```

(12 marks) Investigating the Number of Consecutive Low Pollution Days

We now want to investigate the likelihood of consecutive low pollution days.

12. (3 marks) Using the estimated Markov model, calculate the probability that after any high pollution day, it is over a week until the next high pollution day, i.e. calculate $P(X_1 = 0, .., X_7 = 0|X_0 = 1)$.

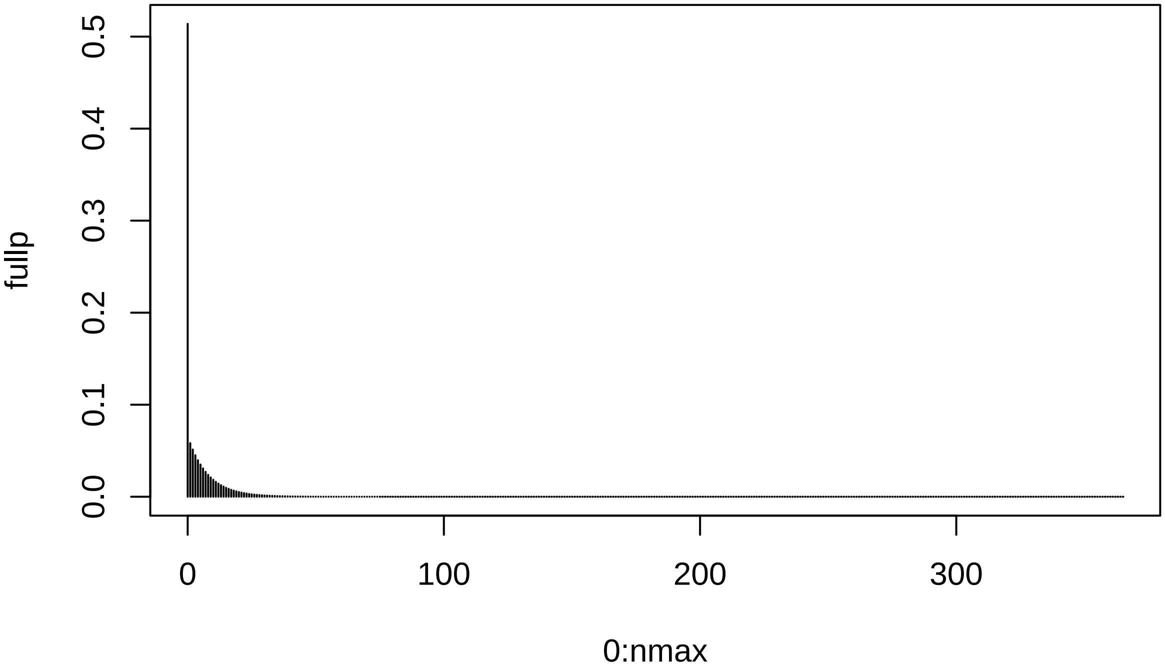
```
# Consider P(X_1=0,.., X_7=0|X_0=1)= P(X_1=0|X_0=1)P(X_2=0|X_1=0)...P(X_7=0|X_6=0)
phat[2,1]*phat[1,1]^6
```

```
## [1] 0.2259629
```

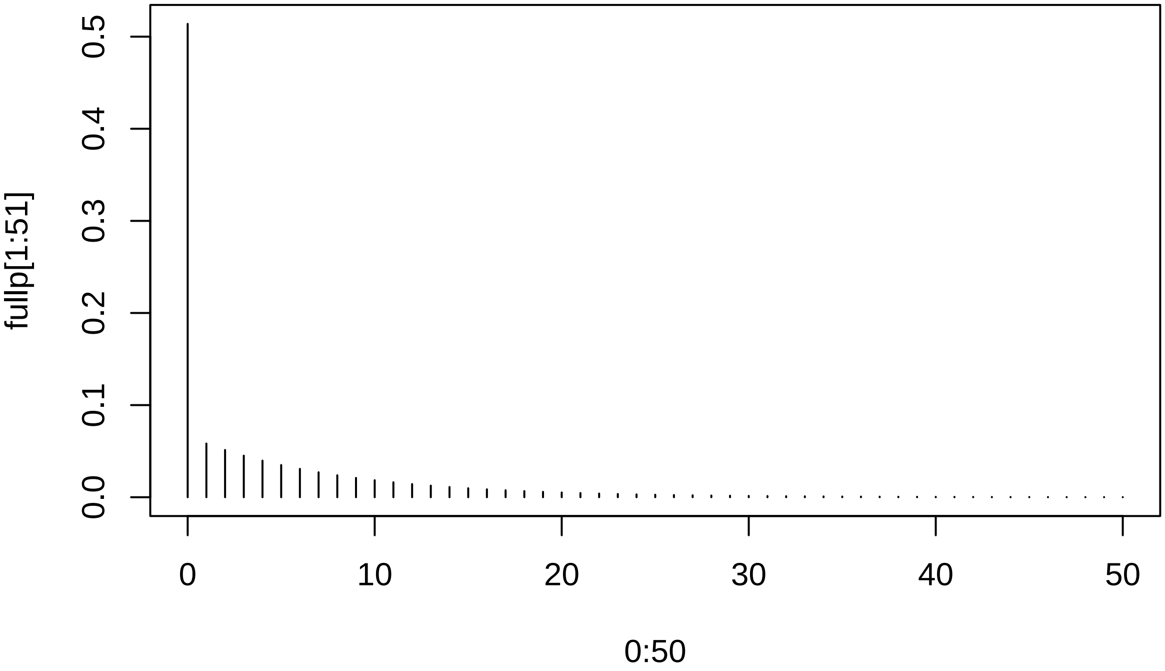
13. (6 marks, Extension) Using the estimated Markov model, plot the probability mass function of the number of consecutive low pollution days immediately after any high pollution day, i.e. let M be the number of consecutive low pollution days starting from day 1, plot $P(M = m|X_0 = 1)$ for all m.

```
# define a function to calculate probability of exactly n consecutive low pollution days
# after a high pollution day
f<-function(n,P){
  ff<-P[2,1]*P[1,1]^{n-1}*P[1,2] # need the P[1,2] term to get prob of exactly n
  # (alternatively subtract (n+1)th prob)
  return (ff)
}

# Use the function to calculate prob for all n
nmax<-365 # need to truncate at some threshold
p<-rep(0,nmax)
for (n in 1:nmax){
  p[n]<-f(n,phat)
}

# need to include probability of 0 consecutive low pollution days to make it valid pmf
```

```r
fullp<- append(phat[2,2],p)

# plot
plot(0:nmax, fullp, type='h')
```



```r
# we can also zoom in a bit
plot(0:50, fullp[1:51], type='h')
```



14. (3 marks) Using the data directly, calculate the average number of consecutive low pollution days between high pollution days.

```
# from every 1 state followed by a 0, calculate how long it is until we see another 1.
# the average number of consecutive 0's we see in the data.
mean(rle(dat$state)$lengths[rle(dat$state)$value==0])
```

```
## [1] 8.138889
```

(3 marks) Conclusion

15. (3 marks) Comment on any limitations of your study.

```
# E.g.
# 1)We have assumed each day is essentially the same, we could also account for seasonality.
# 2)We have not tested whether there is a longer effect on the pollution levels, e.g. the
# pollutants could linger in the environment for several days effecting the pollution levels
# on more than just the next day.
# 3)We have divided the pollution levels into 'high' and 'low', could do something more granular.
```

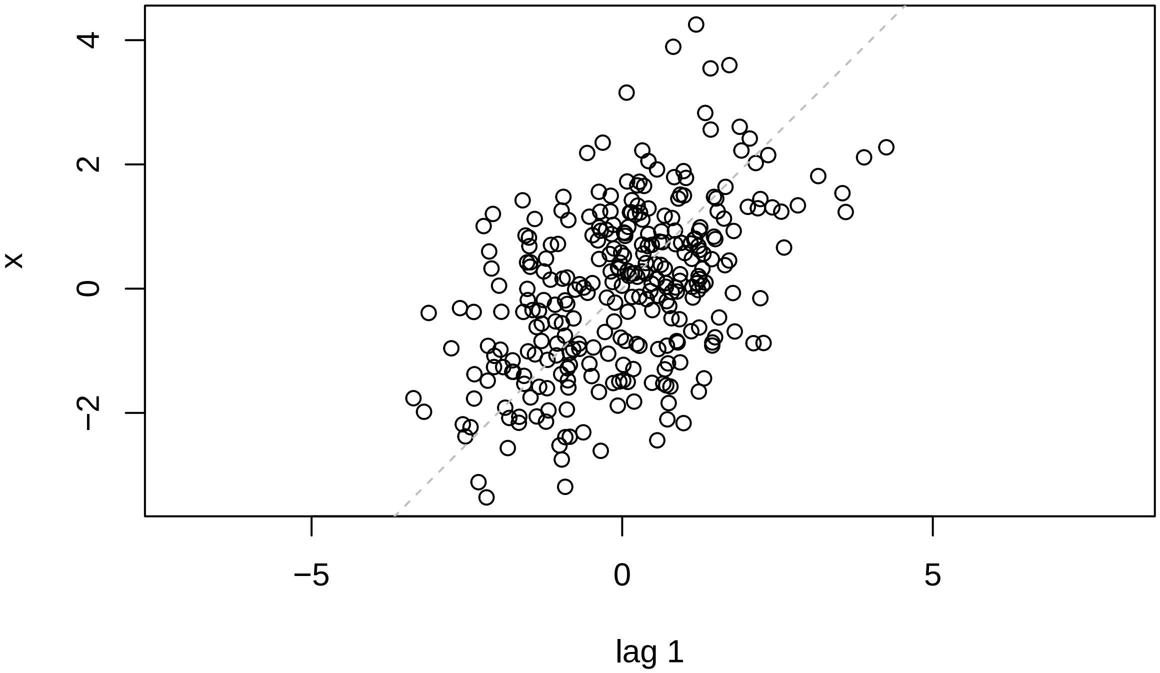
Academic integrity

You are welcome to use any sources (websites, books, etc), but you should cite them. If you make use of chunks of code that you have found, you should also cite the source. You should write your own submission, including all code. Failure to do so will be considered misconduct.

Useful functions

Scatter plot of x_t against x_{t+1}

```
x<-rnorm(300)
x<-x + c(0,x[-300]) #make correlated data
lag.plot(x) # plot
```



Make a vector into a single string.

```
paste(c(0,1,1,0),collapse="")
```

```
## [1] "0110"
```

Count instances of a single letter

```
library(stringr)
str_count("101010",c("0","1"))
```

```
## [1] 3 3
```

Count (overlapping) instances of a double letter

```
str_count("00010100111",paste0("(?=",c("00","01","10","11"),")"))
```

```
## [1] 3 3 2 2
```

Count the lengths of consecutive runs of 1's or 0's

```
vec<-c(0,0,0,1,0,0,1,1,0,0,0,0,1,1,1)
rle(vec)
```

```
## Run Length Encoding
##   lengths: int [1:6] 3 1 2 2 4 3
##   values : num [1:6] 0 1 0 1 0 1
```

In the output, 'lengths' represent the number of consecutive values and 'values' tells us which value is repeated. We can extract each using e.g. `rle(vec)$values`.

Reference:

Ross, S. M. (2020). Introduction to probability and statistics for engineers and scientists. Academic press.