TMA150/MSG400 Assigment 2

Martin Hansson

2020-09-22

1 Introduction

This report covers Assignment 4 in the course TMA150/MSG400. The assignment consists of 2 parts; *Part 1: Stock Investment* and *Part 2: Medical Age Assessment*. Both parts are solved in R.

2 Question 1 - Stock Investment

2.1 Problem

In a data file *stockvalues.txt*, daily stock prices between 2002-06-03 to 2006-06-01 are listed for 7 different stocks. In this part we look at a 100 day investment in these stocks.

a. Generate *S=1000* vectors *V*, from a *multivariate normal distribution* where each vector contains the expected relative change (log-tranformed) in price for a 100 day period for each stock, i.e. each V vector contains one value for each stock. It assumed that there are underlying trends in the data and that some stocks move together but the expected change from one day to another is independent.

b. Write an R-function that for given proportions of each stock, $w_1 - w_7$ (where $\sum_0^7 w_i = 1$), and utility function parameter k, computes the utility u. Use this function to compute the utility for k=[-0.5,0.5,1.5] both when the investment is equally distributed between the stocks and when the stocks with the best expected performance receives all the investment, i.e. $w_{best} = 1$ and all other $w_i = 0$.

c. Assume you are only investing in stocks S3 and S4. Find the optimal weights, w_3 and w_4, when k=-0.5 and k=1.5.

d. Discuss problems and weaknesses of using the model in this assignment as basis for investments decisions.

2.2 Theory and implementation

When doing stock investment portfolio analysis it is common to use *log(Return)*, i.e. $\log(p_{i+1}/p_i)$ instead of using *Raw return*, i.e. $(p_{i+1} - p_i)/p_i)$. There are several benefits of using log-transformed value, such as the possibility to use *Normal distributed returns*, *time-additivity*, *mathematical ease* (for example when integrating) and *numerical stability*. The reasoning behind these are outside the scope of this report.

If X_{ij} is the closing price for stock i at day j, we define, for i=1,...,m and j=1,...,N.

$$Z_{ij} = \log \frac{X_{ij}}{X_{i,j-1}} = \log(X_{ij}) - \log(X_{i,j-1})$$

If we assume there are underlying trends and the some stocks move together and that the day to day changes are independent, the n-day return V (containing all stocks) can be described with a multivariate normal distribution with the following parameters

$$V \sim Normal(n\gamma, n\Sigma)$$

where γ is equal to the average expected log-transformed daily return and Σ is the co-variance matrix for log-transformed data.

To decide upon the best possible portfolio based on a given V, we can compute the *utility* of the outcome using the utility function *u(x)*:

$$u(T) = \frac{1 - (\sum_{i=1}^{m} w_i \exp(V))^{-k}}{k}$$

To take into account that V is random variable, the utility u can be computed for *S=1000* sample vectors $V_{q,..,S}$ and we get the expectation for *u(V)* as

$$E(u(V) = \frac{1}{S}\sum_{q=1}^{S} u(V_q)$$

which also is a measure of the performance of the portfolio.

2.3 Results and discussion

a. The V-vectors were generated with R-function **rmnorm** which simulates from a multivariate normal distribution. The co-variance and mean for the log-transformed data was obtained with R-functions **cov** and **mean**. See R-code for details.

b. See R-code for details of the function computing the utility for given *w,k,V*.

In table 1, the utility are shown for the case when the stocks are equally weighted and the case where only one stock (at a time) gets full weight. It might be expected that the best performing single stock would always have a higher utility than the mixed portfolio, but this is not always the case. It seems like the mixed portfolio have increased performance i relation to the single best stock as k increases. With k=-0.5, the mixed portfolio has lower utility while it is higher for k=1.5. The k is a measure of the investors aversion to risk where a higher k means more risk averse. Money lost is affecting the utility more than corresponding money gained. It is then reasonable that the utility for the mixed portfolio increases (lower variance) compared to the single best stock as k increases.

Table 1: Utility for for equal weighted investment and all investment in single stocks

	$w_i = 1/n$	$w_1 = 1$	$w_2 = 1$	$w_3 = 1$	$w_4 = 1$	$w_5 = 1$	$w_6 = 1$	$w_7 = 1$
k=-0.5	0.048	-0.00096	0.018	0.073	0.065	0.02	0.046	0.040
k=0.5	0.036	-0.019	-0.0036	-0.0011	0.036	-0.0040	0.033	0.027
k=1.5	0.026	-0.039	-0.025	-0.081	0.0086	-0.036	0.021	0.015

c. The utility when only investing in S3 and S4 can be maximized by maximizing the single parameter w_3 using the R-function **nlm**. Then $w_4 = 1 - w_3$. See R-code for algorithm used. Table 2 shows the results. As expected, the utility for the combined S3/S4 is higher than for the corresponding single stocks in table 2. If it was lower then the weight could be set to 1 for one of them and we would get at least the same utility.

Table 2: Weights and utility when only investing in S3 and S4. Utility maximized.

	w_3	w_4	u
k=-0.5	0.604	0.396	0.0798
k=1.5	0.289	0.711	0.0256

d. Some weaknesses with model and the assumption in this assignment are discussed below.

- Assumption on independent data from day-to-day changes are most likely not correct as there will be time dependency that could be useful to incorporate in the model.
- Extrapolating data from historical observations is always tricky. It is hard to determine how well the model actually perform. One way to test the model could be to use some of the data points (for example 30%) as test data.
- The model only includes data from stock prices. It might be useful to include other surrounding factors influencing the economy in general.
- It is difficult to determine for which k (i.e. risk level) to base your decisions on. This is be parameter influenced by the individual investor's preference.

3 Question 2 - Medical Age Assessment

3.1 Problem

In this task *Medical Age Assessment* are studied. A medical feature, in this case the maturity of knees, is used to assess weather a person is an adult (i.e. above age of 18) or a child (i.e. below age of 18). A knee can be reported as *mature* or *immature*. Data on 100 knees (50 immature and 50 mature), together with the actual age, are used in this study.

a. Fit a logistic regression for the available data, by maximizing the likelihood for the given parameters a and b, i.e. find the a and b that maximizes the likelihood function. Plot the data and the logistic regression curve and evaluate the results.

b. Compute and plot the posterior with image plotting function in R, using grid size of 21x21 for a and b evenly spaced between [-0.5,2] and [0.5,3] respectively.

c. Assume that the forensic report is *mature knee*. The age of people subjected to the test are assumed to follow a Gamma distribution with some parameter μ and α (see section 3.2). Given a, b, μ, α and *c(x)*(see section 3.2), write an R-function that computes the difference between the cost of classifying as a child and the cost of classifying as adult. Use a and b that was found in (a). Compute the cost difference for the both cost functions given in section 3.2 and for the following μ and α. Conclude!

Table 3: Age distribution of interest

	μ	α
Case i	18.5	3
Case ii	19.5	6
Case iii	20.58	3

d. Repeat (c) but now take into account the uncertainty in a and b, by using descretization and averaging over the a- and b-intervals used in (b). Compare the results and comment!

3.2 Theory and implementation

A Bernoulli distribution is a discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability *1-p*. In the case of Medical Age Assessment, the probability *p(x)* can be computed with the following formula

$$p(x) = \frac{exp(a + b(x - 18))}{1 + exp(a + b(x - 18))}$$

where a and b are constants. However, using Bayesian inference, the uncertainties in the parameter a and b could be taken into account. With the data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, the probability of this data given a and b, i.e. the likelihood function can be computed as

$$L(a, b) = \prod_{i=1}^{n} \left(\frac{exp(a + b(x_i - 18))}{1 + exp(a + b(x_i - 18))} \right)^{y_i} \left(1 - \frac{exp(a + b(x_i - 18))}{1 + exp(a + b(x_i - 18))} \right)^{1-y_i}$$

By multiplying $L(a, b)$ with the *prior* $\pi(a, b)$, i.e. the function that describes the uncertainty of a and b, we get the posterior distribution up to a constant. In our case we are using a uniform prior, which basically means we are multiplying with a *constant* and the shape of the posterior distribution will have the same shape as the likelihood distribution. In the case of *Medical Age Assessment*, this will be enough to estimate weather the cost difference is above or below 0, which is what we will base or decisions on.

To assess weather a person should be classified as a child or as an adult we need to determine a cost function, c(x), (cost in terms of ethical, economic other possible aspects) for misclassification. In our case we define two possible cost functions, $c_1(x)$ and $c_2(x)$:

$$c_1(x) = \begin{cases} B & \text{if } x \leq 18 \\ 1 & \text{if } x > 18 \end{cases}$$

$$c_2(x) = \begin{cases} B(18 - x) & \text{if } x \leq 18 \\ x - 18 & \text{if } x > 18 \end{cases}$$

In this case we have, $B = 10$, which basically means that the cost of classifying a child as adult is 10 times higher than classify an adult as a child. If a person is classified as a child, the expected cost of a possible misclassification can now be computed with

$$C_c = \int_{18}^{\infty} \pi(x; \mu, \alpha) f_k(x) c(x) dx$$

The same way we get misclassification as adult

$$C_a = \int_{0}^{18} \pi(x; \mu, \alpha) f_k(x) c(x) dx$$

The function $f_k(x)$ depends on weather the report is *mature* or *immature* and $\pi(x; \mu, \alpha)$ is the age distribution defined as

$$\pi(x) = \text{Gamma}(x - 14; \alpha, \alpha/(\mu - 14))$$

3.3 Results and discussion

a. The function $L(a, b)$ was maximized w.r.t. a and b by minimizing $-\log(L(a, b))$ using R-function **nlm**. The reason to use log is to avoid numerical problems as the product will be extremely small. The values of a and b came out to be:

$$\boxed{a_{max} = 0.6846703}$$

$$\boxed{b_{max} = 1.721258}$$

Figure 1 shows the plot of the regression curve $p(x)$, when $a = a_{max}$ and $b = b_{max}$, together with the data points. It shows that if the age is around 15 or below the probability of having a mature knee is close to zero while if the age is 20 or above it is almost certain to have a mature knee. Between age of 15 and 20 it gradually changes from 0 to 1, which is reasonable.

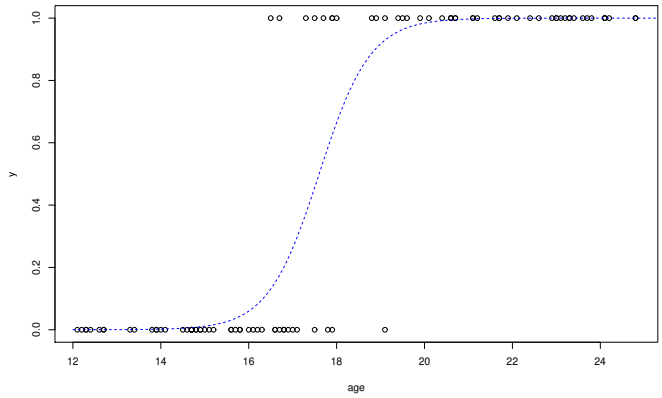

Figure 1: The logistic regression curve together with the data points

b. The plot of the posterior (up to a constant) are shown in figure 2. Here, the results in (a) are confirmed since we see the maximum at a_{max} and b_{max}. Note that this is not proper distribution since the it will not integrate to 1.

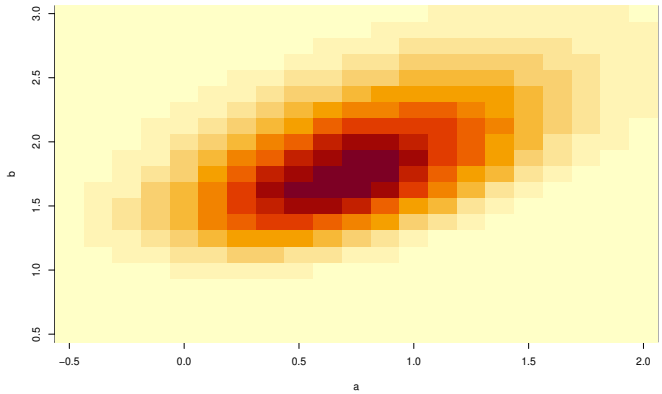

Figure 2: The posterior over the parameters a and b in the logistic regression

c. The function will return the value of $C_c - C_a$, given a, b, *mu*, *alpha* and $c(x)$. Since we are assuming the report is *mature knee* the function f_k is

$$f_k(x) = \frac{exp(a + b(x - 18))}{1 + exp(a + b(x - 18))}$$

The R-function **integrate** is used to compute the integral. In order to get numerical stability f_k is rewritten as

$$f_k(x) = \frac{1}{exp(-(a + b(x - 18))) + 1}$$

The results are shown in table 4. The following can be observed.

- Using cost function, $c_1(x)$, the person should be classified as a child for all age distributions as the cost of misclassification as child is higher.
- Similarly, using cost function, $c_2(x)$, the person should be classified as an adult for all age distributions.
- When μ increases, i.e. the ages of the people subjected to the test increases, the expected cost difference of being misclassified as child and misclassified as adult decreases which is reasonable results.

Table 4: The cost difference, $C_c - C_a$, for different $c(x)$ and age distributions when $a = a_{max}$ and $b = b_{max}$

	$\mu = 18.5, \alpha = 3$	$\mu = 19.5, \alpha = 6$	$\mu = 20.58, \alpha = 3$
$c_1(x)$	0.616	0.219	0.00134
$c_2(x)$	-0.441	-1.229	-2.470

d. The uncertainty of a and b are taken into account by averaging over the intervals in (b). In this case, descretization steps of 0.01 was used for a and b. The procedure is to repeat (c) for every combination of a and b and then divide the sum of all computed cost differences with the number of combinations. The cost differences are shown in table 5. The conclusion from (c) remains, however it can be noticed that the cost difference increases for all combinations compared to (c). This means that it leans more towards classifying a person with mature knee as a child.

Table 5: The cost difference, $C_c - C_a$, for different $c(x)$ and age distributions when averaging over a=[-0.5,2] and b=[0.5,3].

	$\mu = 18.5, \alpha = 3$	$\mu = 19.5, \alpha = 6$	$\mu = 20.58, \alpha = 3$
$c_1(x)$	0.845	0.325	0.124
$c_2(x)$	-0.0291	-1.077	-2.259

Appendix - code

Question 1 R-code

```
library("LearnBayes")
S=as.matrix(read.csv("stockvalues.txt"))
n_st=ncol(S) #number of stocks
n_days=nrow(S) #number of days

Z=matrix(nrow=n_days-1,ncol=n_st) #matrix for log(return)
Z_mean=numeric(n_st)
for (i in 1:{n_days-1}) {
  Z[i,]=log(S[i+1,]/S[i,])
}
for (i in 1:n_st) {
  Z_mean[i]=mean(Z[,i]) #gamma value for multinormal dist
}
Z_cov=cov(Z) #covariance of log(Z)

#1a ##########################################
S=1000
n=100
V <- rmnorm(S, n*Z_mean, n*Z_cov) #generating 1000 V-vectors

#1b ##########################################
#u=function computing utility
u=function(w,k,V) {
  s=matrix(nrow=nrow(V),ncol=1,0)
  for (i in 1:length(w)) {
    s=s+w[i]*exp(V[,i])
  }
  u=(1-s^-k)/k
  return(u)
}

#Case 1 - equal weight
w=rep(1/n_st,n_st) #equal weight
k=c(-.5,.5,1.5) #k-values
u1=u(w,k[1],V) #u for k1
E1=sum(u1)/S #Expected u when simulating S V-vectors
u2=u(w,k[2],V)
E2=sum(u2)/S
u3=u(w,k[3],V)
E3=sum(u3)/S

#Case 2 - "best" stock only
E_s=matrix(ncol=n_st,nrow=3)

for (i in 1:n_st) {
  w=numeric(n_st) #all w=0 except for one stock (at a time)
  w[i]=1
  u1i=u(w,k[1],V) #u for k1 and w_i
  E_s[1,i]=sum(u1i)/S
  u2i=u(w,k[2],V)
  E_s[2,i]=sum(u2i)/S
  u3i=u(w,k[3],V)
  E_s[3,i]=sum(u3i)/S
}
```

```
#1c ##############################################
Ef <- function(w3,k0,V0,S0) {
  -sum(u(c(w3,1-w3),k0,V0))/S0
}
knew=c(-.5,1.5) #k-values of interest
Vnew=V[,3:4] #V vectors with only S3 and S4

w3_1=nlm(Ef,.5,V0=Vnew,k0=knew[1],S0=S) #optimizing over w3 -->w4=1-w3
w3_2=nlm(Ef,.5,V0=Vnew,k0=knew[2],S0=S)
```

Question 2 - R-code

```
library("LearnBayes")
library("Rlab")
MK=as.matrix(read.csv("matureknee.txt"))
IK=as.matrix(read.csv("immatureknee.txt"))
n=length(MK)+length(IK)

#Merge data to one matrix
K=matrix(nrow=n,ncol=2)
K[,1]=rbind(MK,IK)
K[1:{n/2},2]=rep(1,n/2)
K[{n/2+1}:n,2]=rep(0,n/2)

x=K[,1] #age
y=K[,2] #report, Mature=1, Immature=0

#2a
########################################
#Lf_log = -log (Likelihood)
Lf_log <- function(z) {
  a <- z[1]
  b <- z[2]
  -sum(y*log(exp(a+b*(x-18))/(1+exp(a+b*(x-18))))+(1-y)*log((1-exp(a+b*(x
      -18))/(1+exp(a+b*(x-18))))))
  }
max_Lf=nlm(Lf_log,c(1,1)) #minimizing(-log(Lf))
a_max=max_Lf$estimate[1]
b_max=max_Lf$estimate[2]
x0=seq(12,26,0.1)
p=exp(a_max+b_max*(x0-18))/(1+exp(a_max+b_max*(x0-18))) #calculating
    logistic regression curve
plot(x,y,xlab='age')
lines(x0,p,type="l",col = "blue",lty="dashed")

#2b plotting posterior
########################################
Lf_log2 <- function(a,b,x,y) {sum(y*log(exp(a+b*(x-18))/(1+exp(a+b*(x-18)))
    )+(1-y)*log((1-exp(a+b*(x-18))/(1+exp(a+b*(x-18))))))}
s=0
a=seq(-.5,2,length.out = 21) #grid 21x21
b=seq(.5,3,length.out = 21)
Lf=matrix(nrow=length(a),ncol=length(b))
k=0
#for-loop for combinations of a and b
for (ai in a) {
  k=k+1
  l=0
```

```
  for (bi in b) {
    l=l+1
    Lf[k,l]=exp(Lf_log2(ai,bi,x,y)) #Lf(a,b)
  }
}
image(a,b,Lf)


#2c - cost diff when a=a_max, b=b_max
#######################################
cost <- function(a,b,c,mu,alp) {
  B=10
  g <- function(x) {dgamma(x-14,alp,alp/(mu-14))} #age distribution
  f <- function(x) {1/(exp(-(a+b*(x-18)))+1)} #f(x)
  if (c==1) {
    c_low <- function(x) {B}
    c_up <- function(x) {1}
  }
  else {
    c_low <- function(x) {B*(18-x)}
    c_up <- function(x) {x-18}
  }
  f_g_c_low <- function(x) {f(x)*g(x)*c_low(x)} #cost function for x<=18
  f_g_c_up <- function(x) {f(x)*g(x)*c_up(x)} #cost function for x>18
  c_c=integrate(f_g_c_low,0,18)$value #cost of misclass to child
  c_a=integrate(f_g_c_up,18,200)$value #cost of misclass to adult
return(c_c-c_a)
}

q1=matrix(nrow=2,ncol=3) #cost diff for different combinations of age and c
    (x)
mu=c(18.5,19.5,20.58)
alp=c(3,6,3)
for (i in 1:2) {
  for (j in 1:3) {
    q1[i,j]=cost(a_max,b_max,i,mu[j],alp[j])
  }
}
q1

#2d - cost diff when averaging over a and b
#########################################
q2=matrix(nrow=2,ncol=3) #cost diff for different combinations of age and c
    (x)
mu=c(18.5,19.5,20.58)
alp=c(3,6,3)

for (i in 1:2) {
  for (j in 1:3) {
    n=0
    q_ab=0
    for (a in seq(-.5,2,.01)) {
      for (b in seq(.5,3,.01)) {
      n=n+1
      q_ab=q_ab+cost(a,b,i,mu[j],alp[j]) #sum for cost diff for all comb of
          a and b
      }
    }
```

```
    q2[i,j]=q_ab/n  #resulting cost diff for a given age distr and c(x)
  }
}
}
q2
```