1. Министерство науки и высшего образования Российской Федерации
2. Санкт-Петербургский Политехнический Университет Петра Великого
3. —

Институт компьютерных наук и технологий

Высшая школа искусственного интеллекта

Отчёт по лабораторной работе «Линейная регрессия»

по дисциплине «Машинное обучение»

Выполнила

студентка гр. 3540201/20302 Обидина А.И.

Проверил

профессор Уткин Л.В.

Санкт-Петербург

2022

**Задание**

1. Загрузите данные из файла reglab1.txt. Используя функцию lm, постройте регрессию (используйте разные модели). Выберите наиболее подходящую модель, объясните свой выбор.

2. Реализуйте следующий алгоритм для уменьшения количества признаков, используемых для построения регрессии: для каждого  выбрать подмножество признаков мощности , минимизирующее остаточную сумму квадратов *RSS*. Используя полученный алгоритм, выберите оптимальное подможество признаков для данных из файла reglab2.txt. Объясните свой выбор. Для генерации всех возможных сочетаний по m элементов из некоторого множества x можно использовать функцию combn(x, m, ...).

3. Загрузите данные из файла cygage.txt. Постройте регрессию, выражающую зависимость возраста исследуемых отложений от глубины залегания, используя веса наблюдений. Оцените качество построенной модели.

4. Загрузите данные Longley (макроэкономические данные). Данные состоят из 7 экономических переменных, наблюдаемых с 1947 по 1962 годы (*n=16*):

GNP.deflator - дефлятор цен,

GNP - валовой национальный продукт,

Unemployed – число безработных

Armed.Forces – число людей в армии

Population – население, возраст которого старше 14 лет

Year - год

Employed – количество занятых

Построить регрессию lm(Employed ~ .) .

Исключите из набора данных longley переменную "Population". Разделите данные на тестовую и обучающую выборки равных размеров случайным образом. Постройте гребневую регрессию для значений , подсчитайте ошибку на тестовой и обучающей выборке для данных значений λ, постройте графики. Объясните полученные результаты.

5. Загрузите данные EuStockMarkets из пакета « datasets». Данные содержат ежедневные котировки на момент закрытия фондовых бирж: Germany DAX (Ibis), Switzerland SMI, France CAC, и UK FTSE. Постройте на одном графике все кривые изменения котировок во времени. Постройте линейную регрессию для каждой модели в отдельности и для всех моделей вместе. Оцените, какая из бирж имеет наибольшую динамику.

6. Загрузите данные JohnsonJohnson из пакета «datasets». Данные содержат поквартальную прибыль компании Johnson & Johnson с 1960 по 1980 гг. Постройте на одном графике все кривые изменения прибыли во времени. Постройте линейную регрессию для каждого квартала в отдельности и для всех кварталов вместе. Оцените, в каком квартале компания имеет наибольшую и наименьшую динамику доходности. Сделайте прогноз по прибыли в 2016 году во всех кварталах и в среднем по году.

7. Загрузите данные sunspot.year из пакета «datasets». Данные содержат количество солнечных пятен с 1700 по 1988 гг. Постройте на графике кривую изменения числа солнечных пятен во времени. Постройте линейную регрессию для данных.

8. Загрузите данные из файла пакета «UKgas.scv». Данные содержат объемы ежеквартально потребляемого газа в Великобритании с 1960 по 1986 гг. Постройте линейную регрессию для каждого квартала в отдельности и для всех кварталов вместе. Оцените, в каком квартале потребление газа имеет наибольшую и наименьшую динамику доходности. Сделайте прогноз по потреблению газа в 2016 году во всех кварталах и в среднем по году.

9. Загрузите данные cars из пакета «datasets». Данные содержат зависимости тормозного пути автомобиля (футы) от его скорости (мили в час). Данные получены в 1920 г. Постройте регрессионную модель и оцените длину тормозного пути при скорости 40 миль в час.

# Ход работы

## **Пункт 1**

Были построены 3 модели линейной регрессии:

* Зависимость x от y и z

Стандартная ошибка – 0.07888

* Зависимость y от x и z

Стандартная ошибка – 0.06659

* Зависимость z от x и y

Стандартная ошибка – 0.3376

Среднеквадратичная ошибка наименьшая у второй модели, следовательно, модель y(x,z) – наиболее подходящая.

Код:

data<-read.table("D:\\reglab1.txt",header=TRUE)

colnames(data)=c("z","x","y")

lrx<-lm(x ~.,data=data,model=TRUE)

coef(lrx)

summary(lrx)

lry<-lm(y ~.,data=data,model=TRUE)

coef(lry)

summary(lry)

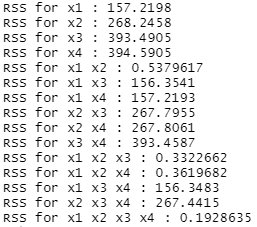
lrz<-lm(z ~.,data=data,model=TRUE)

coef(lrz)

summary(lrz)

## **Пункт 2**

Чтобы выявить зависимость качества работы модели от количества признаков, были построены несколько моделей и посчитаны остаточные суммы квадратов для каждой из них:



Таким образом, наилучшая модель – та, которая была обучена с помощью всех четырех признаков (наименьшая остаточная сумма квадратов). Наибольшая остаточная сумма квадратов наблюдается у моделей, обученных с помощью одного признака.

Код:

data<-read.table("D:\\reglab2.txt",header=TRUE)

colnames(data)=c("y","x1","x2","x3","x4")

col<-c("x1","x2","x3","x4")

n<-dim(data)[1]

rss<-double(20)

i<-1

for(d in 1:4){

com<-combn(col, d)

com

num\_com<-dim(com)[2]

for (k in 1:num\_com){

c<-com[,k]

lr1<-lm(as.formula(paste("y~",paste(c, collapse="+"))),data=data,model=TRUE)

rss[i]=deviance(lr1)

cat("RSS for", c, ":", rss[i], "\n")

i<-i+1

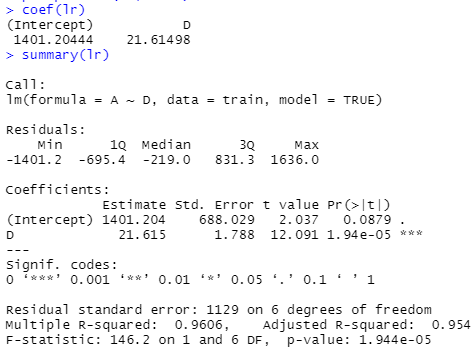
}

}

rss

## **Пункт 3**

Была построена регрессионная модель зависимости исследуемых отложений от глубины залегания:



Значение p-value меньше критического, следовательно, нулевая гипотеза отвергнута, модель можно признать статистически значимой.

Также была посчитан коэффициент детерминации полученной модели (R2):

R2 = 0.942802, что свидетельствует о хорошей точности модели.

Код:

data<-read.table("D:\\cygage.txt",header = TRUE)

colnames(data)=c("A","D","W")

n<-dim(data)[1]

data\_rand<-data[order(runif(n)), ]

n<-dim(data)[1]

delim<-as.integer(n\*0.7)

train<-data\_rand[1:delim, ]

test<-data\_rand[(delim+1):n, ]

lr<-lm(A~D,data=train,model=TRUE)

pr<-predict(lr, test)

coef(lr)

summary(lr)

plot(lr)

RSQUARE = function(y\_actual,y\_predict){

cor(y\_actual,y\_predict)^2

}

LR\_R <- RSQUARE(test[,3], pr)

LR\_R

## **Пункт 4**

Была построена обыкновенная линейная регрессия для набора данных longley, зависимая переменная – Employed.

Получилась модель регрессии, представленная ниже:

После исключения переменной "Population" из набора данных, он был разделен пополам на train и test выборки, с помощью train выборки была построена гребневая регрессия для значений . Затем были вычислены значения RMSE от каждого значения для обеих выборок, графики представлены на рисунках 1, 2.

## 

Рисунок 1 – Зависимость RMSE от для обучающей выборки

## 

Рисунок 2 – Зависимость RMSE от для тестовой выборки

Из полученных графиков видно, что среднеквадратичная ошибка убывает с увеличением значения параметра .

## **Пункт 5**

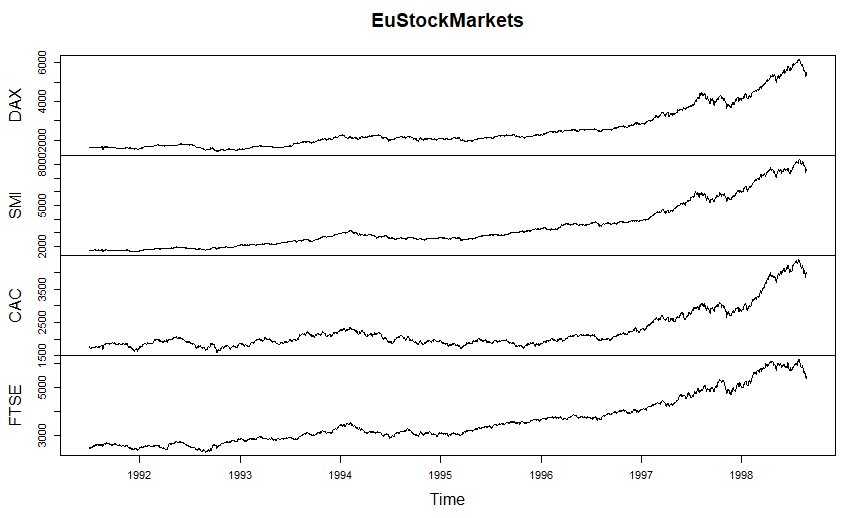


Рисунок 3 – Исходные данные фондовых бирж

Для каждой из бирж и для всех бирж вместе были получены следующие модели линейной регрессии:

Из четырех бирж коэффициент при параметре *Time* больше всего у Switzerland SMI, следовательно, данная биржа имеет наибольшую динамику, что так же видно из графика.

Код:

library(datasets)

data("EuStockMarkets")

data<-as.data.frame(EuStockMarkets)

print(data)

plot(EuStockMarkets)

DAX <- EuStockMarkets[, 1]

SMI <- EuStockMarkets[, 2]

CAC <- EuStockMarkets[, 3]

FTSE <- EuStockMarkets[, 4]

t<-time(EuStockMarkets)

DAXdata<-data.frame(time=t, DAX=DAX)

SMIdata<-data.frame(time=t, SMI=SMI)

CACdata<-data.frame(time=t, CAC=CAC)

FTSEdata<-data.frame(time=t, FTSE=FTSE)

all\_data<-data.frame(time=t, all=DAX+SMI+CAC+FTSE)

lrDAX<-lm(DAX~ ., data=DAXdata)

coef(lrDAX)

summary(lrDAX)

lrSMI<-lm(SMI~ ., data=SMIdata)

coef(lrSMI)

summary(lrSMI)

lrCAC<-lm(CAC~ ., data=CACdata)

coef(lrCAC)

summary(lrCAC)

lrFTSE<-lm(FTSE~ ., data=FTSEdata)

coef(lrFTSE)

summary(lrFTSE)

lr<-lm(all~ ., data=all\_data)

coef(lr)

summary(lr)

## **Пункт 6**

Были построены кривые изменения прибыли во времени с 1960 по 1980 гг. Графики представлены на рисунке 4.

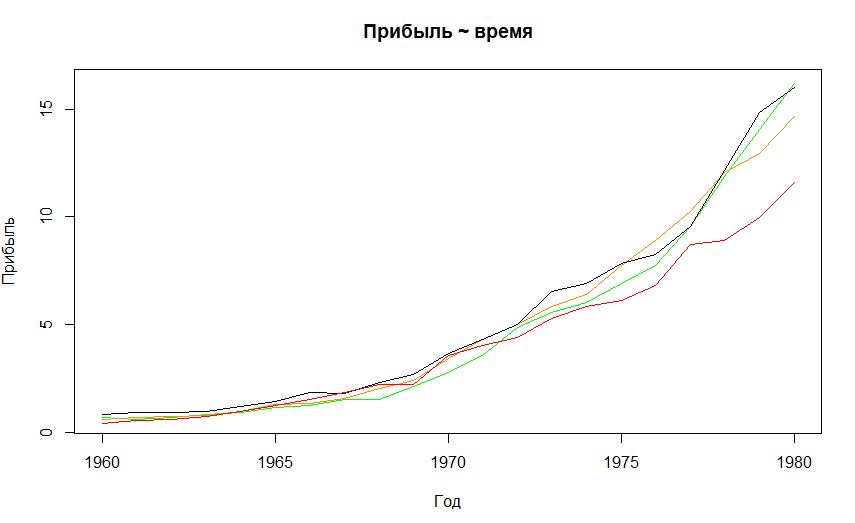


Рисунок 4 – Кривые зависимости прибыли от времени

Были построены 5 моделей линейной регрессии – для каждого квартала по отдельности и для всех кварталов вместе:

По построенным моделям можно сказать, что наибольшая динамика доходности - в третьем квартале, так как коэффициент при параметре time наибольший, наименьшая - в четвертом квартале, так как коэффициент при параметре time наименьший.

Прогноз прибыли в 2016 году:

* первый квартал - 36.75964
* второй квартал - 36.31813
* третий квартал - 37.30176
* четвертый квартал - 28.39273
* средняя по году - 34.92424.

Код:

library(datasets)

data("JohnsonJohnson")

data<-as.data.frame(JohnsonJohnson)

plot(time(JohnsonJohnson)[seq(from = 1, to = length(JohnsonJohnson), by = 4)],

JohnsonJohnson[seq(from = 1, to = length(JohnsonJohnson), by = 4)],

type = "l", col="green", main = "Прибыль ~ время", xlab="Год",

ylab="Прибыль")

lines(time(JohnsonJohnson)[seq(from = 1, to = length(JohnsonJohnson), by = 4)],

JohnsonJohnson[seq(from = 2, to = length(JohnsonJohnson), by = 4)],

type = "l", col="darkorange")

lines(time(JohnsonJohnson)[seq(from = 1, to = length(JohnsonJohnson), by = 4)],

JohnsonJohnson[seq(from = 3, to = length(JohnsonJohnson), by = 4)],

type = "l", col="black")

lines(time(JohnsonJohnson)[seq(from = 1, to = length(JohnsonJohnson), by = 4)],

JohnsonJohnson[seq(from = 4, to = length(JohnsonJohnson), by = 4)],

type = "l", col="red")

n<-length(JohnsonJohnson)

t1<-time(JohnsonJohnson)[seq(1,n,4)]

t2<-time(JohnsonJohnson)[seq(2,n,4)]

t3<-time(JohnsonJohnson)[seq(3,n,4)]

t4<-time(JohnsonJohnson)[seq(4,n,4)]

q1<-JohnsonJohnson[seq(1, n, 4)]

q2<-JohnsonJohnson[seq(2, n, 4)]

q3<-JohnsonJohnson[seq(3, n, 4)]

q4<-JohnsonJohnson[seq(4, n, 4)]

lr1<-lm(q1~., data=data.frame(time=t1, q1=q1))

lr2<-lm(q2~., data=data.frame(time=t2, q2=q2))

lr3<-lm(q3~., data=data.frame(time=t3, q3=q3))

lr4<-lm(q4~., data=data.frame(time=t4, q4=q4))

lr\_all<-lm(JohnsonJohnson[seq(from = 1, to = length(JohnsonJohnson), by = 4)]+

JohnsonJohnson[seq(from = 2, to = length(JohnsonJohnson), by = 4)]+

JohnsonJohnson[seq(from = 3, to = length(JohnsonJohnson), by = 4)]+

JohnsonJohnson[seq(from = 4, to = length(JohnsonJohnson), by = 4)]~

time(JohnsonJohnson)[seq(from = 1, to = length(JohnsonJohnson), by = 4)],

JohnsonJohnson)

lr1

lr2

lr3

lr4

lr\_all

pr1<-lr1$coefficients[1]+lr1$coefficients[2]\*2016

pr1

pr2<-lr2$coefficients[1]+lr2$coefficients[2]\*2016

pr2

pr3<-lr3$coefficients[1]+lr3$coefficients[2]\*2016

pr3

pr4<-lr4$coefficients[1]+lr4$coefficients[2]\*2016

pr4

pr\_all<-(1/4)\*(lr\_all$coefficients[1]+lr\_all$coefficients[2]\*2016)

pr\_all

## **Пункт 7**

Была построена кривая изменения количества солнечных пятен во времени, график представлен на рисунке 5.



Рисунок 5 – Кривая зависимости числа солнечных пятен от времени

Была построена модель линейной регрессии для данных из набора sunspot.year:

Код:

library(datasets)

data("sunspot.year")

data<-as.data.frame(sunspot.year)

years<-time(sunspot.year)

lr\_data<-data.frame(time=years, spots=sunspot.year)

lr\_data

lr<-lm(spots~., lr\_data)

plot(lr\_data$time, lr\_data$spots, col="darkgreen", type="l", main = "Количество солнечных пятен ~ время", xlab = "Год", ylab = "Количество")

coef(lr)

summary(lr)

## **Пункт 8**

Были построены 5 моделей линейной регрессии – для каждого квартала по отдельности и для всех кварталов вместе:

По построенным моделям можно сказать, что наибольшая динамика доходности - в третьем квартале, так как коэффициент при параметре time наибольший, наименьшая - в четвертом квартале, так как коэффициент при параметре time наименьший.

Прогноз прибыли в 2016 году:

* первый квартал - 2230.936
* второй квартал - 1072.375
* третий квартал - 501.9919
* четвертый квартал - 1654.785
* средняя по году - 1372.787.

Код:

library(datasets)

gas <- read.csv("D://UKgas.csv", sep = ',', dec = '.', header=TRUE, stringsAsFactors = FALSE)

plot(gas)

n<-length(gas$time)

n

t1<-gas$time[seq(1,n,4)]

t2<-gas$time[seq(2,n,4)]

t3<-gas$time[seq(3,n,4)]

t4<-gas$time[seq(4,n,4)]

q1<-gas$UKgas[seq(1, n, 4)]

q2<-gas$UKgas[seq(2, n, 4)]

q3<-gas$UKgas[seq(3, n, 4)]

q4<-gas$UKgas[seq(4, n, 4)]

lr1<-lm(q1~., data=data.frame(time=t1, q1=q1))

lr2<-lm(q2~., data=data.frame(time=t2, q2=q2))

lr3<-lm(q3~., data=data.frame(time=t3, q3=q3))

lr4<-lm(q4~., data=data.frame(time=t4, q4=q4))

lr\_all<-lm(gas$UKgas[seq(from = 1, to = n, by = 4)]+

gas$UKgas[seq(from = 2, to = n, by = 4)]+

gas$UKgas[seq(from = 3, to = n, by = 4)]+

gas$UKgas[seq(from = 4, to = n, by = 4)]~

gas$time[seq(from = 1, to = n, by = 4)],

gas)

lr1

lr2

lr3

lr4

lr\_all

pr1<-lr1$coefficients[1]+lr1$coefficients[2]\*2016

pr1

pr2<-lr2$coefficients[1]+lr2$coefficients[2]\*2016

pr2

pr3<-lr3$coefficients[1]+lr3$coefficients[2]\*2016

pr3

pr4<-lr4$coefficients[1]+lr4$coefficients[2]\*2016

pr4

pr\_all<-(1/4)\*(lr\_all$coefficients[1]+lr\_all$coefficients[2]\*2016)

pr\_all

## **Пункт 9**

Была построена модель линейной регрессии для набора данных cars:

С помощью построенной модели была оценена длина тормозного пути при скорости 40 миль в час - 139.7173 футов.

Код:

library(datasets)

data("cars")

data=cars

lr<-lm(dist ~ .,data=data)

coef(lr)

summary(lr)

test<-data.frame("speed"=40,"dist"="")

predict(lr,test)

## **Вывод**

При выполнении данной лабораторной работы были изучены методы линейной и гребневой регрессии. С помощью регрессии можно оценить динамику изменения величины, предсказать ее значение. Выбранная модель и ее параметры влияют на точность регрессии.

Были построены несколько моделей регрессии для различных датасетов, оценено качество их работы, выбраны лучшие параметры моделей, обучающие данные. А статистические критерии помогли оценить значимость этих моделей путем принятия или отвержения гипотез.