МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

Практикум №5

з курсу «Аналіз даних в інформаційних системах»

на тему: «Регресійні моделі»

Викладач:
Ліхоузова Т.А.

Виконав:
студент 2 курсу
групи ІП-14 ФІОТ
Шляхтун Денис

Київ-2023

Тема: Регресійні моделі.

Мета роботи: ознайомитись з різновидами регресійних моделей.

Основне завдання

Завантажити дані про якість червоного вина

1. Дослідити дані, підготувати їх для побудови регресійної моделі
2. Розділити дані на навчальну та тестову вибірки
3. Побудувати декілька регресійних моделей для прогнозу якості вина (12 - quality). Використати лінійну одномірну та багатомірну регресію та поліноміальну регресію обраного вами виду (3-5 моделей)
4. Використовуючи тестову вибірку, з'ясувати яка з моделей краща

Додаткове завдання

Завантажити дані файлу Data4.csv

1. Дослідити дані, сказати чи є мультиколінеарність, побудувати діаграми розсіювання
2. Побудувати декілька регресійних моделей (використати лінійну регресію та поліноміальну регресію обраного вами виду)
3. Використовуючи тестову вибірку з файлу Data4t.csv, з'ясувати яка з моделей краща

Виконання основного завдання.

Виконання комп’ютерного практикуму здійснювалося засобами R та RStudio.

1. Дослідити дані, підготувати їх для побудови регресійної моделі

Для виконання роботи було завантажено дані з файлу “winequality-red.csv”, було перевірено структуру та досліджено на пропущені значення, після перевірки додаткові маніпуляції з даними виявилися непотрібними.

```
> str(data)
'data.frame':   1599 obs. of  12 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
> summary(data)
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
 Min.    : 4.60   Min.    :0.1200   Min.    :0.000   Min.    : 0.900   Min.    :0.01200   Min.    : 1.00       Min.    :  6.00       Min.    :0.9901
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
 Mean    : 8.32   Mean    :0.5278   Mean    :0.271   Mean    : 2.539   Mean    :0.08747   Mean    :15.87       Mean    : 46.47       Mean    :0.9967
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
 Max.    :15.90   Max.    :1.5800   Max.    :1.000   Max.    :15.500   Max.    :0.61100   Max.    :72.00       Max.    :289.00       Max.    :1.0037
       pH          sulphates         alcohol          quality
 Min.    :2.740   Min.    :0.3300   Min.    : 8.40   Min.    :3.000
 1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
 Median :3.310   Median :0.6200   Median :10.20   Median :6.000
 Mean    :3.311   Mean    :0.6581   Mean    :10.42   Mean    :5.636
 3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.    :4.010   Max.    :2.0000   Max.    :14.90   Max.    :8.000
> data[!complete.cases(data),] #пропущених значень немає
 [1] fixed.acidity        volatile.acidity     citric.acid          residual.sugar       chlorides            free.sulfur.dioxide
 [7] total.sulfur.dioxide density              pH                   sulphates            alcohol              quality
<0 rows> (or 0-length row.names)
```

Рис. 1 – дослідження даних

2. Розділити дані на навчальну та тестову вибірки

Дані було розділено на третини, дві з яких складають навчальну вибірку і третя складає тестову.

```
> div <- nrow(data)/3*2
> ndata <- data[data$quality[1:div],] #навчальна вибірка - дві третини даних
> tdata <- data[data$quality[(div+1):nrow(data)],] #тестова вибірка
```

Рис. 2 – розподіл даних на навчальну і тестову вибірки

3. Побудувати декілька регресійних моделей для прогнозу якості вина (12 - quality). Використати лінійну одномірну та багатомірну регресію та поліноміальну регресію обраного вами виду (3-5 моделей)

Було створено 5 різних регресійних моделей:

```
> model1 <- lm(formula = quality ~ alcohol, data = ndata)
> summary(model1)

Call:
lm(formula = quality ~ alcohol, data = ndata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05233  0.00227  0.00227  0.00227  0.42037

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.78535    0.25854  -76.53   <2e-16 ***
alcohol       2.63650    0.02744   96.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08437 on 1064 degrees of freedom
Multiple R-squared:  0.8966,    Adjusted R-squared:  0.8965
F-statistic:  9229 on 1 and 1064 DF,  p-value: < 2.2e-16

>
> model2 <- lm(formula = quality ~ density+citric.acid, data = ndata)
> summary(model2)

Call:
lm(formula = quality ~ density + citric.acid, data = ndata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27050  0.02043  0.02043  0.02043  1.57995

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  142.32665   10.27150   13.86   <2e-16 ***
density     -137.64991   10.29633  -13.37   <2e-16 ***
citric.acid    1.63699    0.06094   26.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1914 on 1063 degrees of freedom
Multiple R-squared:  0.4682,    Adjusted R-squared:  0.4672
F-statistic:   468 on 2 and 1063 DF,  p-value: < 2.2e-16
```

Рис. 3 – регресійні моделі

```
> model3 <- lm(formula = quality ~ pH+density+alcohol, data = ndata)
> summary(model3)

Call:
lm(formula = quality ~ pH + density + alcohol, data = ndata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04208  0.00055  0.00055  0.00055  0.12733

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.89045    5.12907   11.68   <2e-16 ***
pH            0.55981    0.03721   15.04   <2e-16 ***
density     -83.26433    5.23832  -15.89   <2e-16 ***
alcohol       2.78992    0.02886   96.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0746 on 1062 degrees of freedom
Multiple R-squared:  0.9193,    Adjusted R-squared:  0.9191
F-statistic:  4034 on 3 and 1062 DF,  p-value: < 2.2e-16

>
> model4 <- nls(quality ~ a*pH^k, data=ndata, start=list(a=1,k=0.05))
> summary(model4)

Formula: quality ~ a * pH^k

Parameters:
  Estimate Std. Error t value Pr(>|t|)
a 14.92883    0.92224   16.19   <2e-16 ***
k -0.87180    0.04972  -17.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2327 on 1064 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 2.529e-07
```

Рис. 4 – регресійні моделі

```
> model5 <- lm(formula = quality ~ density+I(citric.acid^3), data=ndata)
> summary(model5)

Call:
lm(formula = quality ~ density + I(citric.acid^3), data = ndata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24561  0.02308  0.02308  0.02308  1.41192

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       195.5465     9.2428   21.16   <2e-16 ***
density          -190.9897     9.2653  -20.61   <2e-16 ***
I(citric.acid^3)    6.0355     0.1763   34.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1711 on 1063 degrees of freedom
Multiple R-squared:  0.5754,    Adjusted R-squared:  0.5746
F-statistic: 720.4 on 2 and 1063 DF,  p-value: < 2.2e-16
```

Рис. 5 – регресійні моделі

У всіх моделей $R^2 > 0,3$ і $p < 0,05$, тому ці моделі можна використовувати для опису впливу факторів на відгук.

4. Використовуючи тестову вибірку, з'ясувати яка з моделей краща

Для визначення кращої моделі обраховано сумарний квадрат відхилень від тестової вибірки для кожної моделі.

```
> sort(apply(tdata[13:17],2,function(x) sum(x-tdata$quality)^2), decreasing = FALSE)
   model3    model1    model5    model2    model4
 3.797185  5.209091 33.135677 48.493874 78.019570
```

Рис. 6 – визначення кращої моделі

Чим менше відхилення, тим краще, тому найкраща модель – третя.

Виконання додаткового завдання.

1. Дослідити дані, сказати чи є мультиколінеарність, побудувати діаграми розсіювання

Для виконання роботи було завантажено дані з файлу “Data4.csv”, було перевірено структуру та досліджено на пропущені значення, після перевірки додаткові маніпуляції з даними виявилися непотрібними.

```
> data <- read.csv("Data4.csv",sep=";",dec = ",", fileEncoding = "latin1")
> str(data)
'data.frame':   132 obs. of  7 variables:
 $ Country: chr  "Albania" "Algeria" "Angola" "Argentina" ...
 $ ISO    : chr  "ALB" "DZA" "AGO" "ARG" ...
 $ UA     : chr  "Àëáàí³ÿ" "Àëæèð" "Àíãîëà" "Àðãåíòèíà" ...
 $ Cql    : num  0.974 0.782 0.372 0.884 1.016 ...
 $ Ie     : num  0.605 0.587 0.274 0.7 0.718 ...
 $ Iec    : num  0.539 0.348 0.332 0.282 0.536 ...
 $ Is     : num  0.51 0.498 0.347 0.519 0.486 ...
> summary(data)
   Country             ISO                UA                 Cql               Ie               Iec              Is        
 Length:132         Length:132         Length:132         Min.   :0.2940   Min.   :0.1338   Min.   :0.2499   Min.   :0.2815  
 Class :character   Class :character   Class :character   1st Qu.:0.6804   1st Qu.:0.4209   1st Qu.:0.4111   1st Qu.:0.4315  
 Mode  :character   Mode  :character   Mode  :character   Median :0.9057   Median :0.5974   Median :0.4805   Median :0.5044  
                                                          Mean   :0.9210   Mean   :0.5627   Mean   :0.5034   Mean   :0.5131  
                                                          3rd Qu.:1.1848   3rd Qu.:0.7325   3rd Qu.:0.5875   3rd Qu.:0.5873  
                                                          Max.   :1.4576   Max.   :0.8224   Max.   :0.7860   Max.   :0.6983  
> data[!complete.cases(data),]
[1] Country ISO     UA      Cql     Ie      Iec     Is
<0 rows> (or 0-length row.names)
```

Рис. 7 – дослідження даних

Дослідження на мультиколінеарність:

```
> cor(data[,4:7])
          Cql        Ie       Iec        Is
Cql 1.0000000 0.8836642 0.8755446 0.9391724
Ie  0.8836642 1.0000000 0.6192470 0.7463204
Iec 0.8755446 0.6192470 1.0000000 0.7992111
Is  0.9391724 0.7463204 0.7992111 1.0000000
```

Рис. 8 – дослідження на мультиколінеарність

Майже всі змінні є мультиколінеарними.

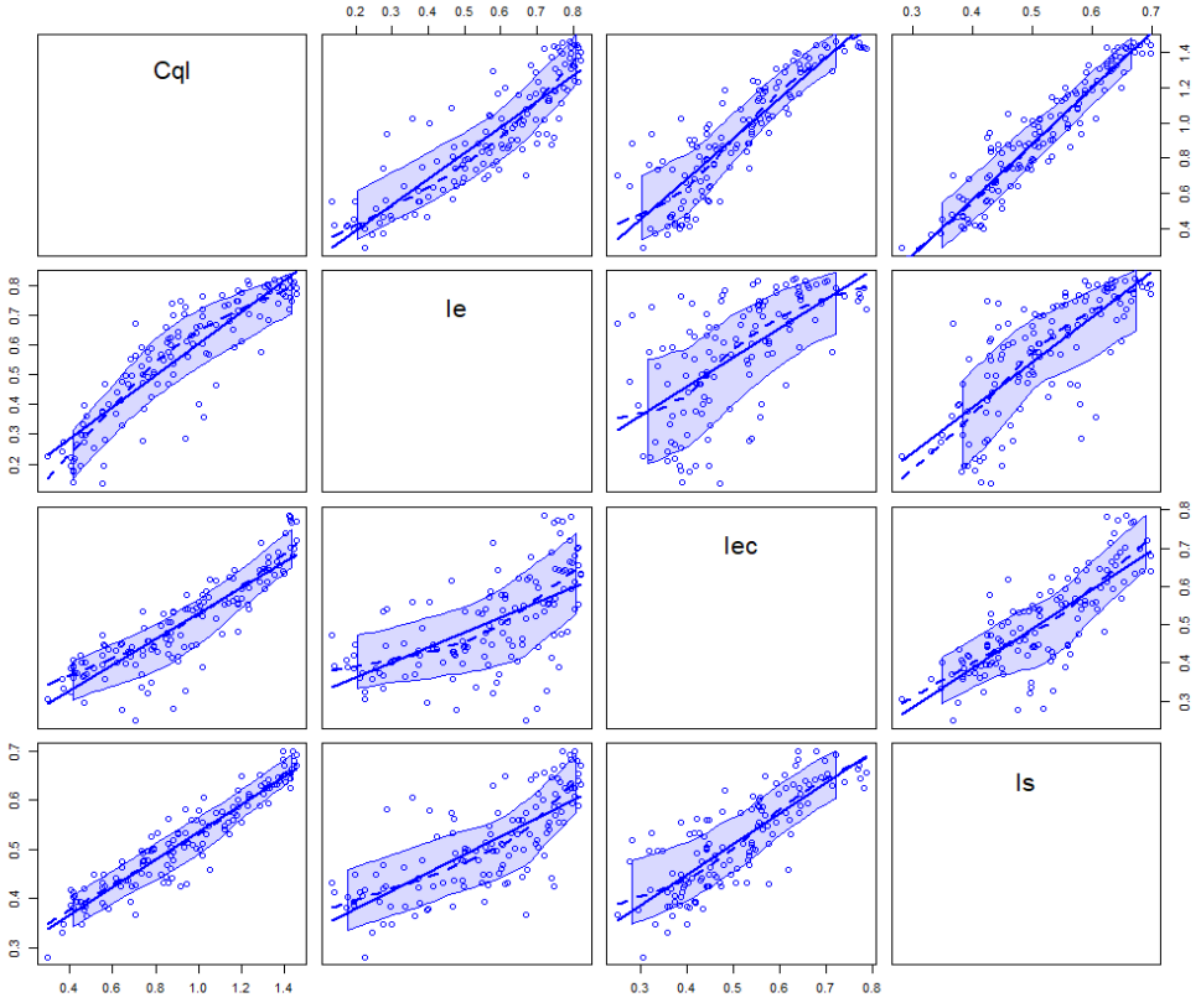

Рис. 9 – діаграми розсіювання

2. Побудувати декілька регресійних моделей (використати лінійну регресію та поліноміальну регресію обраного вами виду)

```
> model1 <- lm(formula = Cql ~ Is, data = data)
> summary(model1)

Call:
lm(formula = Cql ~ Is, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.220706 -0.081071  0.000618  0.070796  0.305205

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.70235    0.05295  -13.26   <2e-16 ***
Is           3.16375    0.10147   31.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1108 on 130 degrees of freedom
Multiple R-squared:  0.882,     Adjusted R-squared:  0.8811
F-statistic: 972.1 on 1 and 130 DF,  p-value: < 2.2e-16

>
> model2 <- lm(formula = Cql ~ Ie+Iec, data = data)
> summary(model2)

Call:
lm(formula = Cql ~ Ie + Iec, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.132573 -0.049996 -0.004477  0.044581  0.220035

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.30002    0.02532  -11.85   <2e-16 ***
Ie           0.92016    0.03920   23.47   <2e-16 ***
Iec          1.39690    0.06189   22.57   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06812 on 129 degrees of freedom
Multiple R-squared:  0.9557,    Adjusted R-squared:  0.955
F-statistic:  1392 on 2 and 129 DF,  p-value: < 2.2e-16
```

Рис. 10 – побудова регресійних моделей

```
> model3 <- nls(Cql ~ a*Ie^k, data = data, start=list(a=1,k=0.05))
> summary(model3)

Formula: Cql ~ a * Ie^k

Parameters:
  Estimate Std. Error t value Pr(>|t|)
a  1.56282    0.04240   36.85   <2e-16 ***
k  0.91774    0.05336   17.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1529 on 130 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 4.773e-06
```

Рис. 11 – побудова регресійних моделей

3. Використовуючи тестову вибірку з файлу Data4t.csv, з'ясувати яка з моделей краща

```
> sort(apply(tdata[8:10],2,function(x) sum(x-tdata$Cql)^2), decreasing = FALSE)
     model2      model1      model3
6.611341e-05 1.872079e-02 1.739798e-01
```

Рис. 12 – визначення кращої моделі

Друга модель найкраща.

Висновок.

При виконанні комп’ютерного практикуму було створено різні регресійні моделі, зокрема лінійні та поліноміальні. Регресії будувалися на навчальних вибірках і перевірялися на тестових за допомогою сумарних квадратів відхилень від тестової вибірки. Для додаткового завдання була проведена перевірка на мультиколінеарність та побудовані діаграми розсіювання.

Додаток А. Код мовою програмування R

```
data <- read.csv("winequality-red.csv", sep=",", header = TRUE, dec = '.')

#Дослідити дані, підготувати їх для побудови регресійної моделі
str(data)
summary(data)
data[!complete.cases(data),] #пропущених значень немає

#Розділити дані на навчальну та тестову вибірки
div <- nrow(data)/3*2
ndata <- data[data$quality[1:div],] #навчальна вибірка - дві третини даних
tdata <- data[data$quality[(div+1):nrow(data)],] #тестова вибірка

#Побудувати декілька регресійних моделей для прогнозу якості вина
model1 <- lm(formula = quality ~ alcohol, data = ndata)
summary(model1)

model2 <- lm(formula = quality ~ density+citric.acid, data = ndata)
summary(model2)

model3 <- lm(formula = quality ~ pH+density+alcohol, data = ndata)
summary(model3)

model4 <- nls(quality ~ a*pH^k, data=ndata, start=list(a=1,k=0.05))
summary(model4)

model5 <- lm(formula = quality ~ density+I(citric.acid^3), data=ndata)
summary(model5)

#Використовуючи тестову вибірку, з'ясувати яка з моделей краща
tdata$model1 <- predict(model1, tdata)
tdata$model2 <- predict(model2, tdata)
tdata$model3 <- predict(model3, tdata)
tdata$model4 <- predict(model4, tdata)
```

```
tdata$model5 <- predict(model5, tdata)
sort(apply(tdata[13:17],2,function(x) sum(x-tdata$quality)^2), decreasing = FALSE)

# додаткове завдання

#Дослідити дані
data <- read.csv("Data4.csv",sep=";",dec = ",", fileEncoding = "latin1")
str(data)
summary(data)
data[!complete.cases(data),]

#перевірка на мультиколінеарність
cor(data[,4:7])

#діаграми розсіювання
library(car)
scatterplotMatrix(~Cql+Ie+Iec+Is, data=data,diagonal=NA)

#Побудувати декілька регресійних моделей (використати лінійну регресію та поліноміальну регресію обраного вами виду)
model1 <- lm(formula = Cql ~ Is, data = data)
summary(model1)

model2 <- lm(formula = Cql ~ Ie+Iec, data = data)
summary(model2)

model3 <- nls(Cql ~ a*Ie^k, data = data, start=list(a=1,k=0.05))
summary(model3)

#Використовуючи тестову вибірку з файлу Data4t.csv, з'ясувати яка з моделей краща
tdata <- read.csv("Data4t.csv",sep=";",dec = ",", fileEncoding = "latin1")
tdata$model1 <- predict(model1, tdata)
tdata$model2 <- predict(model2, tdata)
tdata$model3 <- predict(model3, tdata)
sort(apply(tdata[8:10],2,function(x) sum(x-tdata$Cql)^2), decreasing = FALSE)
```